SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of
Informatics, Systems and Communication
Ph.D. program in Computer Science Cycle XXXVI

Decision-Theoretic Models for Information Retrieval

Peikos Georgios

Registration number: 865290

Supervisor: Prof. Gabriella Pasi

Tutor: Prof. Matteo Palmonari

Coordinator: Prof. Leonardo Mariani

ACADEMIC YEAR 2022/2023

University of Milano-Bicocca
Information and Knowledge Representation, Retrieval and Reasoning Laboratory
IKR3 Lab
Viale Sarca, 336 - 20126, Milano, Italy

Acknowledgements

The finalization of this dissertation marks the end of a three-year commitment to extensive research and creative pursuits. Throughout these years, I had the privilege of experiencing both professional and personal development. As is often the case in life, this growth was made possible by the presence of several exceptional people who accompanied me on this transformative journey. First and foremost, I dedicate this dissertation to my family, tireless pillars of support who have stood by me in every step, decision, and challenge.

I would like to express my gratitude and appreciation to my supervisor, Professor Gabriella Pasi, without whom this work would not have been completed. Moreover, I would like to thank her for creating an exceptional environment that facilitated my academic growth and personal development. I am deeply thankful for the opportunity to learn and evolve under her mentorship.

I would like to sincerely thank Professor Lorraine Goeuriot and Professor Karin Verspoor for reviewing this dissertation and offering invaluable feedback. Their insightful feedback enhanced the current dissertation and paved the way for future research endeavours. I would also like to thank Dr Symeon Symeonidis, the members of the DoSSIER project, and the members of the IKR3 lab for their valuable discussions over these years.

Finally, I would like to express my gratitude to my friends, ever-present listeners and supporters. I am sincerely thankful as, although you have never truly understood the content of this dissertation, you were always there to hear about it.

Milan, January 9, 2024 Georgios Peikos

Abstract

From a user's perspective, Information Retrieval (IR) constitutes a decision-making process. Users motivated by a specific situation engage in search activities to fulfil a related information need. Furthermore, it is common for users to assess the relevance of information items by considering both objective and subjective factors, such as topicality, domain expertise, recency, and others related to the characteristics of the search task. Consequently, there is a notable expectation for IR models to serve as intermediaries in this context and estimate the relevance score of information items by systematically quantifying and aggregating multiple relevance factors. Over the past few years, substantial research has been made into multidimensional relevance estimation, resulting in various proposed approaches. Nevertheless, it remains an ongoing research area with several unresolved issues and challenges.

Motivated by this, in this dissertation, we introduce a Decision-theoretic Multidimensional Relevance Framework (DtMRF), a generalizable IR framework for multidimensional relevance estimation. The framework accounts for positive and negative factors, which are first identified based on the characteristics of a search task, then assessed, and subsequently aggregated to provide an overall relevance estimate of an information item to a considered information need. DtMRF leverages Multiple Attribute Decision-Making (MADM) methods to incorporate user, task, and domain factors in the retrieval process, overcoming the computational complexity limitations of data-driven approaches while offering interpretable rankings. Moreover, we propose Neural-DtMRF, a hybrid framework that leverages neural architectures and a few training data to enhance the functionalities and effectiveness of DtMRF. Specifically, through training, Neural-DtMRF learns the degree to which the considered relevance factors affect the overall relevance in a search task.

To investigate the potential of DtMRF and Neural-DtMRF, we explored a search task within the medical domain, specifically the task of eligibility screening for clin-

ical trials. Our empirical evaluation showed that DtMRF and Neural-DtMRF have enhanced retrieval effectiveness when contrasted with neural models like BERT. Furthermore, as model-driven approaches, both DtMRF and Neural-DtMRF provide rankings that users can comprehensively interpret, a valuable characteristic in professional search contexts. This interpretability feature facilitates informed decision-making and allows for further research and application of these models in complex medical information retrieval scenarios. Finally, we integrate Neural-DtMRF with Large Language Models (LLMs) to enhance patient eligibility assessment and improve retrieval performance for this specific task.
In conjunction with the introduction of DtMRF and its neural extension, we address the challenging task of extracting patient-related information from unstructured medical summaries within Electronic Health Records (EHRs). Our investigation delves into the performance of domain-specific pre-trained language models (PLMs), such as BioBERT, and LLMs, like GPT-3.5, for information extraction and query formulation tasks. Regarding retrieval performance, queries generated by GPT-3.5 outperformed those formulated using the other approaches.
Building on the acquired insights, we designed a conceptual framework tailored to clinical trials retrieval and developed a prototype system for its implementation. The system combines the strengths of GPT-3.5 for information extraction with Neural-DtMRF for multidimensional relevance estimation. The resulting retrieval system can identify relevant clinical trials and provide interpretable rankings, assisting medical professionals in making informed decisions.

Keywords: Decision-theory, Decision-theoretic Retrieval Framework, Multidimensional Relevance Estimation, Interpretable Ranking, Clinical Trials Retrieval, Large Language Models.

Publications

Within the completion of this dissertation, several publications pertinent to the research content have been produced, some of which are already published or are currently submitted for peer review.

1. Peikos et al. [2023a] Georgios Peikos, Daria Alexander, Gabriella Pasi, Arjen P. de Vries, *"Investigating the Impact of Query Representation on Medical Information Retrieval"*, ECIR (2), Lecture Notes in Computer Science, Vol. 13981, pp. 512–521, Springer, 2023.

   Discussed in Chapter 6.

2. Peikos et al. [2023b] Georgios Peikos, Symeon Symeonidis, Pranav Kasela, Gabriella Pasi, *"Utilizing ChatGPT to Enhance Clinical Trial Enrollment"*, CoRR, Vol. abs/2306.02077, 2023. Submitted for peer review.

   Discussed in Chapter 6.

3. Peikos et al. [2022] Georgios Peikos, Oscar Espitia, Gabriella Pasi, *"UNIMIB at TREC 2021 Clinical Trials Track"*, CoRR, Vol. abs/2207.13514, 2022.

   Preliminary results included in Chapter 7.

4. Peikos and Pasi [2021] Georgios Peikos, Gabriella Pasi, *"Multidimensional Relevance in Legal and Health Domains"*, IIR, CEUR Workshop Proceedings, Vol. 2947, CEUR-WS.org, 2021.

   Discussed in Chapter 2.

5. Georgios Peikos, Gabriella Pasi, *"A Systematic Review of Multidimensional Relevance Estimation in Information Retrieval,"* Submitted for peer review.

Discussed in Chapter 3.

6. Georgios Peikos, Gabriella Pasi, *"Integrating Positive and Negative Relevance Factors: A Decision-theoretic Framework to Multidimensional Relevance Estimation,"* Submitted for peer review.

   Discussed in Chapters 5 and 7.

7. Georgios Peikos, Gabriella Pasi, *"Leveraging Neural and Decision-Theoretic Models for Multidimensional Relevance Estimation,"* Submitted for peer review.

   Discussed in Chapters 5 and 7.

8. Georgios Peikos, Gabriella Pasi, *"A prototype Search System for Clinical Trials Retrieval,"* Submitted for peer review.

   Discussed in Chapter 8.

Other research publications related to IR and information processing:

1. Kusa et al. [2022] Wojciech Kusa, Georgios Peikos, Oscar Espitia, Allan Hanbury, Gabriella Pasi, *"DoSSIER at MedVidQA 2022: Text-based Approaches to Medical Video Answer Localization Problem"*, BioNLP@ACL, pp. 432–440, Association for Computational Linguistics, 2022.

2. Askari et al. [2022] Arian Askari, Georgios Peikos, Gabriella Pasi, Suzan Verberne, *"LeiBi@COLIEE 2022: Aggregating Tuned Lexical Models with a Cluster-driven BERT-based Model for Case Law Retrieval"*, CoRR, Vol. abs/2205.13351, 2022.

3. Symeonidis et al. [2022] Symeon Symeonidis, Georgios Peikos, Avi Arampatzis, *"Unsupervised consumer intention and sentiment mining from microblogging data as a business intelligence tool"*, Oper. Res., Vol. 22, No. 5, pp. 6007–6036, 2022.

4. Arampatzis et al. [2021] Avi Arampatzis, Georgios Peikos, Symeon Symeonidis, *"Pseudo relevance feedback optimization"*, Information Retrieval Journal, Vol. 24, No. 4-5, pp. 269–297, Springer, 2021.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter serves as the foundational introduction to our research objectives and aims. It provides an overview of the essential concepts and goals that underpin our work while also specifying the research background and the communities that may benefit from this research. This chapter sets the stage for a comprehensive understanding of the research presented in this dissertation.

1.1 Research Context and Objectives

The research presented in this dissertation is situated within the field of Information Retrieval (IR), with a particular emphasis on estimating the relevance of the information retrieved in response to a user's query. The primary aim is to propose a framework that bridges the gap between user expectations and system output, striving to create user-centric and task-centric IR systems. Furthermore, this dissertation pursues a secondary aim. The second aim of our research involves studying and analyzing a complex professional search task within the medical domain, proposing a comprehensive search solution by leveraging the proposed framework.

The existing research landscape of multidimensional relevance estimation in IR highlights considerable progress over the past years, yielding a spectrum of proposed methodologies. Nonetheless, it remains a dynamic and evolving research field with numerous unresolved issues and challenges. As a result, the necessity of multi-aspect relevance estimation is widely acknowledged in the literature, especially when addressing complex search tasks commonly found in professional search environments. In these search tasks, relevance estimation should be expanded to encompass a broader spectrum of relevance aspects. Furthermore, these searches have specific requirements, such as the need for explainable rankings and enabling user control over the search process. This dissertation contributes towards this direction by proposing a formal decision-theoretic framework for estimating relevance that meets the objectives mentioned above and can be applied to various search tasks. This framework aims to benefit the broader IR community and those involved in designing search systems for professional search. Furthermore, we integrate it with neural and large language models to harness their combined capabilities and enhance its retrieval effectiveness.

As part of our endeavor to evaluate the utility of the proposed framework in a professional search task, our objective is to enhance retrieval effectiveness in clinical trials retrieval and the task of patient eligibility screening. This search task is complex and occurs in professional environments, thus making it sensitive to the previously mentioned requirements. The currently proposed retrieval approaches in this task need to be improved regarding the clarity of the relevance estimation process and the interpretability of the obtained ranking, so that an expert user can have control over the search process. Addressing the evident gaps in current retrieval approaches, our framework is tailored to address these challenges, ensuring clarity in relevance estimation and enhanced interpretability of rankings. The framework developed during the doctoral research can be beneficial to several professional contexts, akin to the one in which we conduct our evaluation.

In clinical trials retrieval, the existence of unstructured medical narratives necessitates applying specific information processing techniques, primarily aiming to extract information. Subsequently, these extracted pieces of information are crucial in enhancing retrieval effectiveness. While the literature offers numerous information extraction (IE) approaches, they are often fine-tuned and evaluated on specific benchmark collections. At the same time, many of these approaches are narrowly tailored to extract specific medical information types, like drugs or medical conditions. The field would benefit significantly from systems capable of comprehensive extraction from medical narratives. Addressing this need, our research makes two contributions. We evaluate the effectiveness of existing state-of-the-art IE approaches in enhancing clinical trial retrieval. Also, we propose a novel approach utilizing Large Language Models (LLMs) for information extraction. This method can extract various medical information, leading to better performance than prior proposed techniques. Our research insights hold significance for scholars exploring Large Language Models within the medical field and pave the way for future research directions.

The third contribution of our research is a search prototype specifically developed to assist medical experts in the patient eligibility screening process for clinical trials. This prototype encapsulates all the techniques we have implemented throughout our research, aiming to offer an end-to-end solution for each stage of the eligibility screening to facilitate patient enrollment in clinical trials. The prototype is designed to diminish the manual review burden for medical experts. It offers seamless integration capabilities with an organization's existing infrastructure, including potential synchronization with its electronic health record-storing system. Our research findings point to the prototype's evolution towards enabling automated initial screenings of patients, achieving this at a significantly reduced cost compared to current screening procedures.

1.2 Research Questions

Our research aims to address the following research questions:

(RQ1) How is the notion of relevance defined, analyzed, and applied across various knowledge domains and search tasks, particularly with respect to its multidimensional nature?

(RQ2) How can decision-making methods, specifically MADM methods, be effectively integrated into IR to enhance and interpret multidimensional relevance estimation?

(RQ3) How can neural-based methodologies enhance and expand standard decision-making methods in the context of IR, and what implications do they present for multidimensional relevance estimation?

(RQ4) How do the presence of various content characteristics, including medical entities, negations, patient and family history, influence retrieval performance in clinical trials retrieval?

(RQ5) How does the deployment of Large Language Models (LLMs) in extracting information from medical narratives compare to standard state-of-the-art methodologies, and what implications arise from their application?

(RQ6) What are the effectiveness and practical implications of employing DtMRF, Neural-DtMRF, and Neural-DtMRF integrated with LLMs in the clinical trials retrieval process, especially regarding retrieval performance and task requirements?

(RQ7) How does the design and implementation of a search prototype tailored for clinical trials retrieval and patient eligibility screening influence the efficiency of the search process?

The previously mentioned primary research questions are further broken down into sub-questions, each presented and addressed in the subsequent chapters of this dissertation.

1.3 Research Contributions

Our research has the following contributions:

1. A comprehensive and systematic literature review covering 70 studies aiming to identify the current research state of multidimensional relevance estimation in IR. This review aims to enhance our understanding of how relevance has been conceptualized and operationalized as a multidimensional concept across various application domains. The findings offer practical guidance for tailoring system designs to achieve closer alignment with users' perspectives of relevance in various search tasks and domains.

2. The definition of a decision-theoretic framework for multidimensional relevance estimation that considers relevance factors with either positive or negative influences on relevance. The proposed Decision theoretic Multidimensional Relevance Framework (DtMRF) leverages Multi-attribute Decision-Making (MADM) methods to incorporate user, task, and domain factors in the retrieval process, overcoming the computational complexity limitations of data-driven approaches while offering interpretable document rankings. The framework exploits scoring-based and distance-based MADM methods showcasing how these methods can be employed for document ranking.

3. An expansion of the DtMRF that incorporates a neural model to enhance multidimensional relevance estimation and add new capabilities. This integration leverages the predictive strengths of neural models while capitalizing on DtMRF's ability to produce interpretable document rankings.

4. A comparative evaluation of widely-used rule-based methods, pre-trained language models, and their hybrid combinations, focusing on information extraction from clinical narratives. This study serves as a performance benchmark for subsequent research in this field.

5. The application of LLMs and the assessment of their effectiveness in extracting information from clinical narratives. The study employs various in-context learning strategies, discusses their practical implications, and benchmarks their performance against previous state-of-the-art methods and medical experts.

6. The definition and implementation of a retrieval pipeline incorporating an LLM on top of the proposed Neural-DtMRF to enhance relevance estimation in the task of clinical trials retrieval. The proposed retrieval system offers a comprehensive solution for identifying eligible patients for clinical trials.

7. The development of a search prototype specifically designed for the requirements of patient eligibility screening and clinical trials retrieval. This prototype addresses the unique challenges in this search task by integrating all of the research contributions mentioned above.

1.4 Thesis Organization

The subsequent sections of this dissertation are divided into four main parts, as outlined below.

Part I: Background

This part is organized into 3 chapters and is dedicated to the fundamental concepts of Information Retrieval and Decision Theory, which are central to our research. It also offers insights into recent advancements in LLMs. Furthermore, it outlines the research field of multidimensional relevance estimation in IR, along with the task of clinical trials retrieval, which serves as the application domain of our research.

Chapter 2: Foundational Concepts and Research Methods

This chapter provides the reader with the necessary background information to comprehend the context of our research; it specifically focuses on IR and decision theory, which are its two pillars. Additionally, it offers insights into tools required in particular parts of our work, such as LLMs and the task of multi-output regression.

Chapter 3: Multidimensional Relevance Estimation: A Systematic Literature Review

This chapter presents the systematic literature review we conducted to assess the current landscape of multidimensional relevance estimation in IR aiming to discern emerging trends and potential avenues for future research.

Chapter 4: Clinical Trials Retrieval

This chapter introduces the task of clinical trials retrieval, which is the professional search task we consider to evaluate the effectiveness of the DtMRF model we propose. Our work is oriented around two directions: processing unstructured patient information from Electronic Health Records (EHRs) and improving its retrieval performance. This chapter reviews existing literature in these areas, identifying research limitations our study seeks to address. Additionally, the chapter introduces the benchmark collections employed to assess the efficacy of our proposed approaches.

Part II: Conceptualizing the Decision Theoretic Framework

This part comprises *Chapter* 5: *A Decision Theoretic Framework for Multidimensional Relevance Estimation* that defines and formalizes the proposed Decision-theoretic Multidimensional Relevance Framework (DtMRF). This chapter introduces and defines the components associated with the DtMRF framework, with illustrations highlighting their applicability in the context of IR. This presentation guides the reader to comprehend how DtMRF can be universally applied across diverse search tasks. The chapter showcases how DtMRF leads to interpretable document rankings and how its end-users can control the relevance estimation process. Additionally, the chapter introduces and formalizes Neural-DtMRF. This approach integrates neural models into DtMRF without compromising its ranking interpretability. The chapter discusses the additional components for leveraging Neural-DtMRF in search. In conclusion, the chapter illustrates how the synergy of neural models can augment the retrieval effectiveness of DtMRF and the benefits of the framework regarding ranking interpretability.

Part III: Putting Theory to the Test: Experimental Insights

This part is composed of two chapters and aims to present the empirical evaluation in the context of clinical trials retrieval, with a specific focus on patient eligibility screening. Our evaluation involves the Information extraction and the evaluation of the DtMRF, Neural-DtMRF and Neural-DtMRF with LLMs.

Chapter 6: *Extracting Information from Electronic Health Records*

This chapter examines our methodologies for information extraction from EHRs, explicitly targeting enhancing clinical trials retrieval. The study offers a comparative assessment of state-of-the-art IE methods applied to medical narratives. Additionally, it defines and presents our methodology for IE that leverages LLMs, and compares its effectiveness to the state-of-the-art approaches and medical experts.

Chapter 7: DtMRF, Neural-DtMRF, and LLMs for Clinical Trials Retrieval

This chapter offers an exhaustive evaluation of experimental outcomes related to the utilization of the DtMRF and its Neural extension in the context of clinical trials retrieval. Moreover, it showcases the retrieval effectiveness of the proposed approach that leverages an LLM on top of Neural-DtMRF.

Part IV: From Conceptualization to Development: A Search Prototype

This part composed of *Chapter 8*: *A prototype Search System for Clinical Trials Retrieval* presents the search prototype that unifies the research findings of our research. Specifically, it introduces the first version of the developed search prototype explicitly designed for clinical trials retrieval, focusing on the eligibility screening process. The prototype is developed to accommodate various requirements inherent to distinct phases of the process.

Part V: Overall Insights

This part, comprises *Chapter 9*: *Conclusions and Further Research*, marks the conclusion of this dissertation and outlines future directions for our research.

Part I Background

Chapter 2

Foundational Concepts and Research Methods

This chapter unveils this dissertation's foundational concepts and research methodologies, emphasizing Information Retrieval and Decision Theory. It delves into important IR aspects, especially the significance of the notion of relevance and the process of its estimation. The exploration extends to Decision Theory, showcasing notable methods within the field. The presented mathematical definitions serve as the theoretical background for the proposed retrieval framework. The chapter also covers a synthetic presentation of Language Models, highlighting their evolution towards in-context learning, which is essential for specific segments of our work. It also touches on multi-output regression, offering a definition and methods to tackle such problems. This chapter aims to provide a thorough understanding of the methods utilized, acting as a reference point for interpreting results in ensuing chapters.

2.1 Information Retrieval

The term "Information Retrieval", often abbreviated as IR, was first introduced by Mooers in 1951. Later that decade, the first IR systems emerged, driven by the rapidly increasing volumes of data and the need for more efficient methods to locate pertinent information within extensive repositories. Mooers conceptualized the term Information Retrieval as follows.

> *"Information Retrieval is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him (user). Information Retrieval embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, techniques, and machines that are employed to carry out the operation."*

Salton [1968], another pioneer in the field, provides his interpretation that explicitly presents the four key elements related to IR.

> *"Information retrieval systems are designed to help analyze and describe the items stored in a file, to organize them and search among them, and finally to retrieve them in response to a user's query. Designing and using a retrieval system involves four major activities: information analysis, information organization and search, query formulation, and information retrieval and dissemination."*

The primary goal of an information retrieval system (search engine) is to enable users to find relevant information within a huge repository in response to a specific information need (query). Queries typically contain keywords or phrases that express a user's perceived information need and are formally represented by an underlying formal language. It is important to note that a query can extend beyond the typical keywords or phrases in specific search situations and encompass an entire text or document.

While traditionally focused on text documents, modern systems have expanded to include various information items (e.g. images, videos). A retrieval process is initiated when a user inputs a query that represents an informational need. The system searches an indexed repository for potential information items that match the expressed request. In order for this process to be performed, both the user's query and the information items have the same representations. The retrieved information items are ranked by means of retrieval algorithms that estimate their relevance to the query. Traditionally, relevance estimation is based on topical similarity, assessed based on different mathematical methods. This process has been significantly enhanced through the adoption of pre-trained language models. Nonetheless, modern search systems have slightly broadened their criteria for estimating relevance by integrating a few additional signals. An excellent example of this advancement is the employment of the PageRank algorithm, which leverages link structure to refine search results [Brin and Page, 1998]. The final stage of a search system entails presenting a ranked list of information items to its end-user for further investigation.

The search process previously outlined pertains to ad-hoc retrieval. However, domain-specific search systems also exist, which are tailored to specialized fields or subjects. Domain-specific search is defined by Lupu et al. [2014] as a search focused on a specific subject area with various modalities (e.g. text, images) that involve a variety of users, tasks, and technical aspects (e.g. specific vocabularies). One example of domain-specific search can be observed in the healthcare sector. In this domain, various medical professionals, such as general practitioners and clinicians, engage in searches to meet their health-related information needs [Kritz et al., 2013]. While these professionals share a certain degree of specialized knowledge in healthcare, they typically engage in distinct search tasks. These tasks may necessitate the retrieval of varied modalities and could involve the use of terminology tailored to their medical specialization. However, even laypeople engage in domain-specific searches for health-related information, often using commercial search engines or social platforms. In constructing a domain-specific search system, several key considerations must be addressed. The employed search system is essential to accommodate the varying information needs expressed by experts and laypeople. It should incorporate a domain-specific vocabulary into the search algorithm and be capable of retrieving different data types. Finally, it is essential to have a user interface designed to facilitate particular domain-specific tasks, thus making the system more user-friendly.

Within the broader category of domain-specific search, a more specialized type of search exists, namely professional search. This type of search is tailored to meet the unique information needs of professionals within a given domain. Unlike general domain-specific searches, which experts and laypeople can use, professional search is designed to handle the intricate queries and requirements that professionals often encounter. Verberne et al. [2018] defines professional search as follows.

> *"Professional search takes place in the work context, by specialists, and using specialist sources, often with controlled vocabularies. (...) Professional search has the key benefit that the task to be solved is, usually, clear; at least to the person who carries out the searches."*

A study exploring common characteristics among professionals across four domains revealed specific search practices and goals among them [Russell-Rose et al., 2018]. The study highlighted the universal emphasis on the need for transparency and repeatability in the ranking algorithm across the various professional domains examined. In the healthcare sector, professionals generally engaged in recall-oriented search tasks, while legal researchers prioritized precision-oriented tasks, seeking recent and credible results. Another study aimed to understand professionals' typical search tasks by coding them based on their characteristics [Verberne et al., 2019]. The study revealed that many professionals conduct searches on behalf of other colleagues. This significantly complicates the relevance assessment of the obtained results and enhance the need of interpretable search systems. Additionally, in professional search, users often engage in extended search sessions that can be interrupted and resumed [Lupu et al., 2014].

In designing a professional search engine, key considerations include algorithmic transparency and repeatability, which are crucial across multiple domains. The system must be tailored to accommodate recall-oriented or precision-oriented search tasks, as the specific professional domain dictates. It should also offer flexibility in result interpretation, as searches might be conducted on behalf of others. Finally, effectively navigating specialized sources using controlled vocabularies is essential as part of domain-specific search. Nonetheless, in a professional search context, there is often a greater emphasis on the quality and usefulness of search results rather than the retrieval speed. That is particularly true in domains where the cost of an inaccurate result can be significant, such as healthcare or legal search. Therefore, a professional search engine could trade off some speed for increased effectiveness, interpretability, and specificity, meeting the complex requirements of professional users.

In each of the aforementioned retrieval contexts, the system must attain various constraints to ensure the retrieval of documents relevant (useful) to the user's needs. The following section aims to clarify the distinction between relevant and useful information within the scope of IR by analyzing the notion of relevance in the field.

2.1.1 The Notion of Relevance in Information Retrieval

From the first search engines in the late 1950's to the present day, the notion of relevance has been a central area of research. Research in this field seeks to explore which information items ought to be deemed relevant in relation to a specific information need, a user, the knowledge domain, or the task that a user aims to accomplish. Relevance refers to a relation between information items and some other concept [Saracevic, 2016b]. A fundamental distinction of the notion of relevance is its dual nature. On one hand, there is "user relevance," capturing the user's perception of what constitutes useful information. On the other hand, there is "system relevance," which is determined algorithmically by the retrieval system itself [Vickery, 1959b,a, Cooper, 1971, Swanson, 1986].

Users' engagement in search activities is commonly motivated by tasks stemming from persistent and evolving problematic situations [Belkin, 2016a]. Search activities can take place in professional settings where individuals often assume varied roles, such as researcher or educator. These roles are associated with specific information needs, whether it be for the purpose of publishing an academic paper or preparing presentation slides [Soufan et al., 2021]. In such searches, as users are presented with information items provided by search systems, a complex cognitive decision-making process is initiated, ultimately leading to them choosing useful items for further examination. The decision-making process is grounded in what Vickery [1959a] termed and what is commonly acknowledged in the field of IR as user relevance. This notion of relevance concerns how users evaluate information as pertinent to their information needs. Changing from a user-centric perspective to a system-oriented one, we focus on the inherent mechanisms by which search systems operate. Central to their operation is a concept highlighted by scholars, as system relevance [Vickery, 1959b, Saracevic, 2016b]. This concept encapsulates a system's ability to retrieve information items in line with an information need and consequently estimate their relevance based on an algorithm or model. This system relevance serves as an approximation to the aforementioned user relevance, aiming to align system outputs with user expectations.

Over time, scholars from varied backgrounds have proposed additional definitions to capture the notion of relevance. These definitions range from *affective relevance* tied to users' emotions and motivations to *situational relevance* addressing specific tasks, *system or algorithmic relevance* determined by query and information matching using an algorithm, *topical relevance* focusing on the relation between the topic expressed in a query and topic covered by information objects, and *cognitive relevance* connecting to a user's knowledge and the information's novelty [Saracevic, 1997, Mizzaro, 1998, Cosijn and Ingwersen, 2000, Borlund, 2003, Ingwersen and Järvelin, 2005, Cosijn, 2009, Belkin, 2016b]. While each definition adopts a distinct viewpoint, they all describe a form of relationship to information. We direct readers interested in a comprehensive understanding of relevance in information sciences to the book by Saracevic [2016b].

The notion of relevance has also been investigated within certain knowledge domains, as researchers have attempted to decompose it and identify the factors that contribute to information's utility (i.e. usefulness) for users. van Opijnen and Santos [2017] provide an in-depth analysis of the concept of relevance in the legal domain, drawing on the relevance classifications presented by Saracevic [2016b]. Similarly, the idea of relevance has been explored in e-commerce. Tsagkias et al. [2021] identify four key dimensions that shape e-commerce relevance: user, time, query, and context, such as a product's category, highlighting the domain-specific nature of relevance. Extending the framework presented by Mizzaro [1998], Crestani et al. [2017] discuss the characteristics of relevance in mobile search settings. Additionally, the study by Balagopalan et al. [2023] investigates the role of relevance in attaining fair rankings. The authors highlight the modifications necessary to meet the specific demands of this task.

Mainly by conducting user studies, numerous scholars have identified factors (i.e. *relevance factors*) that users take into account when assessing relevance in specific search scenarios, i.e. investigating what is referred to as user relevance. While a comprehensive examination of all these studies is beyond the scope of our review, we highlight a few representative ones here. For a more extensive exploration, readers can refer to the book by Saracevic [2016b], as a starting point. Some key studies in this research field are the studies by Cool et al. [1993], Barry and Schamber [1998], and Xu and Chen [2006a], among others. Xu and Chen [2006a] conduct a user study centered on web searches. They investigate the significance of criteria such as information novelty, topicality, reliability, and understandability, among others, in these searches. The findings highlight that topicality and novelty are the foremost criteria for relevance, with understandability being the subsequent priority. Sun et al. [2019] in their systematic literature review identify the criteria and indicators consumers use to evaluate the quality of online health information. Their research highlights multiple criteria, with trustworthiness, expertise, and objectivity being the most important across studies. Additionally, dominant indicators are related to the web page's source, content, and design. Other studies reveal that assessing relevance based only on topicality is not sufficient for medical experts, as they leverage their own knowledge and experience [Tamine and Chouquet, 2017]. Similar studies can be also found in other domains, such as the legal domain. The study by Wiggers et al. [2018] identifies factors affecting relevance assessment in legal professional searches, such as document type, recency, depth level, and legal hierarchy. Also Chu [2011] aims to discern factors influencing relevance judgments and their relative significance, in legal search. The study highlights several relevance factors, with specificity/amount of information, ease of use, and subject matter having being the most essential. The findings from the aforementioned and other related studies hold significant value. Mainly because they can guide the development of retrieval systems specifically tailored to certain search situations, ensuring a better approximation to user relevance in these tasks.

Drawing from the studies and definitions mentioned above and also supported by the study of Schamber et al. [1990], the notion of relevance emerges as a *multidimensional cognitive concept* influenced by users' perceptions of information and their distinct contextual situations. This concept is also *dynamic*, depending on users' perspective of the provided information in time. Nonetheless, as Schamber et al. [1990] conclude, relevance is a complex but *systematic and measurable* concept. In our research, we perceived multidimensional relevance as the estimation of relevance by information retrieval systems (i.e. *algorithmic relevance*) that consider multiple *relevance factors*, including user and task characteristics or other domain-specific requirements. These systems acknowledge that various factors influence relevance estimation, and they aim to integrate them into the retrieval process to better approximate user relevance.

As stated, topical relevance remains the fundamental method by which a search engine evaluates the relevance of information. The content delivered must be pertinent to the search query's subject matter. Therefore, in the subsequent section, we introduce statistical retrieval models commonly used for estimating topical relevance. Due to space constraints, we limit our discussion to the models specifically employed in our research.

2.1.2 Topical Relevance Estimation in Information Retrieval

Information Retrieval systems commonly attempt to approximate user relevance by estimating how closely the content of documents aligns with the textual content of the expressed information need (i.e. query). This form of relevance is termed as topical relevance, and in numerous studies, it is referred to simply as relevance. Another direction of research follows a different perspective, exploring the dynamic nature of relevance, i.e. how the perception of relevance changes over time and through user-system interactions, as seen in studies focused on interactive IR [Liu, 2021]. Other researchers delve into the multidimensional notion of relevance, suggesting that it is shaped by factors related to the user, the undertaken task, and the knowledge domain. Finally, these aspects are also addressed holistically, in systems using relevance models that rely on multiple factors and account for their evolving nature over time. We examine these studies in Chapter 3, given their close relation to the multidimensional relevance framework we introduce in our research. In this section we solely present models that estimate the topical relevance.

The most commonly employed model for estimating topical relevance in retrieval systems is the BM25, referred to as "Okapi BM25," as introduced by Robertson and Walker [1994]. BM25 derives from the 2-Poisson model and the probabilistic binary independence model of relevance. The 2-Poisson model aims to identify the most informative terms of a document. This model is based on the mixture of two Poisson distributions and it requires estimating three parameters for each term in the vocabulary, which is its drawback. However, this model does not need a term weighting algorithm to be implemented. To rank the documents with respect to a query, a measure based on the means of the two Poisson distributions was proposed by Harter [1975a,b]. The binary independence model, introduced by Robertson and Jones [1976], ranks the documents based on the odds of relevance, i.e. the division of the probability of relevance and non-relevance. Here, documents and queries are represented as binary vectors; consequently, terms in a document are considered statistically independent. As a result, a document can be represented as a product of term probabilities. The model assumes that terms that are not appearing in the query have equal frequencies in relevant and non-relevant documents. The BM25 scoring function is:

$$\mathrm{sim}(d_i \mid q_i) = \sum_{t \in q_i} \mathrm{IDF}\left(t\right) \cdot \frac{tf\left(t, d_i\right)\left(k_1 + 1\right)}{tf\left(t, d_i\right) + k_1\left(1 - b + b \cdot \frac{|d_i|}{\mathrm{avgdl}}\right)} \tag{2.1}$$

To measure a term's informativeness and estimate a document's relevance, this model uses the occurrences of individual query terms in a document (term-frequency) and in the whole collection (inverse-term frequency). Although this is a well-performing and popular IR model, it ignores the inter-relationship between the query terms that appeared in a document.

In Equation 2.1 k_1 and b are free parameters related to the query and the collection that are often tuned on a training dataset. $tf(t, d)$ is the frequency of the term t in the document d_i. Also, $|d|$ is the length of the document d_i measured in words, and *avgdl* is the average document length in the text collection. Finally, the $\mathrm{IDF}(t)$ is computed as:

$$\mathrm{IDF}\left(t\right) = \ln\left(\frac{N - n_t + 0.5}{n_t + 0.5} + 1\right) \tag{2.2}$$

where N is the total number of documents in the collection, and n_t is the number of documents containing t. Additionally, a field-based variant of the BM25 model, namely BM25-Field, considers documents as comprising multiple fields, such as title, body, and anchor texts [Robertson et al., 2004].

In our research we also leverage another family of models, namely the Divergence From Randomness (DFR) IR models. In these models, different matching functions can be obtained from a combination of a randomness model with an information gain model, and a term frequency normalization approach [Amati and van Rijsbergen, 2002]. The hypothesis is that the more a term's document frequency diverges from its collection frequency, the more information is carried by this term in the document. In this framework, three components should be considered; a basic randomness model, a first normalization, and a normalization of the term frequencies.

$$\mathrm{sim}(d_i \mid q_i) = \sum_{t \in q} qtw \cdot w_{d_i,t} \tag{2.3}$$

In Equation 2.3 $w_{d,t}$ is the weight of the term t in document d_i, and qtw is the query term weight given by qtf/qtf_{max}; qtf is the query term frequency and qtf_{max}

is the maximum query term frequency among the query terms.

Then, $w_{d,t}$ is calculated based on two different probability distributions (refer to Equation 2.4); Prob_1 measures the information content of t in d_i, specifically, the amount of information is given by $-\log_2 \text{Prob}_1$. In addition, Prob_2 measures the term's information gain for the set of documents it occurs in. In this case, the less the term is expected in the document, the more the amount of information is gained from this term.

$$w_{d,t} = (1 - \text{Prob}_2)\,(-\log_2 \text{Prob}_1) \tag{2.4}$$

Finally, concerning the term frequency normalization, the authors proposed two approaches; the first considers a uniform distribution of the term frequency. The other assumes that the term frequency density is inversely related to the length.

All in all, different models will occur based on the basic model used to calculate Prob_1, e.g. the Poisson model, when combined with a different approach for the first normalization, e.g. Laplace, to calculate the Prob_2, and different term frequency normalization, e.g. the second approach. Putting all together, the formula of the PL2 model is the following.

$$w(t,d) = \frac{tfn^2\left(12\log\left(\frac{tfn}{\lambda}\right) - 12\right) + tfn(6\cdot\log(tfn) + 12(\lambda + 0.92)) + 1}{12\cdot tfn(tfn+1)\log(2)} \tag{2.5}$$

In Equation 2.5 tfn is the normalized term frequency, and λ is equal to F/N, where F is the frequency of t in the whole collection. Under specific assumptions, the DFR model can explain the BM25 ranking formula without requiring the tuning of parameters b and k_1.

Another category of retrieval models uses a set of feedback documents to create language models [Zhai and Lafferty, 2001, Lavrenko and Croft, 2001]. Zhai and Lafferty [2001] use a set of feedback documents to re-estimate the query language model. To this aim, two methods have been proposed that update the initial query model by linear interpolation. The first method estimates the query topic using a feedback document by calculating the maximum likelihood or the regularized maximum likelihood. The second is based on minimizing the Kullback-Leibler divergence between the query language model and the document language model created from the feedback set. The evaluation procedure proved the superiority of the second approach in terms of its efficiency.

Lavrenko and Croft [2001] proposed a relevance model that uses a set of top ranked documents returned from the initial retrieval, denoted as R. Here, a relevance model is formally defined as the probability of observing a term in R, $P(t \mid R)$. It was assumed that both document and query terms were independently sampled from R. In the first introduced approach, namely RM1, each document was weighted based on its query likelihood and the probability of a term is averaged over every document language model. The second method, RM2, was based on the assumption that query terms are associated with document terms. As a result, relevant documents containing query words can be used to compute the association of the their words with the query terms. In both RM1 and RM2, Dirichlet Prior method was used to smooth the language model of each $d \in R$. One of the best performing model has become known as RM3 [Jaleel et al., 2004]. Specifically, this method interpolates the terms selected by RM1 with the original query, instead of using them directly. The final query is used in the same way as in RM1 to produce the final ranking.

Our research employs BM25 as a probabilistic retrieval model and DFR as another statistical model. These are integrated with RM3 to implement a pseudo-relevance feedback mechanism. For readers interested in a more extensive examination of different information retrieval models designed for estimating topical relevance, we direct you to the relevant literature [Schütze et al., 2008, Mitra et al., 2018].

2.1.3 Evaluation of Information Retrieval Systems

The initial development of IR systems almost instantly highlighted the necessity for having evaluation protocols in place for their appraisal. The evaluation assessment spectrum of IR systems is expansive, ranging from a total system-centric focus to a human-centric focus [Joho, 2011]. In our research we employ a system-oriented evaluation protocol based on the Cranfield paradigm [Cleverdon, 1970], that leverages *benchmark collections*. Benchmark collections offer the possibility for fair system performance comparison and, in this evaluation protocol, consist of three crucial components.

Documents. A collection of documents, also referred to as corpus.

Topics. A set of topics or queries which are a surrogate for real information needs. These topics can be derived by analyzing search logs associated with a search task and domain, or by observing real users [Soboroff, 2017]. Depending on the area of focus, queries could be just a few keywords, like in ad hoc searches, or they could be an entire document, which is common in legal searches.

Relevance Judgments. Also referred to as relevance assessments, ground-truth, labels, or *qrels*. These are (q, d, r) triples that assess if a document d is relevant r with respect to the query q. Commonly, human assessors carry out these annotations, making the relevance annotation process both time-consuming and expensive, especially when the annotators are domain experts (e.g. clinicians). Relevance assessments can be binary or graded-scale, and commonly the higher the value the more relevant the document to the query. Most commonly relevance assessments measure only topical relevance. Nonetheless, there are collections that assess multidimensional relevance, i.e. relevance is grounded with respect to multiple relevance factors. Finally, as Belkin et al. [2009] propose, each document within the collection could be evaluated for its utility in fulfilling the user's task, essentially incorporating a "usefulness" (utility) judgment.

Based on the search task being addressed, the size of a benchmark collection may differ. Nonetheless, it is advisable for a benchmark collection to comprise over ten documents and exceed 250 queries, as suggested by Spark-Jones [1975]. Having stated that, obtaining complete relevance judgments, i.e. for each query-document pair, for large benchmark collections is almost never feasible. As a result, relevance judgments are obtained using pooling, where the human assessors annotate only the top documents (without repetition) retrieved by the many retrieval systems [Buckley and Voorhees, 2004]. However, the exploration of pooling approaches remains an open research issue.

Due to the pooling process, a novel retrieval approach exhibiting significant methodological deviation from those contributing to the original pool might retrieve documents that have not been assessed by human annotators. In such instances, the prevalent approach is to regard these documents as non-relevant; under this scenario, the new system tends to underestimate its performance. In contrast, these documents can be considered relevant and overestimate the new system's retrieval performance; this is not a common practice in IR. Lastly, the evaluation of the new system can be conducted based on condensed measures as suggested by

Sakai [2007], serving as a method to address retrieved yet unjudged documents. In condensed evaluation, retrieved but unjudged documents are removed from the ranking before the estimation of a measure.

2.1.3.1 Evaluation Measures

The necessity for evaluation measures emerges from the objective of enhancing the retrieval performance of systems to accurately and efficiently address users' information needs. Evaluation measures are used to quantify the effectiveness of IR systems by utilizing relevance judgments. The choice of evaluation measures largely depends on the requirements of the given search task. Different search tasks may prioritize different retrieval aspects, such as precision, recall, or the quality of the ranking of retrieved documents. For instance, a search task that aims to provide a comprehensive set of relevant documents might prioritize recall, while a task focused on retrieving the most relevant documents at the top positions might prioritize measures like precision at k or Normalized Discounted Cumulative Gain (nDCG). As a result, choosing the proper set of evaluation measures is crucial as it directly impacts the understanding and the subsequent improvement of the retrieval performance of an IR system.

In this section, we briefly introduce the measures employed to evaluate the retrieval performance achieved by the proposed approaches. For a more comprehensive description of various evaluation measures, please refer to Manning et al. [2008] and Mitra and Craswell [2018].

Precision. Precision is order-unaware measure that is used to estimate a system's ability to retrieve information items relevant to a query. It is estimated as the proportion of relevant documents retrieved out of the total retrieved documents. It is defined as:

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of retrieved documents}}$$

In computing precision, multi-graded relevance assessments should be converted to binary values utilizing a specified relevance threshold. It is commonplace for precision to be estimated at a pre-determined cut-off of the rank k representing the ratio of relevant items found within the top-k retrieved results.

Recall. Recall measures a system's ability to retrieve all documents that are relevant to a query. It is estimated as the proportion of relevant documents retrieved out of the total relevant documents in the collection. It is defined as:

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}}$$

Similarly to precision, it is an order-unaware measure that relies on binary relevance judgments, often measured at a pre-determined cut-off k.

R-precision. The calculation of this metric necessitates the identification of the total number of relevant documents, R, corresponding to a specific query. It is essentially calculated as precision at k, where $k = R$.

$$R\text{-precision} = \frac{\text{Number of relevant documents retrieved in the top R positions}}{R}$$

R-Precision serves as a unique measure because it is equal to both the precision at the R-th position and the recall at the R-th position when a system retrieves exactly R documents.

Bpref. Bpref is a preference-oriented measure emphasizing the relative ranking of relevant documents over non-relevant ones. It is developed to be robust to the challenges posed by incomplete relevance assessments. The formula for $bpref$ is given by:

$$bpref = \frac{1}{R} \sum_{r=1}^{R} \left(1 - \frac{|n \text{ ranked higher than } r|}{R}\right)$$

In this equation, R represents the total number of relevant documents for a particular query. The term $|n$ ranked higher than $r|$ denotes the count of non-relevant documents that appear higher in the ranking than each relevant document r among the top R retrieved results.

Reciprocal Rank. The Reciprocal Rank is heavily influenced by the position of the first relevant item. It is estimated based on the reciprocal of the rank at which the first relevant document is retrieved. If the first relevant item is found in a low position in the ranking, the reciprocal rank score will be low.

It is defined as:

$$\text{Reciprocal Rank} = \frac{1}{\text{Rank of first relevant document}}$$

Since it only considers the position of a single relevant document, it may not be suitable for evaluating a system's performance in search tasks where a user will need to assess more than one relevant result.

nDCG (normalized Discounted Cumulative Gain). In contrast to the previously presented measures, this one is particularly designed for graded relevance assessments. To analyze it, we will focus on its three constituent components.

From Cumulative Gain (CG), one can derive Discounted Cumulative Gain (DCG) by introducing a logarithmic discounting element to account for the position of each information item, acknowledging that items retrieved earlier are more valuable to the user. The formula transitions from:

$$\text{CG@}k = \sum_{i=1}^{k} \text{rel}_i \text{ to DCG@}k = \sum_{i=1}^{k} \frac{\text{rel}_i}{\log_2(i+1)},$$

where rel_i is the relevance score of item i, and k is related to rank positions. This change addresses the limitation of CG that it does not consider the position of retrieved items, which is crucial for user satisfaction and system effectiveness.

Normalized Discounted Cumulative Gain (nDCG) further refines DCG by normalizing it against a perfect ranking to ensure the values lie between 0 and 1, making comparisons across queries and systems fairer. The formula is:

$$\text{nDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k},$$

where IDCG@k is the Ideal DCG at position k, obtained by sorting all items by relevance in descending order. This progression from CG to DCG addresses the positional relevance, and from DCG to nDCG adds normalization, each step overcoming the limitations of the previous metric to provide a more accurate and fair evaluation of retrieval system performance.

2.2 Decision Theory: An Overview and Applications

Decision theory primarily focuses on methodologies for optimal decision-making among various alternatives [Berger, 2013]. It offers a robust framework for rational decision-making, mainly when the outcomes of a selection process are not entirely predictable [North, 1968]. Within this framework, the goal is often to identify the most advantageous alternative, especially in situations characterized by uncertainty, risk, or incomplete information.

Building on this foundational framework, the methodologies inherent in decision theory offer quantitative tools for evaluating and comparing various alternatives. Such capabilities make the field invaluable for a broad spectrum of practical applications where a robust framework for systematic decision-making under uncertainty is needed. For instance, in healthcare, Decision Theory is employed to design optimal treatment plans for chronic diseases like diabetes and cancer, hospital resource allocation, or assess service quality in hospitals [Mardani et al., 2019]. In economics and business, decision-making is essential in applications related to optimizing investment portfolios and ranking banking performance, among others. It helps firms evaluate the risks and rewards associated with different business strategies, guiding them to make more informed decisions [Zavadskas and Turskis, 2011].

Decision Theory provides a universal and rigorous foundation for dealing with various decision problems across multiple domains. Its methodologies are beneficial for quantifying and analyzing uncertainties and trade-offs, often inherent in practical scenarios. The applications in healthcare and economics exemplify the theory's breadth and depth, demonstrating its crucial role in guiding individual and organizational decision-making processes.

2.2.1 Multi-criteria Decision-Making

Multi-criteria Decision-Making is a branch of decision-making that encompasses Multiple Objective Decision-Making (MODM) and Multiple Attribute Decision-Making (MADM) [Triantaphyllou, 2000, Alinezhad and Khalili, 2019]. The decision space in MODM is continuous, while MADM concentrates on problems with discrete decision spaces, where the set of decision alternatives is predetermined. Therefore, MADM has been widely used to solve problems in which a decision-maker aims to choose among different alternatives (set of possible solutions) those that better fulfill her/his preferences based on the evaluation of a set of predefined attributes (or criteria) [Triantaphyllou, 2000, Alinezhad and Khalili, 2019]. In the literature, several methods to address MADM problems have been proposed, which ground on some common notions but differ in their mathematical formalization [Aruldoss et al., 2013, Alinezhad and Khalili, 2019]. Independently of the employed method, a MADM problem is formalized by a finite set of m alternatives, denoted by $A = \{a_1, a_2, \cdots, a_m\}$, where each alternative is evaluated according to a finite set of n criteria, $C = \{c_1, c_2, \cdots, c_n\}$. Moreover, each criterion may be associated with an importance weight w_i; in the approaches we considered in this paper, $w_i \in [0, 1]$ and, $\sum_{i=1}^{n} w_i = 1$.

In addition, several MADM methods allow a particular objective to be assigned to each criterion. An objective is something to be pursued to its fullest, and it indicates the desired direction of change [Hwang and Yoon, 1981]. A criterion can be associated with either a positive or negative objective, based on the decision-maker's preference; i.e these criteria have either a positive or a negative effect in the decision process. Criteria that are associated with positive objectives are called beneficial criteria, while criteria associated with negative objectives are called non-beneficial. To clarify the notions of beneficial and non-beneficial criteria, let us consider a simple decision-making scenario. In this case, an individual (decision-maker) seeks to purchase a camera with the lowest possible cost (criterion 1) and weight (criterion 2) while simultaneously desiring the best possible battery life (criterion 3). Consequently, the cost and weight criteria are non-beneficial, as the decision-maker aims for the most affordable and lightweight option, whereas battery life is a beneficial criterion.

Usually, the information related to a MADM problem is represented in a decision matrix, $M_{m \times n}$, as depicted in Table 2.1, where each element x_{ij} represents the degree to which an alternative a_i satisfies a criterion c_j. A x_{ij} value is called performance score and it is calculated by an evaluation function. The $M_{m \times n}$ decision matrix, along with the criteria weights, objectives, and evaluation functions, are usually the fundamental inputs for a MADM problem. To solve a MADM problem and rank the alternatives, one has to calculate a global performance score by aggregating, for each alternative a_i, the performance scores computed for that alternative by

Table 2.1: An example of an $M_{m \times n}$ decision matrix.

	c_1	c_2	$\cdots$	c_n
a_1	x_{11}	x_{12}	$\cdots$	x_{1n}
a_2	x_{21}	x_{22}	$\cdots$	x_{2n}
$\cdots$	$\cdots$	$\cdots$	x_{ij}	$\cdots$
a_m	x_{m1}	x_{m2}	$\cdots$	x_{mn}

using an appropriate decision-making method, e.g. MADM methods. The selection of an appropriate method is usually based on the characteristics of the considered problem. In fact, over the years, various MADM methods have been proposed, which, depending on their properties, have been classified into various categories, such as, scoring-based, distance-based, compromising, and outranking methods, among others [Hwang and Yoon, 1981, Triantaphyllou, 2000, Alinezhad and Khalili, 2019]. In our research, we focus on a category of methods that allow the association of objectives with the criteria and can be employed for computationally complex problems, such as multidimensional document ranking, which is the aim of our work. In the following Section, the four considered MADM methods are presented in detail.

2.2.2 Multi-attribute Decision-Making Methods

In our research we exploit four established and widely used MADM methods, namely the Weighted Sum Model (WSM) [MacCrimmon, 1968], the Complex Proportional Assessment method (COPRAS) [Zavadskas et al., 1994], the Technique of Order Preference by Similarity to the Ideal Solution (TOPSIS) [Hwang and Yoon, 1981], and the VIseKriterijumska Optimizacija I Kompromisno Resenje method (VIKOR) [Opricovic and Tzeng, 2004]. A common characteristic of these four methods is their low computational complexity, making them suitable to be used in our applicative context, i.e. document ranking, where a huge quantity of items must be managed. It is worth noting that other MADM methods are available, for instance outranking MADM methods like PROMETHEE [Brans et al., 1986]. However, outranking methods tend to be less computationally efficient than the ones considered in this work due to need of performing pairwise comparisons and optimization operations [Calders and Van Assche, 2018]. As a result, these methods can not be efficiently employed for large-scale problems, such as document ranking.

Of the considered methods, WSM has been categorized as scoring or utility-based MADM method, as it selects the alternative that has the highest score, i.e the maximum estimated utility [Hwang and Yoon, 1981, Penadés-Plà et al., 2016]. Similarly, COPRAS, which has been introduced as an extension of WSM, falls into the same category. Regarding TOPSIS and VIKOR, these methods have been categorized as compromising MADM methods [Tzeng and Huang, 2011, Alinezhad and Khalili, 2019].

The following sections describe the mathematical properties and assumptions at the basis of the considered MADM methods.

2.2.3 Scoring-based Methods

Scoring-based MADM methods are the simplest methods to assess the overall performance of the considered alternatives. Both WSM and COPRAS methods are compensatory (i.e. the under-satisfaction of a criterion is compensated by the over-satisfaction of other criteria). Also, these methods assume that the considered criteria are independent, and they allow for objectives to be associated with the criteria. The main difference between them is that the COPRAS method allows to consider both beneficial and non-beneficial criteria, while WSM is designed only for modeling beneficial criteria. As a result, in the WSM method, the non-beneficial criteria must be converted to beneficial ones.

2.2.3.1 Weighted Sum Model (WSM)

Given a decision matrix as the one presented in Table 2.1, a set of weights associated with the criteria, and assuming that all criteria are beneficial criteria, the global performance score of an alternative, a_i, is estimated by employing a weighted sum as an aggregation function, as follows:

$$Q_i = \sum_{j=1}^{n} w_j x_{ij}, \text{where } w_j \text{ is the weight associated with the } c_j \text{ criterion.} \quad (2.6)$$

As mentioned above, this method accounts only for beneficial criteria, and therefore non-beneficial criteria must be properly expressed as beneficial ones, e.g. price can be either evaluated as cheap or expensive. In addition, all x_{ij} values must be expressed in the same unit by, for instance, normalizing the values of the decision matrix. Finally, the alternatives are ranked in descending order based on the obtained Q_i scores.

2.2.3.2 Complex Proportional Assessment (COPRAS)

Given a decision matrix as the one presented in Table 2.1, a set of weights associated with the considered criteria, and considering the criteria either beneficial or non-beneficial, a global utility score of an alternative is calculated by the following steps:

Step 1: In this initial step, a weighted normalized decision matrix, $D_{m \times n}$, is obtained as follows:

$$d_{ij} = \frac{x_{ij} w_j}{\sum_{i=1}^{m} x_{ij}} \tag{2.7}$$

In Equation 2.7, $i = 1, \ldots, m$ is the number of alternatives, and $j = 1, \ldots, n$ is the number of criteria.

By applying Equation 2.7, the sum of the d_{ij} values of each criterion, is equal to the assigned weight of that criterion. That means that the value of the weight is proportionally distributed among all alternatives, based on their performance score x_{ij}.

Step 2: In the second step the sums of the weighted normalized values, d_{ij} are computed, for both the beneficial criteria, Eq. 2.8a, and the non-beneficial criteria, Eq. 2.8b, for each alternative a_i, as follows:

$$S_{i^+} = \sum_{j=1}^{\kappa} d_{ij}, \text{where } \kappa \text{ indicates the number of beneficial criteria.} \tag{2.8a}$$

$$S_{i^-} = \sum_{j=\kappa+1}^{g} d_{ij}, \text{where } g = n - \kappa \text{ indicates the non-beneficial criteria.} \tag{2.8b}$$

These two values express the degrees to which each alternative attains the problem's beneficial and non-beneficial constraints.

Step 3: Finally, the global performance score for each alternative, a_p, is calculated using Equation 2.9:

$$Q_p = S_{i^+} + \frac{\min_i(S_{i^-}) \sum_{i=1}^{m} S_{i^-}}{S_{i^-} \sum_{i=1}^{m} \frac{\min_i(S_{i^-})}{S_{i^-}}} \tag{2.9}$$

In Equation 2.9, $\min_i(S_{i^-})$ is the minimum S_{i^-} value, across all alternatives.

Ultimately, the alternatives are ranked in descending order on the basis of their Q_i scores.

2.2.4 Compromising Methods

Compromising MADM methods are based on the notion of compromise solution, established by Yu [1973] and by Zeleny [1982]. A compromise solution can be defined as a feasible solution that is the closest to the ideal solution, where, in this context, compromise means an agreement established by mutual concessions. The term "ideal solution" refers to a hypothetical solution, i.e an alternative whose properties fully meet the problem's requirements.

Both TOPSIS and VIKOR methods are compensatory and allow for weights and objectives to be associated with the considered criteria. In particular, the VIKOR method is based on an aggregation function representing "closeness to the ideal solution", while the TOPSIS method introduces two reference points, representing a positive-ideal solution (PIS) and, additionally, a negative-ideal solution (NIS). An extensive comparison between them has been conducted by Opricovic and Tzeng [2004] and by Shekhovtsov and Salabun [2020], while their mathematical properties and main computational steps are analyzed below.

2.2.4.1 Technique of Order Preference by Similarity to the Ideal Solution (TOPSIS)

The TOPSIS method assumes that each criterion is monotonically increasing or decreasing an alternative's utility. Given a decision matrix, a set of weights, and considering that a criterion can be either beneficial or non-beneficial, the computational steps of TOPSIS are described as follows:

Step 1: creation of the normalized decision matrix, $R_{m\times n}$, by Equation 2.10.

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}}, i = 1, \ldots, m; j = 1, \ldots, n \tag{2.10}$$

Step 2: In this step, the weighted normalized decision matrix, $D_{m\times n}$, is obtained as follows:

$$d_{ij} = w_j r_{ij}, i = 1, \ldots, m; j = 1, \ldots, n \tag{2.11}$$

Step 3: The positive-ideal and negative-ideal solutions are determined and expressed by two distinct vectors, V^+ and V^-, respectively. The positive-ideal solution, Eq. 2.12a, maximizes the set of beneficial criteria, k, and minimizes the set of non-beneficial criteria g, where $g = n - k$ and n is the total number of criteria. In detail, for a criterion l, if it is a beneficial criterion, i.e. $l \in k$, $v_l^+ = max(d_{il}),\ i = 1, \ldots, m$; if l is a non-beneficial criterion, i.e. $l \in g$, then $v_l^+ = min(d_{il}),\ i = 1, \ldots, m$.

In contrast, the negative-ideal solution, Eq. 2.12b, minimizes the set of beneficial criteria, k, and maximizes the set of non-beneficial criteria, g. Specifically, for a criterion l, if it is a beneficial criterion, i.e. $l \in k$, $v_l^- = min(d_{il}),\ i = 1, \ldots, m$; if l is a non-beneficial criterion, i.e. $l \in g$, then $v_l^- = max(d_{il}),\ i = 1, \ldots, m$.

$$V^+ = (\mathrm{v}_1^+, \mathrm{v}_2^+, \ldots, \mathrm{v}_\mathrm{n}^+) \tag{2.12a}$$
$$V^- = (\mathrm{v}_1^-, \mathrm{v}_2^-, \ldots, \mathrm{v}_\mathrm{n}^-) \tag{2.12b}$$

Step 4: For each alternative, the euclidean distance of the vector representing the alternative from the vectors of both the positive (S_{i^+}) and the negative (S_{i^-}) ideal solutions is computed, as follows:

$$S_{i^+} = \left[\sum_{j=1}^{n} (d_{ij} - V_j^+)^2\right]^{0.5} \tag{2.13a}$$
$$S_{i^-} = \left[\sum_{j=1}^{n} (d_{ij} - V_j^-)^2\right]^{0.5} \tag{2.13b}$$

Step 5: The relative closeness of each alternative to the ideal solution is computed as the global performance score of the alternative (Eq. 2.14).

$$Q_i = \frac{S_{i^-}}{S_{i^+} + S_{i^-}}, 0 \leq Q_i \leq 1 \tag{2.14}$$

Finally, the alternatives are ranked in decreasing order of their computed Q_i scores.

2.2.4.2 VIseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR)

The VIKOR method has been proposed to solve decision problems with conflicting and non-commensurable criteria, assuming that compromise is acceptable for conflict resolution. The alternatives are evaluated based on their distance from a positive-ideal solution, by employing the $L_p - metric$ (Eq. 2.16) proposed by Yu [1973]. In detail, given a decision matrix, a set of weights and objectives associated with each criterion, the VIKOR procedure consists of the following steps:

Step 1: For each criterion $j = 1, \ldots, n$, its best (X_{j^+}) and worst (X_{j^-}) performance scores are obtained as follows:

$$X_{j^+} = \max_j x_{ij}, \quad X_{j^-} = \min_j x_{ij}, \text{ if the } j_{th} \text{ is a beneficial criterion.} \tag{2.15a}$$

$$X_{j^+} = \min_j x_{ij}, \quad X_{j^-} = \max_j x_{ij}, \text{ if the } j_{th} \text{ is a non-beneficial criterion.} \tag{2.15b}$$

Step 2: This step involves the aggregation of the performance scores, x_{ij}, of each alternative, a_i.

By using the $L_p - metric$, and by setting different p values in Equation 2.16 two distinct distance measures are derived. In particular, for $p = 1$, Equation 2.16 becomes a weighted and normalized Manhattan distance, denoted as S_i (Eq. 2.17). The $\min_i(S_i)$, across all alternatives, represents a maximum group utility ("majority" rule).

$$L_{P,i} = \left\{ \sum_{j=1}^{n} \left[w_j \left(X_{j^+} - x_{ij} \right) / \left(X_{j^+} - X_{j^-} \right) \right]^p \right\}^{1/p}, \quad 1 \leqslant p \leqslant \infty. \tag{2.16}$$

$$S_i = \sum_{j=1}^{n} \left[w_j \left(X_{j^+} - x_{ij} \right) / \left(X_{j^+} - X_{j^-} \right) \right] \tag{2.17}$$

For $p = \infty$, Equation 2.16 estimates $L_{\infty,i}$, i.e a Chebyshev distance, denoted as R_i (Eq. 2.18). The $\min_i(R_i)$, across all alternatives, represents a minimum individual regret.

$$R_i = \max_j \left[w_j \left(X_{j^+} - x_{ij} \right) / \left(X_{j^+} - X_{j^-} \right) \right] \tag{2.18}$$

In this method, both the S_i and R_i can be used to obtain two distinct rankings of the alternatives. However, the global performance scores produced by the above two equations is combined to produce an overall performance score that weights the “strategy of group utility” for each alternative using a weighted aggregation approach presented in Equation 2.19.

$$Q_i = \nu \frac{S_i - S_+}{S_- - S_+} + (1 - \nu) \frac{R_i - R_+}{R_- - R_+}, \text{where, } \nu \text{ is a balancing parameter and} \tag{2.19}$$

$$S_+ = \min_i(S_i) \qquad S_- = \max_i(S_i) \tag{2.20a}$$

$$R_+ = \min_i(R_i) \qquad R_- = \max_i(R_i) \tag{2.20b}$$

Based on its values, ν can either represent an optimistic ($\nu > .5$), pessimistic ($\nu < .5$) or neutral ($\nu = .5$) viewpoints. In the literature, the VIKOR method estimates the distance from a positive-ideal solution; therefore, the best alternative is the one with the minimum Q_i score.

The original VIKOR method involves further steps dedicated to the proposal of the compromise solution or a set of compromise solutions, determination of the weight stability intervals, and trade-off analysis (please refer to the cited related work for further details [Papathanasiou and Ploskas, 2018]). In this paper, we are investigating whether these MADM methods can be used as aggregation mechanisms within DtMRF. Therefore, to that aim, similarly to Shekhovtsov and Salabun [2020], we only need the ranking provided by Equation 2.19.

2.3 Large Language Models: From Prompting to In-context Learning

Information retrieval systems commonly utilize methods derived from Natural Language Processing (NLP) to enhance the ranking of documents. A notable example of applying NLP techniques in information retrieval is query expansion. This approach extends user queries beyond their initial form by incorporating synonyms, related terms, and contextually relevant words. Through this expansion, the recall of relevant documents is significantly improved, showcasing the adaptability and effectiveness of NLP in refining the search process. Other common NLP approaches frequently employed in information retrieval systems include sentiment analysis, coreference resolution, and text summarization, each playing a distinct role in enhancing retrieval effectiveness and efficacy.

This dissertation employs NLP techniques to process domain-specific unstructured information, aiming at extracting essential information and formulating queries. Specifically, Large Language Models are employed to extract information from unstructured medical notes through specifically designed prompts. This section introduces the primary concepts associated with prompting and outlines the main techniques proposed in the literature.

Prompting is a technique that leverages language models to the aim of generating content by directly predicting the probability of text. In these models, the original input x is modified using a *template* into a textual string, i.e. *prompt* x', that contains some unfilled slots (often related to the desired prediction). Then, the language model is used to fill the unfilled information and outputs a final string x'', from which the final output y (desired prediction) can be derived [Liu et al., 2023a]. Considering the following example; the original input (i.e., text) is "A patient with [symptom] is diagnosed with [condition]," and a template is provided as "A patient with [ ] is diagnosed with [ ]." When used as a prompt, the modified input becomes "A patient with cough is diagnosed with [ ]." Applying a language model to fill in the unfilled information results in the final string: "A patient with cough is diagnosed with pneumonia." The desired prediction is derived from this final output as "Pneumonia." Based on the literature, there are two distinct types of prompts, namely *soft prompts* and *hard prompts*. Soft prompts are learned embeddings or vectors that guide the model towards generating desired responses [Lester et al., 2021]. Hard prompts refer to human-engineered textual inputs that provide context, instructions, or examples to guide the model toward generating an appropriate response. In this dissertation, the term *prompts* refers always to *hard prompts*, unless specified otherwise.

The development of LLMs like GPT-2 [Radford et al., 2019] and GPT-3 [Brown et al., 2020], among others, allowed the use of prompts (i.e. hard prompts) that contain task-related instructions or demonstrations (i.e. task-specific input-output example pairs). These prompts are provided as natural language in the LLMs during inference time, and the models are expected to complete the provided text by generating a likely textual completion. This process is referred to as "in-context learning" [Brown et al., 2020]. In-context learning mainly leverages three techniques; "few-shot learning", where the prompt contains a few demonstrative examples (often between 10 to 100), "zero-shot learning" where only a task description with no tasks related examples is provided, or "one-shot learning" where one single example along with the task description is given to the model [Brown et al., 2020].

The selection of the appropriate prompt for in-context learning is essential to the overall LLM's effectiveness in the considered task [Zhao et al., 2023, Perez et al., 2021, Reynolds and McDonell, 2021]. Due to that, various research works introduce new prompting strategies for more efficient hard prompt construction, such as chain-of-thought [Wei et al., 2022], least-to-most prompting [Zhou et al., 2022], instruction prompt tuning [Singhal et al., 2022], self-consistency [Wang et al., 2022], and chaining multiple LLM prompts together [Wu et al., 2022].

Other works have been focused on prompt tuning, i.e. soft prompt construction or propose hybrid prompting approaches [Lester et al., 2021, Nye et al., 2021, Keskar et al., 2019]. Other research endeavors focus on addressing multiple limitations of LLMs, such as enhancing their capacity for reasoning [Zhou et al., 2022, Creswell et al., 2022, Kojima et al., 2022].

Wei et al. [2022] introduce the Chain-of-Thought (CoT) method as a way to address the limitations of large language models in arithmetic, commonsense, and symbolic reasoning [Rae et al., 2021]. The CoT prompting technique is a form of few-shot prompting that includes an <input, chain-of-thought, output> triplet. The chain-of-thought component comprises a series of natural language reasoning steps, for instance, human-like thoughts of solving a mathematical problem, which guide the model to produce the desired output. By applying this approach to three different LLMs, the authors showed its effectiveness in enhancing performance across arithmetic, commonsense, and symbolic reasoning tasks. According to Zhou et al. [2022], chain-of-thought prompting performs poorly when the requested problem is more complex than those included as demonstrations in the prompt. To overcome this, they proposed Least-to-Most prompting, a two-step approach that simplifies complex problems into more manageable sub-problems. The first step, problem reduction, supplies the model with examples and a specific question to break them into sub-problems. In the second step, problem-solving, the model sequentially addresses these sub-problems using constant examples, previously answered sub-questions, and generated solutions for guidance. Occasionally, the two stages can be combined into a single-pass prompting. Experimental findings in symbolic manipulation, compositional generalization, and mathematical reasoning demonstrate that least-to-most prompting substantially surpasses both standard prompting and chain-of-thought prompting in performance. Self-consistency is another strategy that aims to improve the performance of chain-of-thought prompting [Wang et al., 2022]. The intuition behind this approach is that a complex reasoning problem often has multiple ways of thinking that lead to the same correct answer. Therefore, by considering diverse reasoning paths and focusing on the most consistent answer, the self-consistency method aims to enhance the model's ability to solve complex reasoning tasks. In this approach, the final answer is the one with the majority vote. The empirical evaluation conducted by the authors suggests that the self-consistency method significantly improves the performance over the chain-of-thought prompting on a range of popular arithmetic and commonsense reasoning benchmarks. Moreover, the method also outperforms the Least-to-Most Prompting approach on the arithmetic reasoning task, based on the results obtained using the GSM8K dataset [Cobbe et al., 2021]. Wu et al. [2022] introduced the concept of Chaining LLM steps together. In this prompting approach the output of one prompt becomes the input of the next one, thereby combining the benefits gained at each step. Through the utilization of Chaining, a complex problem is decomposed into various smaller sub-tasks, each associated with a separate prompt. A user study conducted by the authors showed that the implementation of Chaining not only resulted in improved task outcomes but also contributed to users' satisfaction, sense of control, collaboration, and enhanced transparency of the LLM-based system.

Lester et al. [2021] present prompt tuning as a simple and computationally efficient technique for adapting LLMs to particular downstream tasks. This method involves learning soft prompt vectors through back-propagation while keeping the rest of the LLM frozen. During the tuning process, a task-specific copy of the entire pre-trained model is created for each downstream task, and inference is conducted in separate batches. As a result, prompt tuning requires only a small task-specific prompt per task, enabling mixed-task inference using the original pre-trained model. The authors' experimental results indicate that adapting frozen pre-trained language models to downstream tasks using prompt tuning helps prevent overfitting to a specific domain. Following the research work mentioned above, Nye et al. [2021], introduce instruction prompt tuning. This method combines a soft prompt learned through prompt tuning with a task-specific human-engineered hard prompt. The authors evaluated the performance of their method in MultiMedQA multiple-choice datasets, and their approach surpasses prior state-of-the-art by 17%. Another model that allows for task-specific adaptation is the CTRL model introduced by Keskar et al. [2019]. It is trained with *control codes* that can be related to a domain, subdomain, entities, relationships between entities, dates, and task-specific behavior (e.g. question answering or translation). As a result, the text generation process during inference is easily controlled by its end users.

LLMs have already reached state-of-the-art (SoA) performance in various tasks, and selecting an appropriate prompt has a significant impact. As a result, a significant number of related works investigate techniques to improve prompt construction further and their effectiveness [Liu et al., 2022, Rubin et al., 2022, Shin et al., 2021], explore their robustness to permutations of the demonstrative examples [Zhao et al., 2021], their sensitivity to negations [Jang et al., 2022], and their ability to generalize across different LLMs [Rakotonirina et al., 2023].

To begin with, Liu et al. [2022] investigate the sensitivity of GPT-3's performance to the selection of in-context demonstrative examples. The authors propose KATE, a retrieval-based approach for prompt construction that, given a test query (i.e. required question to the model), selects semantically-similar examples to the query and uses them to construct the final prompt. Their findings suggest that this approach consistently outperforms random prompt selection on various NLP benchmarks, with notable gains observed in tasks such as table-to-text generation. Similarly, Rubin et al. [2022] found that retrieving semantically-similar examples to the query and adding them in the final prompt improves effectiveness on three sequence-to-sequence tasks that map utterances to meaning representations. Similar conclusions have been drawn by Shin et al. [2021]. This evidence suggests that using semantically-similar demonstrative examples to the final query, rather than randomly selected, is a better practice for prompt construction.

Zhao et al. [2021] show that GPT-3's few-shot learning could be unstable due to the selected prompt format, number of training examples, and example order. To address this issue, the authors introduced a contextual calibration procedure that significantly improves GPT-3 and GPT-2's accuracy and stability across various prompt choices. LLM's sensitivity to ordering a prompt's examples has also been investigated by Lu et al. [2022], where the authors showed that it significantly impacts effectiveness across various NLP tasks. Jang et al. [2022] explore LLMs ability to comprehend and respond in negated prompts. The authors experimented with several LMs and LLMs, e.g. GPT-3, InstructGPT [Ouyang et al., 2022], among others and their findings suggest that scaling LMs does not enhance their ability to understand negations. In certain situations, LLMs benefit from in-context learning to understand negation, while fine-tuning is effective in all scenarios. However, fine-tuning negatively impacts the performance of the original task. Despite these approaches, the LLMs used in the study still fell short compared to human performance. Another study explored the behavior of ChatGPT in terms of semantic, negation, and symmetric consistency [Jang and Lukasiewicz, 2023]. Semantic consistency implies that a model should make coherent decisions in contexts that have the same meaning. Symmetric consistency is a type of consistency that relies on symmetric inference, meaning that for a given function f, if $f(x, y)$ yields a result, then $f(y, x)$ should produce the same result. Finally, negation consistency revolves around the logical negation property. Their findings suggest that ChatGPT exhibits improved language understanding, particularly in negation expressions and antonyms, compared to other LMs. However, it displays self-contradictory behavior by frequently changing its predictions when presented with paraphrased inputs. Finally, ChatGPT tends to generate different outcomes when the order of input sentences is altered, violating symmetric consistency. The authors emphasize the significance of human inspection in AI-generated content, particularly for risk-sensitive applications, as revealed by their findings.

Rakotonirina et al. [2023] investigate whether prompts can generalize across different LMs and LLMs focusing on the slot filling NLP task. The authors experimented with manual, semi-manual, and automatic methods for prompt creation. Their empirical evaluation suggests that prompts are generally more stable across different sizes of the same model. They modified the AutoPrompt algorithm [Shin et al., 2020] so that one LM generates candidate prompts, and then a second model evaluates them and chooses the best one.

2.4 Multi-output Regression

Regression is a form of predictive modeling task that aims to estimate a numerical output based on one or more input variables. Multi-output regression models, also known as multi-variate or multi-target regression, aim to predict multiple output variables simultaneously based on one or more input variables [Borchani et al., 2015, Watt et al., 2020]. Unlike single-output regression where each input is mapped to a single output, multi-output regression maps each input to a vector of output variables. The general mathematical formulation for multi-output regression with a generalized function F can be written as:

$$\mathbf{Y} = F(\mathbf{x}; \Theta) + \epsilon \tag{2.21}$$

In Equation 2.21 $F(\mathbf{x}; \Theta)$ aims to approximate the relationship between the input variables $\mathbf{x}$ and the output variables $\mathbf{Y}$. The function F serves as the generalized model that can be a linear equation, a polynomial model, or a more complex models such as deep neural networks. Θ represents the set of parameters that define the function F. These parameters are adjusted during the model training process to minimize the error between the predicted and actual output variables. Usually, the choice of F depends on the complexity of the relationship between the input and output variables. $\mathbf{Y}$ represents multiple output variables, with each element corresponding to a different output variable that the model aims to predict. For instance, $\mathbf{Y} = [y_1, y_2, y_3]$ would represent three output variables y_1, y_2, and y_3. The model is not an exact representation of the underlying relationship, therefore possible unexplained factors are captured by the error term ϵ. ϵ is a vector of error terms, one for each output variable. It accounts for the model's limitations and the unexplained variance in the output variables. Besides solving directly a multi-output regression problem, other common approaches involve breaking it down into single-output independent regression problems, or sequential regression models [Borchani et al., 2015].

Independent Regression Models. Researchers have devised independent regression models to simplify the complexity of multi-output regression, which transform the multi-output problem into independent single-output problems, each solved using a conventional single-output regression algorithm. Following the earlier mathematical formulation, this approach employs separate models for each output variable as follows:

$$
\begin{aligned}
y_1 &= f_1(\mathbf{x}; \theta_1) + \epsilon_1 \\
y_2 &= f_2(\mathbf{x}; \theta_2) + \epsilon_2 \\
y_3 &= f_3(\mathbf{x}; \theta_3) + \epsilon_3
\end{aligned}
$$

Sequential Regression Models (Chaining). Sequential Regression Models, commonly known as chaining approaches, are a subset of problem transformation methods that predict each output variable sequentially, using the predicted values of previous output variables as additional inputs. The mathematical formulation for chained models using generalized functions f_1, f_2, f_3 is:

$$
\begin{aligned}
y_1 &= f_1(\mathbf{x}; \theta_1) + \epsilon_1 \\
y_2 &= f_2(\mathbf{x}, y_1; \theta_2) + \epsilon_2 \\
y_3 &= f_3(\mathbf{x}, y_1, y_2; \theta_3) + \epsilon_3
\end{aligned}
$$

The core components of training a regression model include defining the loss function, selecting an optimization algorithm, and executing the model training process.

Defining a Loss Function. The loss function quantifies the difference between a predicted $\mathbf{Y}$ and an actual $\hat{\mathbf{Y}}$ output for each data point and is crucial for a model's training. The Mean Squared Error (MSE) is most commonly used in regression tasks due to its ease of computation and differentiable properties.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{Y}_i - \hat{\mathbf{Y}}_i)^2 \tag{2.22}$$

MSE measures the average of the squares of the differences between the predicted and actual output values. Other alternatives include Mean Absolute Error (MAE) and Huber Loss, which might be more robust to outliers. However, the selection of a particular loss function relies on the characteristics of the problem [Chai and

Draxler, 2014].

Optimization Algorithm. Once the loss function is defined, the next step is to choose an optimization algorithm to adjust the model parameters to minimize this loss. Gradient Descent is the most commonly used optimization algorithm in regression tasks [Ruder, 2016]. More advanced optimization algorithms like Adam [Zhang, 2018] are also used, particularly in neural network-based regression models.

Training Process. After defining the loss function and choosing an optimization algorithm, the model is trained using a training dataset. The optimization algorithm iteratively adjusts the model parameters to minimize the loss function. This process can be straightforward for some models. In contrast, neural models rely on back-propagation, commonly used to update the weights during the training phase. The training process often involves multiple iterations, or epochs, through the training dataset until the loss converges to a minimum value. In regression models that utilize neural networks, an activation function for the output layer is generally only necessary if one aims to impose certain constraints on the output, such as bounding it between 0 and 1.

In our research, we formulate the objective of predicting importance weights corresponding to relevance factors as a multi-output regression problem, employing the previously discussed methodologies to address it. This formulation is incorporated into Neural-DtMRF, which utilizes a neural regression model.

Chapter 3

Multidimensional Relevance Estimation: A Systematic Literature Review

Our research introduces a framework for multidimensional relevance estimation in IR. Therefore, this chapter presents a systematic literature review we conducted aiming to enrich the understanding around this research field. Through our systematic review of 70 studies, we have categorized research based on domain specificity and the distinct relevance aspects employed for estimating multidimensional relevance. Moreover, we highlight the approaches used to aggregate scores related to these factors, and rank information items.

3.1 Introduction

Our survey systematically examines 70 studies that have proposed and experimentally evaluated multidimensional relevance models. We synthesize these studies based on the knowledge domains and search tasks, their employed relevance factors, and the utilized benchmark collections. The specific research questions tailored to our study are presented in Section 3.2. This review aims to aid the development of future multidimensional models by identifying current necessities and paving the way for future research.

The chapter is structured as follows. Section 3.2 outlines the systematic methodology employed for collecting and synthesizing literature studies. In Section 3.3, we present the outcomes of our synthesis, including key characteristics of the reviewed studies, such as their geographic and temporal distributions, among others. Additionally, we identify the knowledge domains in which relevance has been perceived and modeled based on multiple factors. We analyze the identified relevance factors based on their definitions and operationalizations, aiming to highlight their commonalities and differences across and within domains. Furthermore, Section 3.3.4 discusses the benchmark collections used to evaluate multidimensional relevance models in the included studies. Subsequently, Section 3.4 offers an in-depth discussion of our systematic literature examination findings, pointing to potential avenues for future research directions. Section 3.5 discusses our study's prospects and limitations, while Section 3.6 concludes our study.

3.2 Method

The main objective of this systematic review is the examination of studies that consider relevance a multidimensional notion, as described in Section 2.1.1. We will categorize these studies based on their applied knowledge domain (e.g. health, legal, academic) and the relevance factors they utilize. Additionally, we will analyze the methods employed to aggregate these relevance factors. Furthermore, we will group the different relevance factors used in the reviewed studies according their definitions and operationalization, i.e. how the authors estimated or measured these factors. We will compile a comprehensive list of benchmark collections that have been utilized in the reviewed studies. These benchmark collections will be characterized based on the annotated relevance factors, the knowledge domain, their size, and availability. Finally, we will provide an overview of various initiatives that offer shared tasks centered around multidimensional relevance. The ultimate goal of this systematic review is to shed light on the multidimensional nature of relevance and to highlight the various approaches and benchmark collections used to study this important concept across different knowledge domains. By doing so, we aim to contribute to a clearer understanding of multidimensional relevance and its practical and theoretical implications.

Following the methodological approach proposed by Cooper et al. [2019], the systematic review conducted in this study consists of the following steps. (1) Formulation of the research questions, (2) establishment and clarification of the inclusion and exclusion criteria associated with the selection of research papers, (3) development of a retrieval strategy (e.g. involved sources and databases, keywords), (4) proposal of a coding scheme for paper annotation, (5) synthesizing the findings to answer the research questions.

3.2.1 Step 1: Research Questions

This section introduces the research questions that guide our systematic review. By addressing these questions, we aim to gain valuable insights into how relevance is perceived, decomposed into several factors, and estimated in different knowledge domains. The answers to these questions will not only deepen our understanding of multidimensional relevance but also contribute to the advancement of research and practical applications within the domain of Information Retrieval. To this end, this systematic review seeks to answer the following research questions:

(RQ1) How is relevance conceptualized and operationalized as a multidimensional concept (as defined in Section 2) in the identified studies?

- (1.1) What are the different knowledge domains (e.g. health, legal, academic) in which, multidimensional relevance has been explored?
- (1.2) What are the relevance factors utilized by researchers in the reviewed studies?
- (1.3) What are the diverse approaches employed to aggregate relevance factors in the context of multidimensional relevance estimation?

(RQ2) How do authors define and operationalize relevance factors (i.e. estimate a score to be associated with them) in the reviewed studies?

(2.1) How have the relevance factors been defined within the studies incorporated in the review?

(2.2) What methodologies and techniques are used to operationalize the identified relevance factors?

(RQ3) Which benchmark collections have been used to estimate multidimensional relevance, and how are they characterized based on their annotated relevance factors, size, and availability?

3.2.2 Step 2: Inclusion and Exclusion Criteria

The aim of this step was to establish and evaluate the inclusion and exclusion criteria that were utilized to systematically select and reject articles for review. The development of the inclusion and exclusion criteria commenced by compiling criteria that align with the target study type: multidimensional relevance estimation in IR. Although the initial list of criteria was seen as provisional and subject to refinement throughout the review process (i.e. after processing 10% of total included articles), no further adaptations to the criteria were implemented. Table 3.1 presents a comprehensive list of the final criteria.

Table 3.1: List of selection criteria.

Inclusion/Exclusion Criteria
Including studies focused on text retrieval
Including empirical studies that utilize a minimum of two relevance factors
Including scholarly publications subject to peer-review
Including both full-length research articles and short papers
Excluding studies solely focused on operationalizing a relevance factor
Sources are confined to journals, conference proceedings, and workshops
No specific time frame
Studies must be written in English

This review exclusively included studies focusing on text retrieval systems (i.e.

document retrieval), as studies involving other types of information objects (e.g. audio, video) would significantly expand the scope of the study. This review encompassed empirical studies (i.e. use experimental methods) that utilized a minimum of two relevance factors for document ranking, with topical relevance being one of those factors. Consequently, we omitted studies that employed neural models for document re-ranking, as these studies rely solely on topical relevance signals. In this review, we excluded studies in which researchers solely operationalized a relevance factor, without utilizing it to estimate multidimensional relevance and perform document ranking (hereafter ranking). We applied this exclusion criterion since the studies primarily aimed to predict a single score for a relevance factor rather than estimate multidimensional relevance. Our review specifically investigated how relevance factors have been operationalized only when they were utilized for retrieval.

Table 3.2: List of evaluation initiatives.

Initiatives	
CLEF	Conference and Labs of the Evaluation Forum
TREC	Text Retrieval Conference
FIRE	Forum for Information Retrieval Evaluation
INEX	Initiative for the Evaluation of XML Retrieval
NTCIR	NII Testbeds and Community for Information Access Research

To ensure the selection of higher quality articles, the inclusion criteria were restricted to scholarly publications that had undergone peer-review. Consequently, sources were confined to journals, conference proceedings, and workshops, encompassing both full-length research articles and short papers. This criterion lead to the potential exclusion of essential initiatives' proceedings such as those mentioned in Table 3.2. Nonetheless, several of these papers were still included as they were later published in peer-reviewed journal or conferences. Moreover, our systematic review reports on benchmark collections that are often associated with the aforementioned initiatives, providing a reference point for interested readers. Ultimately, to capture the complete scope of relevant articles, a specific time frame was not imposed, and all studies included in the review were required to be written in English.

3.2.3 Step 3: Search Strategy and Paper Selection

We used the inclusion and exclusion criteria outlined previously to acquire publications on multidimensional relevance estimation in IR. These were obtained through searches across multiple research publication search engines and databases. The process of searching for potentially relevant articles for this review consists of the following steps, as shown in Figure 3.1.

Similarly to the systematic review conducted by McGregor et al. [2023], we initiated the search process by searching within the selected literature databases (journals and conferences/workshops) shown in Table 3.3.

To facilitate the database search we created the following query, *("multidimensional relevance" OR "relevance factors" OR "relevance dimensions" OR "relevance aspects" OR "multi aspect relevance") AND ("information retrieval").* For the majority of the resources, the search was refined to "title" and "abstract" search. However, in cases that this was not feasible, we conducted the search using the "full-text" option. To avoid missing relevant articles, we additionally conducted searches in Google Scholar, Springer Link, ACM Digital Library, IEEE Xplore, and Science Direct, similarly to previous studies [Liu, 2021, Vakkari, 2020]. These searches also utilize the same query, with slight modifications tailored to their specific requirements. We tried different combinations of the aforementioned keywords, aiming to cover most, if not all, of the relevant research for further analysis. Following the aforementioned search process, a total of 1,387 studies have been identified. Those articles have been manually screened by reviewing their title and abstracts to determine their relevance to this study. At this point, we were interested in reducing the initial document pool to include those focused on document retrieval and excluding those studies that solely estimate a score to be associated with a relevance factor without using it for ranking. As a result, a total of 134 studies have been selected for further examination. These studies have been evaluated based on the whole set of inclusion/exclusion criteria listed in Table 3.1, and a total of 62 studies have been identified as eligible for this study. The majority of the papers have been excluded because they were not focused on text retrieval. Finally, similarly to Liu [2021] and McGregor et al. [2023], we performed a forward and backward citation chaining on the final pool of the 62 eligible studies and, 8 additional studies were included for review.

Table 3.3: Sources examined in database and other searches.

Journals
Information Processing & Management (IP&M)
Journal of the Association for Information Science & Technology (JASIS&T)
International Journal on Digital Libraries
Information Retrieval Journal (IRJ)
Journal of Information Science
Journal of Documentation
ACM Transactions on Information Systems (TOIS)
Conferences/Workshops
ACM/IEEE Joint Conference on Digital Libraries (JCDL)
European Conference on Digital Libraries (ECDL)
European Conference on Information Retrieval (ECIR)
ACM International Conference on Information and Knowledge Management (CIKM)
Proceedings of the Association of Information Science & Technology (ASIS&T)
ACM Special Interest Group on Information Retrieval Conference (SIGIR)
ACM SIGIR Conference on Human Information Interaction & Retrieval (CHIIR)
Information Interaction in Context Conference (IIiX)
ACM International Conference on Web Search & Data Mining (WSDM)
International Conference on the Theory of Information Retrieval (ICTIR)
ACM Conference on Recommender Systems Conference (RecSys)
Other Sources
Google Scholar
Springer Link
ACM Digital Library
IEEE Xplore
Science Direct

3.2.4 Step 4 and 5: Coding Scheme and Paper Synthesis

The categories for coding and analysis were designed in accordance with the research questions (RQs) we aimed to address. The employed coding scheme consists of general information related to publication characteristics such as authors' affiliations, publication venues and year. The purpose was to provide insights into the distribution of research across different areas and over time. Aiming to address RQ1, we coded studies based on the knowledge domain exploited in their experimental evaluations, the employed relevance factors, and the exploited approach to aggregate the relevance scores and rank the documents. The identified aggregation

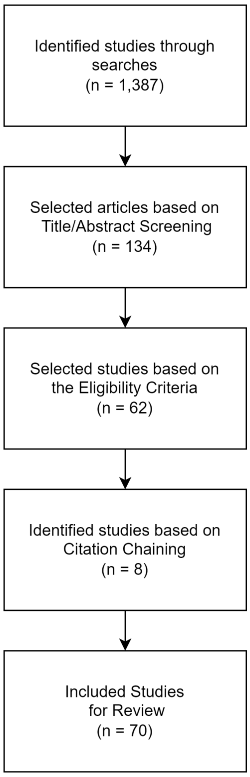

Figure 3.1: Overview of the followed search process.

approaches were categorized to data-driven, model-driven, or other (for those that did not clearly fit in the other categories). The second research question aimed at providing insights related to the investigated relevance factors. To that aim, we highlighted similarities and differences between the definitions and operationalizations of the identified relevance factors across studies and knowledge domains. By comparing and contrasting the identified studies based on how they exploit the associated relevance factors, we obtained a clearer understanding regarding conceptual and experimental differences. Finally, regarding the third research question, the included studies have been coded based on their employed benchmark collection, which have been further analyzed regarding their characteristics. As a result, we obtained insights regarding the available benchmark collections that can be used to investigate multidimensional relevance models.

Following the coding schema as described above, we were able to identify commonalities and differences regarding multidimensional relevance estimation, across knowledge domains and search tasks. Through this systematic review, we delved a better understanding of the limitations and potentials of exploiting relevance as

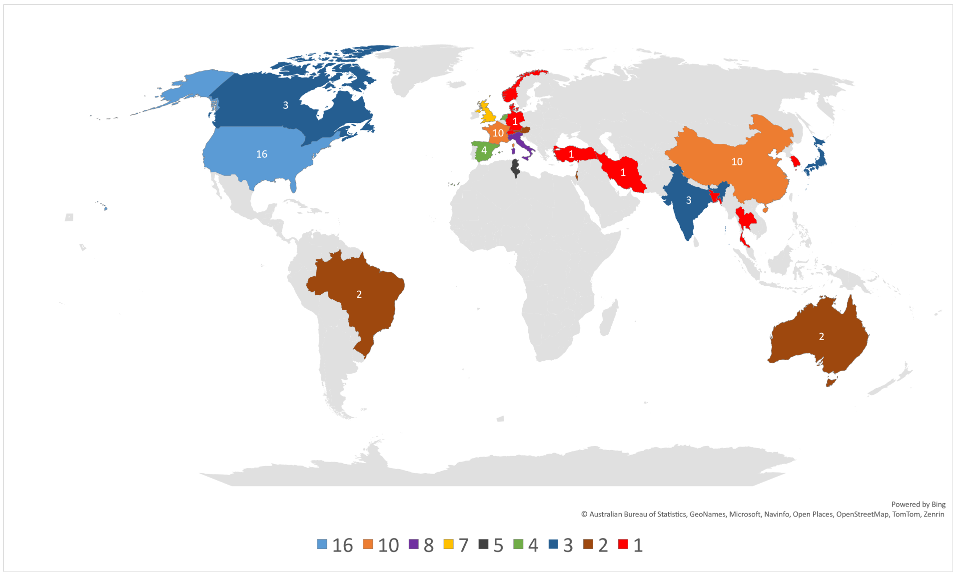

Figure 3.2: Map illustrating the number of publications on Multidimensional Relevance Estimation by country. The geographic location is determined by the authors' affiliations and not their nationalities.

multidimensional concept in IR. Our analysis allows to compare studies in terms of the aggregation methods, the application domains, and the relevance factors (definition, operationalization). Synthesizing them based on the application domain, we draw insights regarding the definition and operationalization of the employed relevance factors. Synthesizing them based on the relevance factors, we investigate how these factors are exploited across domains. Finally, by analyzing their datasets, we draw insights regarding their similarities and differences and we highlight future necessities.

3.3 Results

This section delves into the synthesis and comparative analysis conducted on the body of literature under review. The outcomes are in alignment with our predetermined coding scheme and the posed research questions. Our analysis provides a comprehensive review of the studies in question, laying a foundation for future research and exploration.

3.3.1 Overall Publication Characteristics

This section presents key characteristics of the publications under study. We explore a multifaceted view of the research landscape in multidimensional relevance estimation, by examining the publications based on: their geographical distribution; the collaborative efforts between industry and academia highlighting synergies; the diversity in types of publication venues; and the temporal distribution that offers insights into the evolution of research in this domain.

Central to our review, we identified 199 researchers who have significantly contributed to the literature on this subject. These researchers represent a wide spectrum of expertise, originating from varied academic and professional backgrounds. Our review reveals a diverse geographical distribution of research on multidimensional relevance estimation. A detailed representation of the number of papers per country, based on authors' affiliations, is provided in Figure 3.2. As illustrated in the figure, the USA leads in contributions with 16 studies, closely followed by China and France. Similarly, several European countries have shown significant contributions, with Italy, UK, France, Spain, and the Netherlands collectively accounting for 33 studies. Notably, Tunisia stands out in the North African region with 5 contributions, while Asia's presence is also marked by contributions from countries such as China, Japan, India, and South Korea. The global map illustrating this geographic distribution provides a comprehensive snapshot of the worldwide research landscape in the examined area, highlighting the strong collaboration among researchers.

A noteworthy observation from our review is the synergy between academia and industry, as shown in Figure 3.3.

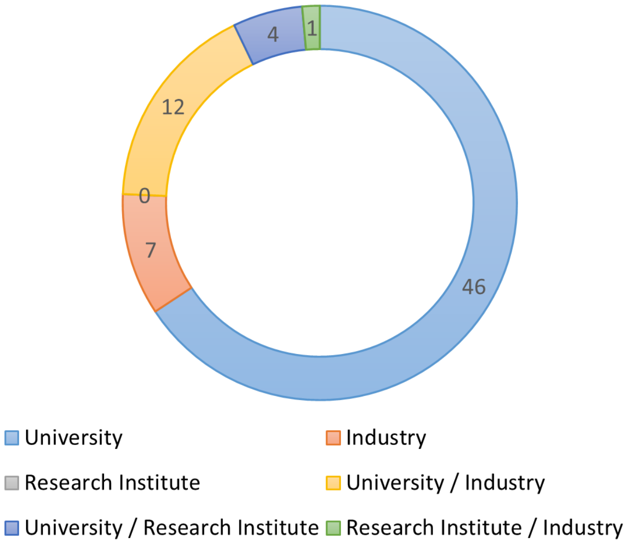

Figure 3.3: A representation of synergies between universities, research institutions, and industry in the studied literature.

We quantified the collaborations and found 12 of the included publications showcasing a partnership between academia and industry. A total of 7 publications is being authored by researchers working in industry. Several studies were conducted by major corporations in the field such as Microsoft [Craswell et al., 2005, Collins-Thompson et al., 2011], Google [Zhuang et al., 2021], Yahoo [Kang et al., 2012], and Amazon [Mandayam Comar and Sengamedu, 2017, Carmel et al., 2020, Yang et al., 2021], as well as other companies collaborating with universities to address information retrieval tasks in domain-specific search [Sasaki et al., 2016, Wiggers et al., 2023]. Such collaborations are indicative of the practical applications and real-world significance of estimating relevance by considering several factors that affect it under specific contextual situations.

Regarding the distribution of publication venues over time, this is illustrated in Figure 3.4. As we previously discussed, the idea of considering relevance as a multidimensional concept is rooted in the origins of information search systems [Saracevic, 2007]. Contributions by researchers such as Goffman and Newill [1966], Cooper [1971], Mizzaro [1998], among many others, lead to a shift towards recognizing its dynamic and multidimensional nature. Following this recognition, several researchers conducted user studies to identify contributing relevance factors, with key studies being from Barry and Schamber [1998], Cool et al. [1993], Xu

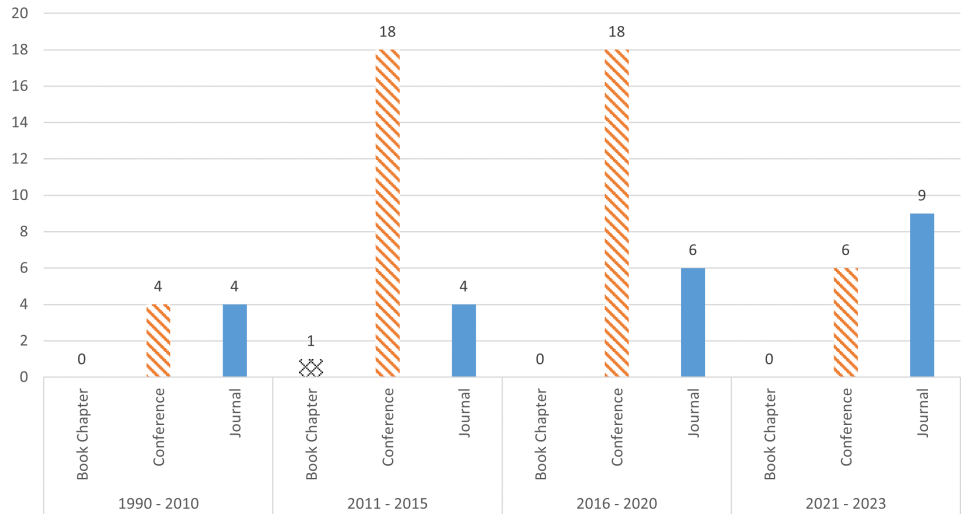

Figure 3.4: Time-based distribution of research studies across venue types.

and Chen [2006a], among others. Subsequently, experimental evaluations were pursued by multiple scholars [Brin and Page, 1998, Craswell et al., 2005, Ashoori and Lalmas, 2007, Sieg et al., 2007, Farah and Vanderpooten, 2008], with the most notable contribution that utilized multiple relevance signals for ranking is the integration of the PageRank algorithm in commercial web search [Brin and Page, 1998]. In the following years (2011-2020), we observe a consistent trend in publication output, with both the periods 2011-2015 and 2016-2020 showing nearly identical numbers of conference and journal publications. This suggests a stable and sustained research interest in the topic throughout the decade. From 2021-2023, there one can observe an upward trend in journal publications. However, this observation might not provide a full comparison with the previous years for two reasons: 1) the time span under consideration is shorter, and 2) several of the identified publications in 2023 have not been peer reviewed and have been excluded from our review.

Among the 26 identified conferences, the ACM SIGIR Conference on Research and Development in Information Retrieval stands out as a primary venue, hosting 10 out of the 70 surveyed papers, followed by the Conference on Information and Knowledge Management (CIKM) with 6 publications. There are 18 distinct Journals, from which the Journal of the Association for Information Science and Technology emerges as a leading venue with 3 publications, followed by journals such as Information Fusion and the Information Processing and Management

Journal that have 2 publications. In the subsequent sections, we delve further into our analysis, targeting the specific answers to our research questions.

3.3.2 How is relevance conceptualized and operationalized as a multidimensional concept?

In this section, we aim at providing insights related to the conceptualization of relevance across different knowledge domains. Our primary goal is to identify and describe these domains, and highlight particular search task in which relevance has been treated as a multidimensional notion. Following that, we mention the specific relevance factors that are utilized within each domain and search task. Finally, we classify the various methods researchers have used to combine relevance scores associated with distinct factors, to obtain an overall multidimensional relevance score.

3.3.2.1 What are the different knowledge domains in which a multidimensional notion of relevance has been explored?

Table 3.4 presents a detailed breakdown of studies conducted across diverse knowledge domains and search tasks, from which we have identified 18 domains. The observed domains span from academic and medical to web and social, with some emphasizing specific search tasks, like consumer health and biomedical article retrieval tasks within the medical field. Notably, while some domains have only one study, research areas like web search dominate with 25 studies, reflecting possible research emphasis and potential complexity of investigating multidimensional relevance in the other domains. Further result analysis, presented in Section 3.3.4, deepens our comprehension of the underlying reasons for the observed long-tailed distribution of the identified domains.

Having established the distribution of various knowledge domains, we now focus on each domain and highlight the specific retrieval tasks in the identified studies. In web search, 25 studies met our criteria and are incorporated into our systematic review, all of which developed models for multidimensional relevance. Specifically, Lioma et al. [2016] explore how the factuality and objectivity of documents relate to document relevance, and integrate them as query-independent features in a retrieval model. Undoubtedly, Page Rank is a fundamental feature integrated into commercial web search systems [Brin and Page, 1998]. Expanding on that, Craswell et al. [2005] implement sigmoid transformations on PageRank, URL Length, and ClickDistance and combine them with topical relevance signals such as BM25. Other scholars explore how external knowledge from knowledge graphs can be combined with topical relevance signals to improve retrieval performance [Rinaldi, 2009, Li et al., 2021]. Focusing on specific web search tasks, several studies propose retrieval models that integrate topicality with other relevance factors such as information freshness [Dai et al., 2011, Bambia and Faiz, 2015], content's quality [Bendersky et al., 2011], content's readability [Sasaki et al., 2016], source's popularity, recency, and reputation [Badache and Boughanem, 2014]. Other scholars proposed models to retrieve child-friendly content [Eickhoff et al., 2013a], information related to programming search tasks [Silva et al., 2019], and web tables [Shraga et al., 2020]. Several studies leverage user-related relevance factors for web retrieval, i.e. personalized web search [Sieg et al., 2007, Collins-Thompson et al., 2011, Sahraoui and Faiz, 2017, Li et al., 2017b, Uprety et al., 2018]. Moreover, research efforts have been made to tackle the challenge of obtaining a diverse set of retrieved documents, ensuring they address multiple query aspects while reducing redundancy (topic distillation) [Farah and Vanderpooten, 2008, van Doorn et al., 2016, Vargas et al., 2012, Shajalal et al., 2020]. Finally, several studies proposed frameworks that leverage multiple relevance signals for document ranking and use web search as an application domain [Komatsuda et al., 2016, Eickhoff and de Vries, 2014, Zhuang et al., 2021].

Within the medical domain, two distinct search tasks where relevance is interpreted as a multidimensional concept have been identified: the retrieval of biomedical articles [Znaidi et al., 2016, Xu et al., 2016, Alsulmi and Carterette, 2018, Qu et al., 2020, 2021] and consumer health search [van Doorn et al., 2016, Zhang et al., 2015, Palotti et al., 2019, Putri et al., 2021]. In addition, research endeavors prioritize retrieving health information that is topically relevant, credible, and reliable [Upadhyay et al., 2022, Fernández-Pichel et al., 2022]. Additional domains that have attracted the attention of researchers with respect to multidimensional relevance estimation include social and e-commerce searches. In social search, studies explore Twitter (now referred to as X Corp) search and integrate topical relevance with signals like recency, authority, trustworthiness [Jabeur et al., 2012, Ravikumar et al., 2013, Moulahi et al., 2014a]. Other studies focus on retrieving content related to events, disasters or opinions [Madisetty and Desarkar, 2022, Putri et al., 2020], or leverage social content to improve ranking [Tamine et al., 2011]. E-commerce has risen to significant prominence in recent years. In this domain, the notion of relevance is influenced by domain-specific factors that are related to products, temporal contextual information (referred to as *seasonality*), reviews, and users' intents, among others [Mandayam Comar and Sengamedu, 2017, Karmaker Santu et al., 2017, Feng et al., 2018, Carmel et al., 2020, Yang et al., 2021, Bassani and Pasi, 2021].

Research on multidimensional relevance estimation spans a variety of other domains, reflecting the diverse nature of information needs across different contexts. For example, in academic search, researchers such as Jomsri and Prangchumpol [2015], Arastoopoor [2018], Singh and Dave [2019] have put forth models incorporating recency alongside other domain-specific criteria. Meanwhile, math search is another domain where the complexity of relevance estimation necessitates the combination of multiple signals, as shown by Yan et al. [2022]. Blog post search involves the aggregation of signals related to a source authority or level of opinion [Eickhoff et al., 2013a, Gerani et al., 2012, Chenlo et al., 2015, Huang et al., 2018]. In newswire search, researchers have proposed models that leverage recency, reliability and coverage signals [Lioma et al., 2016, da Costa Pereira et al., 2009, 2012b, Dumitrescu and Santini, 2021]. Geographic IR is distinguished by its integration of temporal, spatial, and topical relevance signals most commonly used in the domain [Palacio et al., 2010, Daoud et al., 2013]. Community question answering is another domain in which topical relevance mainly refers to text passages and is combined with factors like recency and context's quality [Yulianti et al., 2018, Amancio et al., 2021]. Another identified domain is referred to as educational search in which primary school children are considered as users [Usta et al., 2021]. Legal search has witnessed recent explorations, as reflected by the included studies [Wiggers et al., 2023, Ma et al., 2023], while, in this domain, the conceptualization of relevance significantly diverges from other domains, as we analyzed in Section 2.1.1.

Additional domains include expert finding, with specific areas like expert translator finding [Rekabsaz and Lupu, 2014], local search [Kang et al., 2012], mobile Search [Bouidghaghen et al., 2011], personalized bookmark search [Eickhoff et al., 2013a], personalized contextual search [Moulahi et al., 2014b], and XML Retrieval [Ashoori and Lalmas, 2007].

Table 3.4: Table representing the various knowledge domains and search tasks alongside the number of studies conducted in each category.

Knowledge Domain and Search Tasks	Number of Studies
Web Search - Personalization (N=5) - Topic Distillation (N=4) - Child-Friendly Content Retrieval (N=1) - Programming Related Search (N=1) - Table Retrieval (N=1) - Other (N=13)	25
Medical Search - Biomedical Articles Search (N=5) - Consumer Health Search (N=4) - Other (N=2)	11
Social Search - Twitter Search (N=3) - Disaster Related Search (N=1) - Event Related Search (N=1) - Opinion Related Search (N=1) - Scientific Community Search (N=1)	7
E-commerce Search	6
Academic Search - Math Search (N=1) - Other (N=3)	4
Blog Post Search - Opinions Search (N=4)	4
Newswire Stories Search	4
Community Question Answering	2
Geographic Information Retrieval	2
Legal Search	2
Educational Search	1
Expert Finding - Expert Translator Finding (N=1)	1
Local Search	1
Math Search	1
Mobile Search	1
Personalized Bookmark Search	1
Personalized Contextual Search	1
XML Retrieval	1

From our analysis, distinct trends and patterns emerge across various domains. E-commerce research is mainly driven by industry stakeholders, implying a close relationship between real-world application needs and research advancements in this field. Academic institutions have been at the forefront of research in the medical domain, pointing to an academic interest in addressing its challenges. Furthermore, the chronological progression of research across domains reveals a dynamic evolution of focus areas. Web and newswire research, spanning from 1998 to 2021, underscores its longstanding and persistent relevance. The medical domain's concentrated activity between 2015 and 2022, with a peak in 2016, signifies an increased interest in recent years. Social and e-commerce search reflect the last decade's technological and commercial shifts, spanning 2011-2021 and 2017-2021, respectively.

3.3.2.2 What are the relevance factors utilized by researchers in the reviewed studies?

Table 3.5 presents the identified relevance factors across the knowledge domains and their associated search tasks. For example, web search considers factors such as topicality, reputation, and PageRank, among others. Specific search tasks of web search, like personalization and table search, consider their own sets of relevance factors such as user interest and multi-modal table properties, respectively. As it can been seen in Table 3.5, each domain has its unique set of relevance factors, some of which are shared across domains. This showcases the multidimensional and task-specific nature of information retrieval across diverse domains and search tasks. The analysis of the included studies revealed that certain domains are dominated by identical relevance factors; for instance, medical searches are often influenced by factors associated with the credibility of the information. Furthermore, some relevance factors remain consistent across multiple domains, exemplified by the usage of the recency factor regardless of the domain or task. Notably, there are relevance factors that essentially convey similar relevance signals but are mentioned differently, underscoring the need for future formalization to bring consistency. This is seen in terms such as credibility, trustworthiness, and genuineness, which although distinct in wording, often intersect in their conveyed meaning. In Section 3.3.3, addressing our second research question, we aim at analyzing relevance factors that fall in the aforementioned category by analyzing their definitions and operationalization.

Table 3.5: Table representing the various knowledge domains and search tasks alongside the exploited relevance factors for multidimensional relevance estimation.

Knowledge Domain and Search Tasks	Relevance Factors
Web Search	Topicality, Reputation, Readability, PageRank, Authority, Objectivity, Knowledge, Content Quality, Popularity, Freshness, Factuality, Coverage, Anchor Text, User's Actions, Temporal Relevance, Syntactic Relevance, Other Task-based Features
- Personalization (N=5)	Topicality, User's Interest, Scope, Reliability, User's Habit, Novelty
- Topic Distillation (N=4)	Topicality, Rareness, Proximity, Prominence, Position, Frequency, Document Length, Content Diversity, Authority
- Child-Friendly Content Retrieval (N=1)	Topicality, Appropriateness for Children
- Programming Related Search (N=1)	Topicality, Semantic Similarity, API Method-based score, API Class-based score
- Table Retrieval (N=1)	Topicality, Multi-modal Table Properties
Medical Search	Topicality, Passage Level Reliability, Passage Level Topicality, Genuineness
- Biomedical Articles Search (N=5)	Topicality, Content Diversity, Other Task-based Relevance
- Consumer Health Search (N=4)	Topicality, Understandability, Credibility, Readability
Social Search	
- Twitter Search (N=3)	Topicality, Trustworthiness, Temporal Relevance, Recency, Authority, User's Social Importance
- Disaster Related Search (N=1)	Topicality, Informativeness, Interestingness, Credibility, Opinionatedness
- Event Related Search (N=1)	Topicality, Hashtag-based Similarity, Event-based Topicality
- Opinion Related Search (N=1)	Topicality, Informativeness, Interestingness, Credibility, Opinionatedness
- Scientific Community Search (N=1)	Topicality, User-related Social Features, Popularity, Freshness
E-commerce Search	Topicality, Temporal Relevance (*Seasonality*), Sales, Reviews, Purchase User Intent, Node Compatibility, Item Popularity, Category Compatibility, Other Task-based Features
Academic Search	Topicality, Reliability, Recency, Readability, Coverage
- Math Search (N=1)	Image Similarity and Context Similarity based on Math Formulas
Blog Post Search	
- Opinions Search (N=4)	Topicality, Topical Evidence, Temporal Relevance, Social Features, Opinion, Authoritative Evidence
Newswire Stories Search	Topicality, Reliability, Objectivity, Freshness, Coverage, User-related Appropriateness, Factuality
Community Question Answering	Topicality, Recency, Passage Quality
Geographic Information Retrieval	Topicality, Temporal Relevance, Spatial Relevance
Legal Search	Document's Usage, Citations, Other Task-based Features
Educational Search	Task-based Features
Expert Finding	
- Expert Translator Finding (N=1)	Topicality (as a proxy to Language Proficiency), Price, Number of Cooperation Times, Duration of the translation
Local Search	Topicality, Reputation, Distance
Math Search	Taxonomic Distance of Functions, Data Type Hierarchical Level, Match-Depth, Coverage, Other Task-based Features
Mobile Search	Topicality, Location, User's Interest
Personalized Bookmark Search	Topicality, User-based Relevance
Personalized Contextual Search	User's Interest, Location
XML Retrieval	Topicality, Specificity, Exhaustivity

In Table 3.5, we reference the terms *Task-based Features* and *Task-based Relevance*. Recognizing that these terms might hold varying interpretations, we highlight their meaning within the framework of our review. We use the term *Task-based Features* when a study incorporates a considerable volume of features to estimate multidimensional relevance, often in a learning to rank (LtR) setting. This was evident in two studies in the e-commerce domain. In the study by Karmaker Santu et al. [2017], a set of 562 features is utilized, focusing on aspects related to the query, the document (in this case, a product), and the query-document relationship. These features encompass metrics such as BM25 scores, user ratings, and total sales. Similarly, Feng et al. [2018] deploy a variety of features to determine relevance. While the exact number of these features is unspecified, some illustrative examples include the item's popularity and rating score. Moving to the educational search domain, Usta et al. [2021] leverage 50 domain-specific and generic features. These related to queries (e.g. the name of a course), documents (for instance, the document's course), their relationship (like BM25), and also they leverage session data. In legal search, Ma et al. [2023] also generated a set of domain-specific features. Specifically, the authors, leveraging the structure of legal documents, they split them in three core segments, namely *Facts*, *Holding*, and *Decision*. By doing that, they create a token-level representation for each of the segments, concatenate them, and use them to train a LtR model. Similarly, in web search, Zhuang et al. [2021] propose the use of generalized additive models (GAMs) for ranking, in an approach that also leverages a vast amount of domain-specific and generic features. Lastly, in the medical search, Alsulmi and Carterette [2018] leverage 74 features for biomedical articles search. Regarding the term *Task-based Relevance*, this is used to describe three studies from biomedical articles search [Qu et al., 2020, 2021, Znaidi et al., 2016]. In those studies, the authors model relevance estimation by considering several relevance signals, and the characteristics of the search tasks. Specifically, the authors propose approaches that mimic the user's workflow and decision-making processes and develop search models that follow the same steps to predict a document's relevance.

A more detailed examination of the identified relevance factors can be found in Section 3.3.3, where we discuss proposed definitions and operationalization methods.

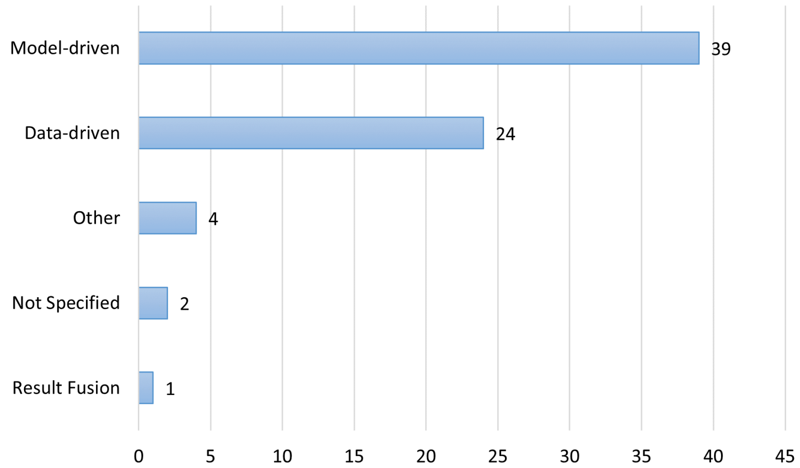

Figure 3.5: Number of studies categorized based on the employed aggregation approach.

3.3.2.3 What are the diverse approaches employed to aggregate relevance factors in the context of multidimensional relevance estimation?

In this section, we focus on the methodologies authors have adopted to aggregate multiple relevance factors into a unified relevance score. Based on our review, we categorized these methodologies as *data-driven*, *model-driven*, and *other* that includes studies that do not fall in either of these categories. Data-driven methods primarily use learning to rank or other machine learning techniques. Model-driven methods have been employed in the majority of the reviewed studies. Notably, the most frequent approach is a simple linear combination of the consider relevance factors. While our review primarily explores multidimensional relevance models, we acknowledge studies that leverage score fusion techniques, as it is a popular method for aggregating scores from distinct relevance factors. Figure 3.5 presents the distribution of studies based on their aggregation approach types, indicating that 39 studies employ a model-driven approach, 24 adopt a data-driven approach, and the remaining utilize result fusion, other methods, or do not specify their aggregation technique.

Model-driven Approaches. With a few exceptions, the majority of the model-driven approaches exploit a weighted linear combination to obtain an overall relevance score. Nonetheless, some exceptions do exist. *Linear combination* (or weighted linear combination) has been exploited to aggregate scores related to distinct relevance factors in web search [Craswell et al., 2005, Silva et al., 2019, Rinaldi, 2009, Lioma et al., 2016, Sahraoui and Faiz, 2017], math search [Zhang and Youssef, 2014], academic search [Singh and Dave, 2019, Jomsri and Prangchumpol, 2015], blog post search [Huang et al., 2018, Gerani et al., 2012, Chenlo et al., 2015], medical search [Upadhyay et al., 2022], social search [Putri et al., 2020, Madisetty and Desarkar, 2022, Tamine et al., 2011], e-commerce [Bassani and Pasi, 2021], community question answering [Yulianti et al., 2018], and geographic information retrieval [Daoud et al., 2013]. In consumer health search, Zhang et al. [2015] introduced a custom formula that applies an exponential weighting to the readability score and then multiplies it with a power-weighted topical relevance score. Another popular model-driven aggregation technique relies on *Copulas*. Copulas is a class of probability density functions that can be used to describe the dependence between multiple variables, separate from their individual behaviors or distributions. They excel at capturing complex, non-linear relationships, including the intricate connections seen at extreme values, known as tail dependencies. Due to that, several studies in our reviewed leverage copulas for multidimensional relevance estimation [Eickhoff and de Vries, 2014, Sieg et al., 2007, Sasaki et al., 2016, Komatsuda et al., 2016, Eickhoff et al., 2013a]. In their study, da Costa Pereira et al. [2009] proposed the usage of a *prioritized scoring aggregating* operator for multidimensional relevance estimation that assumes order of importance among the relevance factors. In detail, the importance weight of a certain criterion is dependent upon the satisfaction or score of a previous or higher-priority criterion. Bouidghaghen et al. [2011] introduced another operator for multidimensional relevance estimation, namely the *prioritized "and" operator*. The distinguishing aspect of this operator is the extent to which the least satisfied criterion is considered. Since their introduction, these operators have been used in several studies [da Costa Pereira et al., 2012b, Znaidi et al., 2016]. In math search, Yan et al. [2022] leverage the *hesitation fuzzy set* to obtain an interpretable document ranking. Other scholars have modified traditional *language models* by incorporating additional relevance factors. Specifically, they integrated these factors as prior probabilities or made specific adjustments to existing models [Badache and Boughanem, 2014, Bambia and Faiz, 2015, Ashoori and Lalmas, 2007]. Other studies introduce *relevance models*, e.g. probabilistic models, that account for several relevance aspects, based on the characteristics of the applied domains [Vargas et al., 2012, Bendersky et al., 2011, Jabeur et al., 2012, Uprety et al., 2018]. Finally, some studies consider the task of multidimensional relevance estimation as a multi-criteria decision-making (MCDM) problem [Moulahi et al., 2014b,a, Farah and Vanderpooten, 2008]. Therefore, these studies leverage MCDM methods such as the *Choquet Integral* or *ELECTRE*. It is worth noting that several studies mentioned above exploit some training data to predict a set of importance weights associated with the relevance factors.

Data-driven Approaches. For the data-driven approaches, their diversity makes it challenging to identify commonalities and categorize them; thus, we provide a concise description of each study. Numerous data-driven techniques have emerged in the field of e-commerce. Mandayam Comar and Sengamedu [2017] leverage users' search intents and propose a Multi-intent Poisson-beta model for product ranking. The model, which identifies users' purchase intentions based on observed click patterns, is trained using click logs data collected over 30 days from the Amazon product search dataset. Karmaker Santu et al. [2017] experimented with several LtR methods and found LambdaMART as the best performing for product search. The authors emphasize the efficacy of popularity-based features and found that click rates are more predictable than add-to-cart ratios. Their experimentation shows that model optimization based on order rates frequently yields the most consistent predictions, indicating a potential advantage in transitioning to order rate-centric models. Feng et al. [2018] propose the Multi-Agent Recurrent Deterministic Policy Gradient (MA-RDPG) tailored for multi-scenario ranking in the e-commerce domain. The model uses an online learning system that dynamically updates based on real-time user logs and a replay buffer mechanism. Consequently, it can continuously adapt to changing user behaviors. The study by Li et al. [2021] presents the Topic-enhanced Knowledge-aware retrieval model, which incorporates three dimensions of relevance, i.e. semantic similarity, knowledge relevance, and topical relatedness, to assess the relevance between a query and a document. The model aims to minimize simultaneously a ranking loss that ensures good semantic relevance, and the loss of the neural model that ensures topical relatedness. Yang et al. [2021] introduced LogSR and VelSR features based on neural models to capture product seasonality in e-commerce search. They incorporated these features into a standard LtR setup, validated their approach through offline and online experiments, and highlighted its efficacy. Finally, Carmel et al. [2020] address the challenge of optimizing multiple objectives, including maximizing product relevance and purchase likelihood simultaneously (a problem known as Multi-Objective Ranking Optimization - MORO). To that aim, the authors introduce a novel approach, namely stochastic label aggregation. This method randomly assigned labels to training examples based on a given distribution over the labels. Theoretical analysis and empirical experiments on different datasets revealed that MORO with stochastic label aggregation consistently outperformed deterministic label aggregation methods. Label aggregation has also been exploited as an approach in local search by Kang et al. [2012]. The authors define a label aggregation function that quantitatively combines multi-aspect relevance values into an overall score. To train this function, they use relative preference data, where one document is preferred over another. Once the aggregation function is learned, it is applied to a larger dataset containing ranking features and multi-aspect relevance vectors. This process generates an expanded dataset with overall relevance scores. Subsequently, they train a ranking function using this expanded dataset, enabling effective handling of multiple relevance dimensions in ranking models. In their study, van Doorn et al. [2016] perceive multiple relevance factors as objectives and aim to learn a set of rankers that provide different trade-offs concerning these objectives. They use a combination of gain-based evaluation and multi-objective optimization techniques, including Optimistic Linear Support (OLS) and dueling bandit gradient descent (DBGD), to find optimal rankers. In medical search, several learning to rank and machine learning approaches have been introduced, with few emphasizing interpretability. Applying LtR techniques for retrieving biomedical articles, Alsulmi and Carterette [2018] exploit a wide range of general and domain-specific features for ranking. Notably, among the algorithms investigated in the research, Coordinate Ascent emerged as the top-performing when combined with a feature selection strategy. For biomedical article retrieval, Xu et al. [2016] introduced a framework that combines multiple LtR techniques. This framework aims to optimize document ranking by considering topical relevance and diversity. The authors utilized label aggregation approaches to merge these two aspects and train the LtR models. Among all of the evaluated models, LambdaMART exhibited the best performance. Qu et al. [2020, 2021] propose a model that leverages structured search strategies to build an effective, explainable, and label-efficient retrieval algorithm for professional search tasks. This model utilizes machine learning classifiers to predict different aspects of the query and then combines these predictions using a logical function to determine document relevance. The experimental results show that their model performs as well as complex LtR models, even with limited labeled documents. In consumer health search, Putri et al. [2021] introduce a Multi-Task Learning model that simultaneously estimates relevance based on topicality and another factor, such as readability or credibility. This model combines a neural retrieval model for topical relevance estimation with a classification model that categorizes documents based on the aforementioned factors. Both of these models share certain model parameters during training and inference. Also in the context of consumer health search, Palotti et al. [2019] explored various methods to incorporate understandability and topicality into ranking. Of the tested methods, the authors concluded that LtR is the most effective. Fernández-Pichel et al. [2022] also leverage a LtR approach to rank health related documents by considering several factors. Their experiments revealed that result fusion methods, such as CombSUM, outperformed LtR in terms of effectiveness. The web search domain has also witnessed the advent of various data-driven techniques. The study of Li et al. [2017b] stands out as the most exhaustive one regarding the utilization of relevance factors and the depth of their feature engineering efforts. In the context of web search, the authors identify seven relevance factors, operationalize them using multiple features, and incorporate these features into a LtR model, i.e. LambdaMART. In their work, Zhuang et al. [2021] present interpretable ranking models that utilize generalized additive models (GAMs). These models can integrate both list-level and item-level features, making them well-suited for LtR tasks. In the context of web search, their experiments show that the proposed ranking GAMs outperform conventional GAMs while preserving their interpretability. Dai et al. [2011] introduced CS-DAC, a LtR methodology that optimizes topical relevance and freshness. The approach enhances the divide-and-conquer ranking technique by using hybrid labels and leveraging a new query-document importance factor that the authors introduced. Also, in web search, Collins-Thompson et al. [2011] propose a LtR method to re-rank topically relevant web pages according to their reading level. That is achieved by estimating the reading proficiency of users and the complexity of web pages, and by training a LambdaMART ranking model. Shraga et al. [2020] proposed a deep-learning retrieval technique for web table retrieval that considers web tables as multimodal entities. Their neural ranking model leverages Gated Multimodal Units (GMUs) to represent queries and table modalities jointly. Experiments indicate the potential of viewing web tables as multimodal structures in future research. In expert finding, Rekabsaz and Lupu [2014] develop a translator-expert retrieval system that leverages domain-specific features such as price and delivery time, among others, for ranking. Through empirical evaluations, they determined that a ranking model based on linear regression leads to superior performance. The study by Amancio et al. [2021] introduced a ranking approach for community question answering, leveraging quality and recency features. The authors experiment with nine LtR algorithms, from which Coordinate Ascent and LambdaMart lead to the best performance. Usta et al. [2021] employed a LtR approach specifically tailored for educational search. The model exploits features related to queries, documents, user's session, and their relationships. Furthermore, instead of using a single general model to rank all queries, the authors introduced query-dependent ranking models, grouping queries based on common characteristics, like association with a course user's grade level. These models lead to significant performance improvements. In legal search, Ma et al. [2023] proposed a structured LtR model to retrieve the most relevant legal cases for a given query. Their method uniquely combines semantic-level and charge-level relevance signals by integrating internal case details with external structural information about charges. Utilizing the Lightgbm model, they effectively aggregate these factors to produce a ranked list of cases, using nDCG as a training objective.

Other Approaches. The studies discussed below estimate multidimensional relevance by considering various relevance factors depending on the specific search tasks they address. In the context of academic and social search, the works by Arastoopoor [2018] and Ravikumar et al. [2013] both propose a retrieval pipeline that re-ranks an initial set of documents based on the considered relevance factor(s). Similarly, the work of Shajalal et al. [2020] sequentially re-ranks a set of documents, aiming to reduce information redundancy. Dumitrescu and Santini [2021] created a custom function highly tailored to the characteristics of the studied search task (i.e. newswire search). Lastly, the study of Palacio et al. [2010], apply rank fusion techniques to combine relevance scores, while, due to limited information, we can not classify the methods exploited by Wiggers et al. [2023] and Brin and Page [1998].

3.3.3 How do authors define and operationalize relevance factors (i.e. estimate a score to be associated with them) in the reviewed studies?

As highlighted in Section 3.3.2.2, some factors are recurrent across multiple domains, whereas others convey the same relevance signals but differ in terminology. To elucidate this, Sections 3.3.3.1 and 3.3.3.2 are dedicated to illustrating how the most frequent used relevance factors are defined and applied. To facilitate our analysis, in Table 3.6 we present a synthesis of these factors. On the left, it clusters similar relevance factors for easy comparison, and on the right, it enumerates the specific domains and search tasks where each factor has been operationalized.

3.3.3.1 How have the relevance factors been defined within the studies incorporated in the review?

Examining Table 3.6 we address this research question by presenting and discussing the diverse definitions of the listed relevance factors.

Topicality. Across the included studies, following the standard paradigm in IR, it has been defined as the degree to which the content of a document matches or relates to a query posed by a user.

Appropriateness, **User's Interest**, **Personal Relevance**, **Appropriateness for children**, **User Related social Features**, **User Intent**, **User's Habit**. *Appropriateness* was introduced by da Costa Pereira et al. [2009] and later adopted in da Costa Pereira et al. [2012b], both utilizing the same definition and operationalization. It has been defined as a relevance factor that estimates how appropriate a document is to the user's interest. The concept of *user's interest* has been referenced in multiple studies [Li et al., 2017b, Uprety et al., 2018, Sieg et al., 2007, Sahraoui and Faiz, 2017, Dumitrescu and Santini, 2021, Bouidghaghen et al., 2011, Tamine et al., 2011]. Yet, not every study provides a formal definition for it. Relying on previously introduced definitions, Tamine et al. [2011] consider that user interest expresses the cognitive background of the user. Li et al. [2017b] define interest as the extent to which the user prefers the retrieved documents according to their topics of interest, whereas Uprety et al. [2018] adopt the same definition in their study. In their investigation, Bouidghaghen et al. [2011] utilize Park [1994]'s definition, which assesses the "Interest" criterion as the degree to which a retrieved document aligns with the user's interest, a concept akin to appropriateness introduced by da Costa Pereira et al. [2009]. *User's Habit* has been defined by Li et al. [2017b] as the extent to which the retrieved documents are preferred by a user according to their sources, genre, and language, among others. This definition has been adopted also by Uprety et al. [2018]. Both the notions of *personal relevance* and *appropriateness for children* have been mentioned by Eickhoff et al. [2013a]. However, due to limited details in the paper, it is challenging to further analyze them in the context of our review. Within the e-commerce domain, Mandayam Comar and Sengamedu [2017] identify and utilize two distinct

Table 3.6: Table representing the most frequently used relevance factors across the included studies. The left column groups similar relevance factors together, while the right mentions the domains and search tasks in which they have been employed.

Relevance Factors	Knowledge Domain and Search Tasks
Topicality (N=64)	Exploited in the vast majority of the included studies
Appropriateness (N=2), User's Interest (N=7), Personal Relevance (N=1), Appropriateness for children (N=1), Users' Intent (N=1), User's Habit (N=2)	Web Search (Personalization, Child-Friendly Content Retrieval), Social Search (Scientific Community Search), E-commerce, Newswire Stories Search, Mobile Search, Personalized Bookmark Search, Personalized Contextual Search
Freshness (N=4), Temporal Relevance (N=3), Recency (N=3), Novelty (N=2)	Web Search, Social Search (Twitter Search, Scientific Community Search), E-commerce Search, Academic Search, Blog Post Search, Newswire Stories Search, Community Question Answering, Geographic Information Retrieval
Reliability (N=6), Credibility (N=2), Trustworthiness (N=1), Genuineness (N=1), Factuality & Objectivity (N=1)	Web Search (Personalization), Medical Search (Consumer Health Search), Social Search (Twitter, Disaster Related and Opinion Related Searches), Academic Search, Newswire Stories Search
Readability (N=6), Understandability (N=3)	Web Search, Medical Search (Consumer Health Search)
Content Diversity (N=4), Exhaustivity (N=1), Scope (N=2)	Web Search (Topic Distillation), Medical Search (Biomedical Articles Search), XML Retrieval
Authority (N=4), PageRank (N=2), Authoritative Evidence (N=1)	Web Search (Topic Distillation), Social Search, Blog Post Search
Coverage (N=5), Specificity (N=1)	Web Search, Academic Search, Newswire Stories Search, Math Search, XML Retrieval
Spatial Relevance (N=2), Location (N=3)	Geographic Information Retrieval, Local Search, Mobile Search, Personalized Contextual Search
Objectivity (N=1), Opinionatedness (N=1), Opinion (N=3)	Web Search, Social Search (Disaster Related and Opinion Related Searches), Blog Post Search
Content Quality (N=1), Passage Quality (N=2), Web Page Quality (N=2)	Web Search, Community Question Answering
Popularity (N=2), Reputation (N=2)	Web Search, E-commerce Search, Local Search

user intents —purchase and explore—to rank products. The authors perceive purchase intent as akin to the navigation intents in standard web searches but with the user's goal directed towards finding a specific product. When users are curious to explore the variety of items displayed by the retrieval system, it is considered an exploration intent.

Freshness, **Temporal Relevance**, **Recency**, **Novelty**. Based on our review, the terms highlighted earlier correspond to relevance factors that estimate comparable relevance signals [Badache and Boughanem, 2014, Bambia and Faiz, 2015, Dai et al., 2011, Dumitrescu and Santini, 2021, Yang et al., 2021, Jabeur et al., 2012, Amancio et al., 2021, Moulahi et al., 2014a, Daoud et al., 2013, Jomsri and Prangchumpol, 2015, Li et al., 2017b, Omidvar-Tehrani et al., 2022]. To enhance web search using social cues, Badache and Boughanem [2014] present a domain-specific interpretation of *freshness*, defining it as "a date of each social action (e.g. date of comment, date of share) performed on a resource on social networks can be exploited to measure the recency of these social actions, hence freshness of information." Another study in web search aiming to answer real-time sensitive queries defines a document's freshness relying solely on its content and specifically by including "fresh words" [Bambia and Faiz, 2015]. The authors consider as fresh words those that are trending on the social web and are topically relevant to the query, typically found in new social posts, micro-blogs, or breaking news. In a LtR approach for web search, Dai et al. [2011] define freshness as a concept sensitive to query temporal content, such as when users search for breaking news or events. In the context of newswire search and personalization, by considering also the notion of freshness, Dumitrescu and Santini [2021] argue that an item is considered fresh if it falls within a semantic domain of a user's interest that has not been encountered in the recent history. *Recency* is another term used in the literature. Amancio et al. [2021] conceptualize recency in community question answering by assessing the recency of the topics or terms present in an answer, i.e. the answer's content. This definition aligns with the one given by Bambia and Faiz [2015]. Lastly, Li et al. [2017b] and Uprety et al. [2018] exploit the term *novelty*, drawing on the definition put forth by Xu and Chen [2006a] who defined novelty as "the extent to which the content of a retrieved document is new to the user or different from what the user has known before" and argue that recentness can be regarded as one possible way of ensuring novelty, but not the only one.

Reliability, **Credibility**, **Trustworthiness**, **Genuineness**, **Factuality**. The terms mentioned above have been used in multiple studies of our review and point to similar relevance signals. The notion of *reliability* in web search has been defined in the studies of Li et al. [2017b] and Uprety et al. [2018] as the extent to which users trust a source, and it is associated with the wisdom of population. Similarly, in newswire stories search da Costa Pereira et al. [2009, 2012b] define reliability as the extent to which a user trusts a document's source, i.e. a source's reputation. In a different direction, Fernández-Pichel et al. [2022] perceive the reliability of web content as a combination of content correctness and source credibility. In Twitter search, Ravikumar et al. [2013] employ the term *trustworthiness*, associating it with both the source and the content of a tweet, whereas Putri et al. [2021], in consumer health search, use the terms *credibility* and trustworthiness interchangeable. Upadhyay et al. [2022] introduce the concept of *genuineness* as a new abstract term that encompasses the various aspects introduced above (credibility, trustworthiness, among others). In their study, Lioma et al. [2016] use *factuality* and *objectivity* as proxies to estimate credibility.

Readability, **Understandability**. In their study, Sasaki et al. [2016] adopt the *readability* definition introduced by Klare [2000], in which "text readability can be formally defined as the sum of all elements in textual material that affect a reader's understanding, reading speed, and level of interest in the given material." The other studies in our review that utilize readability for document ranking do not mention a formal definition. Concerning *understandability*, both Li et al. [2017b] and Uprety et al. [2018] treat the term as synonymous with readability. They adopt the definition from Xu and Chen [2006a], which describes understandability as a "complex cognitive concept that measures the extent to which the user perceives the content of a retrieved document as easy to read and understand." In their work on consumer health search, Palotti et al. [2019] differentiate the notions of readability and understandability so that readability measures how easy it is to understand a text. Understandability is a broader term that encompasses the text's readability and presentation, such as its legibility, layout, and even the use of visuals to clarify complex ideas.

Content Diversity, **Exhaustivity**, **Scope**. In biomedical article retrieval, Xu et al. [2016] incorporate *diversity* to maximize the coverage of query-related aspects in retrieved documents. Both Shajalal et al. [2020] and Singh and Dave [2019] exploit information topicality and coverage, as described above, as a proxy to retrieve documents with diverse topics. Based on the studies mentioned before, we observe a connection between coverage and topical diversification in the result list, where coverage serves as a means to attain diversification. In XML retrieval, Ashoori and Lalmas [2007] define *exhaustivity* based on the degree (i.e. how much) an XML element discusses the topic of the user's query. Li et al. [2017b] and Uprety et al. [2018] leverage the notion of *scope* in their experiments. Relying on the definition of Xu and Chen [2006a], the scope factor is defined as the extent to which the topic covered by a retrieved document is appropriate to the user's information need, that is, both breadth (similar to coverage/specificity) and depth (similarly to exhaustivity).

Authority, **PageRank**, **Authoritative Evidence**. In web search, the term *authority* relates to the source and reputation of a web page, frequently estimated using *PageRank* as an indicative measure [Zhuang et al., 2021, Eickhoff and de Vries, 2014]. *PageRank* has been defined by Brin and Page [1998] as a measure that quantifies the importance or "authority" of a web page based on the number and quality of links pointing to it. In social search, Moulahi et al. [2014a] define authority as the influence of tweets' authors on the platform. In blog search, Huang et al. [2018] interpret *authoritative evidence* as the relatedness of a blogger/feed's content to controversial topics and used it as a proxy to estimate *opinion*, as we will describe later in our analysis. Controversial topics refer to those that may cause controversy, argument and polarized opinions.

Coverage, **Specificity**. Both of these concepts are related to textual content. Specifically, da Costa Pereira et al. [2009, 2012b] define *coverage* as a measure related to the degree a user's interests are included in a document. A similar definition is provided by Dumitrescu and Santini [2021], who perceive it as the proportion of a user's interests represented by the documents retrieved from the stream, i.e. news streams, within a specific time span. Singh and Dave [2019] characterize minimum coverage as the shortest segment of the document, which covers all the user query terms that appear in that document. In math search, Zhang and Youssef [2014] estimate coverage by measuring the portion of a mathematical expression mentioned in a query and a given document. Shajalal et al. [2020] describe coverage in the context of their study as a measure that considers both the relevance of a subtopic to a query and how frequently that subtopic appears in documents. *Specificity*, in the context of XML retrieval, refers to how focused an XML element is on the topic of request, meaning it does not discuss other topics, irrelevant to the user's query [Ashoori and Lalmas, 2007].

Spatial Relevance, **Location**, **Distance**. Each of the terms highlighted above relates to geographic locations; however, our analysis will explore their specific interpretations. Specifically, Daoud et al. [2013] assess *spatial relevance* by concentrating on the query intent rather than the actual geographic *location* of the user, which is the focus of studies by Kang et al. [2012], Bouidghaghen et al. [2011], Moulahi et al. [2014b].

Objectivity, **Opinionatedness**, **Opinion**. In the context of web search, Lioma et al. [2016] use the notion of *objectivity* along with the concept of *factuality* as proxies to credibility. The authors consider objectivity as the degree to which text meaning depends on the author's perspective, i.e. the exact opposite notion of subjectivity. Regarding the concept of *opinionatedness*, Putri et al. [2020] define it based on the likelihood of a document to express an opinion about a query, a synonym to the term *opinion*.

Content Quality, **Passage Quality**, **Web Page Quality**. According to Bendersky et al. [2011], *quality* of a web page can be evaluated based on multiple criteria including its originality, trustworthiness, content relevance, metadata accuracy, interlinked resources, and user-centric layout design. From the provided description, it is evident that the concept of quality is broad, incorporating multiple of the previously described relevance criteria. The domain of community question answering has also utilized the concept of *passage quality* regarding the retrieved answers [Yulianti et al., 2018, Amancio et al., 2021]. Nonetheless, the domain has yet to offer a clear definition of the concept of quality.

Popularity, **Reputation**. From our review of the included studies, the concepts of *popularity* and *reputation* emerge within e-commerce, social search, web search, and local search contexts. Badache and Boughanem [2014] treat them as two distinct notions that characterize a document, and define popularity as a measure of how well-known a resource is among the public, primarily driven by sharing and commenting activities on social networks; while reputation reflects the general opinion or appreciation of that resource, determined by positive social actions, such as number of likes. In e-commerce, Bassani and Pasi [2021] exploit products' popularity as ranking feature. A product's popularity is reflected by how often users choose it.

Even though we have made significant efforts to combine all the relevance factors mentioned in Table 3.5 based on their conceptual similarity, there remain certain domain-specific factors, like document's usage and citations in legal search [Wiggers et al., 2023], which we could not assimilate with other factors. Therefore, we direct readers interested in these specific factors to the original papers.

3.3.3.2 What methodologies and techniques are used to operationalize the identified relevance factors?

Based on Table 3.6, we address this research question by presenting and discussing the methodologies that have been leveraged to operationalize the identified relevance factors (i.e. estimate a score).

Topicality. In the majority of the included studies, *topicality* has been estimated by the BM25 model or other lexicon-based retrieval models. However, in studies that leverage a LtR approach, topicality has been represented by several lexicon-based retrieval models.

Appropriateness, **User's Interest**, **User based Relevance**, **Appropriateness for children**, **User Intent**, **User's Habit**, **User's Familiarity**. *Appropriateness* has been calculated by examining the similarity between term-based vector representations of a given document and user's interest [da Costa Pereira et al., 2009, 2012b]. To operationalize *user's interest* Li et al. [2017b] and Uprety et al. [2018] estimate it by capturing terms and topics from SAT-Clicked documents in a session, a day, and long term, based on previously published methods. Bouidghaghen et al. [2011] built upon prior research to estimate user's interest and use a method that conceptualizes it as a collection of weighted concepts. To determine a document's interest scores, they measured the cosine similarity between a document representation and the highest k-ranked concepts from the user profile. Although Sieg et al. [2007] do not explicitly define the notion of user's interest, they propose the creation of an ontological user profile, which is updated during the search session to reflect changes in the user's interests. Unlike previous studies, the authors assume that a user's interest is not static. Similarly, Sahraoui and Faiz [2017] consider the user's interest as a dynamic notion during a search session, in this sense the authors also perceive relevance as a multidimensional and dynamic notion. In their study, the authors define users' interests implicitly from their social Web activities and represent them as vectors of weighted terms. Recognizing the evolving nature of interests, they suggested adjusting term weights based on their recency and frequency to capture new and persistent interests. Dumitrescu and Santini [2021] introduce a set of algorithms to dynamically filter a stream of documents, ensuring they align with a user's interests and provide a diverse range of content. To achieve that, they employed an adapted version of a previously published algorithm, creating a user model from a representative collection of documents. Their approach distinguishes new content from areas the user has not recently engaged with and ensures comprehensive coverage of their varied interests. Tamine et al. [2011] implicitly capture a user's interest in literature retrieval based on social network analysis by measuring co-authorship based on the assumption that collaborators have shared interests. Moving to the *user's habit* factor, this has been operationalized by Li et al. [2017b] and Uprety et al. [2018] using three different methods that leverage behavioral signals. The first evaluates users' preference for a particular source website, drawing from their historical interactions and overarching global query logs. The other models aim to capture user's preference towards documents of specific lengths and language. In e-commerce, Mandayam Comar and Sengamedu [2017] operationalize *user intents* and incorporate them in their relevance model by looking at how often users clicked on results at different positions, estimating the click-through rate (CTR) of user profiles. Users with purchase intent typically have a rapidly declining CTR as position increases. In contrast, exploration intent shows a consistent CTR across positions.

Freshness, **Temporal Relevance**, **Recency**, **Novelty**. To estimate multidimensional relevance incorporating the notion of *freshness*, Badache and Boughanem [2014] propose a model that relies on counting specific social actions (i.e. like, share, comment) conducted on a resource (i.e. document). This model adjusts the count based on when an action occurred, so resources with more recent actions are promoted. Based on their domain-specific definition Bambia and Faiz [2015] assume that *freshness* can be described by a set of known terms extracted from current search trends or other sources. Then, using a language model, the authors evaluated the closeness of query terms to those terms and estimated a freshness score for each document. Constructing a set of features that leverage a temporal contextual profile of queries constructed based on a set of pseudo-relevance retrieved documents to a query, Dai et al. [2011] assess freshness. Dumitrescu and Santini [2021] exploit the notion of freshness alongside the notions of user's interest (i.e. personalization) and coverage. To incorporate the notion of freshness in search, the authors integrate the timestamp of an item in their estimations. Similarly, in academic search, Jomsri and Prangchumpol [2015] integrate the *recentness* of a publication into their ranking, utilizing a normalized version of the paper's publication year. In social search, Moulahi et al. [2014a] estimate a tweet's recency by considering the time lapse between its publication and the submission time of a query. Likewise, in community question answering, Amancio et al. [2021] employ features like the answer's creation date, the most recent date mentioned in a referenced web page text, to train a LtR model to associate a recency score for an answer. Jabeur et al. [2012], although they do not define the notion of *temporal relevance*, they estimate it based on the occurrence of query term configuration in temporal neighbor tweets under predefined temporal intervals. Yang et al. [2021] propose a domain-specific notion of temporal relevance, namely *seasonality* of products. Even without a formal definition, its implication is intuitively understood. To predict seasonality, the authors train a model that utilizes the annual sales data for a calendar year and create vector representations based on product-month relationships. To estimate a temporal relevance score in geographic IR, Daoud et al. [2013] use a probabilistic ranking model that considers the temporal frequency of terms within the document and the weight of the temporal query context. To estimate *novelty* in their models, Li et al. [2017b] and Uprety et al. [2018] exploit four features grounded in both temporal and psychological views of novelty. Those were related to the divergence between the language model of a retrieved document and previously viewed documents or estimated the time gap between a document's creation and retrieval.

Reliability, **Credibility**, **Trustworthiness**, **Genuineness**, **Factuality**. According to da Costa Pereira et al. [2009, 2012b], a source's reliability could be assessed based on past observations and user's-source interaction data. To estimate reliability, Li et al. [2017b] and Uprety et al. [2018] employ seven features based on SAT-clicks. Notably, the authors leverage the PageRank score of a web page as a proxy for its reliability. Fernández-Pichel et al. [2022] argue that a document's reliability needs to be estimated primarily relying on query-related document's content. To estimate a reliability score, the authors propose two approaches, one based on a fine-tuned Mono T5 model that classifies a passage as reliable or unreliable and an unsupervised approach that measures the similarity of a document's passage to true and false query-related handcrafted claims. In academic search, Jomsri and Prangchumpol [2015] associate a document's reliability based on the type of research paper publication, which varies from Journal to file. In social search, Putri et al. [2020] exploit a model-driven approach based on multi-criteria decision-making, initially proposed by Pasi and Viviani [2018], to estimate a document's credibility score. To estimate a *trustworthiness* score, Ravikumar et al. [2013] use a set of features related to a user's profile (number of followers, verified profile, among others) and to the content of the tweet, e.g. length, or hashtags. All these features are used in a LtR setting to predict an overall score. To estimate credibility/trustworthiness in consumer health search, Putri et al. [2021] leverage a set of features related to the presence of internal and commercial links and commercial content in a document. Upadhyay et al. [2022] propose an unsupervised method to evaluate the genuineness of online health information using a set of scientific articles that can support the claims made in a document. The authors compute a genuineness score by estimating and aggregating the cosine similarity values between the context of a document and a selection of k medical articles that cover the same topic. Finally, [Lioma et al., 2016] estimate credibility based on indicators of *factuality* and *objectivity*. The authors constructed two distinct data collections and trained two models that predict these scores.

Readability, **Understandability**. Sasaki et al. [2016] propose a method to evaluate a document's *readability* by assessing its complexity across three dimensions: vocabulary (e.g. syllables), sentence structure (e.g. length), and overall document structure (e.g. depth of heading tags). The probability of a document's readability was then determined using logistic regression. Another study by Arastoopoor [2018] explore the application of classic readability measures to scientific texts in Persian. To estimate readability, the study utilized Flesch–Dayani's formula, specifically designed for Persian. In their research, Putri et al. [2021] employed eight established readability formulae, such as the Gunning fog formula, proposed in prior studies. In their research focusing on consumer health search task, van Doorn et al. [2016], recognizing that conventional readability metrics may not align well with real-world readability in medical contexts curated a list of medical terms, derived from an English Wikipedia page, along with the Coleman-Liau index, Gunning fog index, and document length, they utilized machine learning model to predict the understandability score of a document. Zhang et al. [2015] suggest a two-layered approach to assess document readability: one based on surface content (using readability formulas like Putri et al. [2021]) and another on underlying document's topics. They introduced a method that considers both these levels, using tools like Topic Trace to follow topics and Topic Scope to calculate how much a document covers a topic. Regarding readability, Collins-Thompson et al. [2011] explore predicting a web page's reading level, leveraging its search result 'snippet' and the entire body text as features in a classifier. Moreover, the authors inferred users' reading proficiency from their search behavior, aiming to incorporate both in their ranking model. In their attempt to estimate *understandability* related features for their models, Li et al. [2017b] and Uprety et al. [2018] introduce seven features grounded in established understandability and readability metrics and user click-through data. Finally, Palotti et al. [2019] investigate various methods for estimating the understandability of health-related web pages, showing that machine learning techniques that leverage natural language, HTML structure, and domain-specific features outperform traditional readability metrics.

Content Diversity, **Exhaustivity**, **Scope**. To achieve *diversity* in results, van Doorn et al. [2016] employ MMR and cluster-based ranking techniques from the literature that re-ranks a set of documents based on their topical diversity. Xu et al. [2016] use a group-wise learning to rank framework that retrieves topically relevant and diverse documents. Their model relies solely on features related to topically, while diversity has been incorporated during the training phase. Specifically, during the training phase, the authors divided relevant documents into groups based on the different aspects they covered. Each group consisted of a document that covered more aspects (labeled as 1) and several others with fewer aspects (labeled as 0). The document with more aspects encompassed all the aspects found in the other, less comprehensive documents. Vargas et al. [2012] propose a relevance model that unifies previous approaches in the literature (i.e. the IA-Select and xQuAD models) to integrate result diversification based on users' intent. To estimate an *exhaustivity* score, Ashoori and Lalmas [2007] again rely on a topic segmentation algorithm, as described above. Even though Li et al. [2017b] and Uprety et al. [2018] leverage the *scope* relevance factor that encompasses both coverage/specificity and exhaustivity, they operationalize it solely the coverage aspect. Specifically, they exploit features that, for example, estimate the number of query term appearances across a document.

Authority, **PageRank**, **Authoritative Evidence**. In the studies examined, features associated with *authority*, including *PageRank*, utilize LtR methods to order web pages based on various relevance factors [Zhuang et al., 2021, Eickhoff and de Vries, 2014]. Furthermore, some research calculates a unique PageRank score, linking it to topical relevance [Craswell et al., 2005, Farah and Vanderpooten, 2008]. A prime example is Google, which initially based its algorithm on topical similarity and PageRank Brin and Page [1998]. Moulahi et al. [2014a] estimate a user's authority on Twitter by considering the volume of tweets the user has published and the number of times the user has been mentioned or cited by others.

Coverage, **Specificity**. da Costa Pereira et al. [2009, 2012b] measure *coverage* using a fuzzy inclusion, determined by the cardinalities of the fuzzy sets that represent both the user interests and the document. Dumitrescu and Santini [2021] estimate coverage by monitoring a dynamically changing "interest" parameter for different semantic areas. As content related to a topic is engaged, the interest in that topic and nearby semantic areas decreases. After a series of items, the overall level of remaining interest indicates the coverage of those items. Singh and Dave [2019], exploit a formula that considers the length of the user's query, the coverage of search terms, and the number of search terms missing from the document. Mathematical coverage has been estimated by simply counting the number of covered terms [Zhang and Youssef, 2014]. For coverage estimation, Shajalal et al. [2020] introduce a formula that multiplies the fraction of the number of snapshot terms by the number of complete document terms, with the inverse of its rank normalized by the total number of documents. To measure *specificity*, Ashoori and Lalmas [2007] utilize a topic segmentation algorithm based on lexical cohesion. The foundational idea behind the algorithm is that a vocabulary shift indicates a topic change.

Spatial Relevance, **Location**, **Distance**. To estimate the *spatial relevance* of a document to a query, Daoud et al. [2013] first extract query's geographic context (i.e. locations) from a topically relevant pseudo-relevant documents. The geographic score of a document is determined using a probabilistic ranking model, where instead of inverse document frequency, the frequency of documents with a geographic expression is used. In geographic IR research, Palacio et al. [2010] utilize a specially designed document index that contains spatial information. Using this index, the authors compute a relevance score reflecting the spatial relationship between the documents and the query. Estimating a user's actual physical *location* is a simpler task. Kang et al. [2012] employ a LtR model to predict multiple aspects, specifically focusing on location-related queries. They trained their model using label aggregation across the three relevance aspects they investigated. In the context of mobile search, Bouidghaghen et al. [2011] use a geographic weighting function previously introduced in the literature. A relevance score is estimated considering a geographic hierarchy, a user's geographical location, and a set of documents. In the study by Moulahi et al. [2014b], document scores are computed based on the location factor by assessing the distance between identified places in the documents and the specific user's context.

Objectivity, **Opinionatedness**, **Opinion**. To estimate *objectivity*, Lioma et al. [2016] estimate a document's objectivity using a subjectivity detection approach proposed in the literature. The authors trained an objectivity classifier using extracted patterns, lexicon entries, and POS features, and then applied this model to determine the objectivity of each document based on the proportion of its objective sentences. In social search, [Putri et al., 2020] estimate opinionatedness using two scores: a term-based score, which assesses opinionated terms (identified from a lexicon), and a stylistic-based score, which evaluates elements such as emoticons and exclamation marks in a tweet. A similar approach has been followed by Gerani et al. [2012] to measure *opinion*. Eickhoff et al. [2013a] calculate an opinion related score for a document using a state-of-the-art classifier in the literature. In blog post search, Huang et al. [2018] follow a more complicated process to estimate opinion in blog search. The authors estimated opinion in blog feeds and posts by associating them with the degree to which they relate to controversial topics. They employed a language model to determine an opinion score based on the generation probability of topical terms present in the post. Owing to space constraints, we direct readers seeking further details to the original publication.

Content Quality, **Passage Quality**, **Web Page Quality**. To incorporate a document's *quality* in their Markov Random Field model, Bendersky et al. [2011] used a set of features related to document's content (e.g. entropy of the page content), structure (e.g. depth of the URL path), and presentation (e.g. measuring the fraction of visible text on the rendered page). In their research on community question answering, Yulianti et al. [2018] use a mix of features, such as term overlap, sentence count, and term importance, to estimate *passage quality*. Similarly, Amancio et al. [2021] harness a total of 186 features and exploit a classifier for quality prediction. In their retrieval models that utilize numerous features to rank web pages, Zhuang et al. [2021], Eickhoff and de Vries [2014] incorporate *web page quality* scores derived from classifiers.

Popularity, **Reputation**. To measure a *popularity* score to be associate with a product, Bassani and Pasi [2021] use the n-root of the total number of times the item has been purchased. Badache and Boughanem [2014] measure popularity based on the number of comments on social platforms, number of tweets, and shares, while *reputation* was determined by social activities with positive meanings, such as likes. Kang et al. [2012] leverage the concept of *reputation* to improve document ranking in local search. To estimate a reputation score, the authors consider user reviews as a primary source, and along the other two relevance factors, namely topicality and location, they train a LtR model.

The relevance factors previously mentioned encompass most of the factors that have been identified and implemented to estimate multidimensional relevance in the reviewed studies. Despite the fact that we make a big effort to merge them based on their conceptual similarity, there are still some very domain-specific factors such as document's usage, citations in legal search [Wiggers et al., 2023], that we could not merge with other factors. For these factors, we refer the interesting readers to the original publications.

3.3.4 Which benchmark collections have been used to estimate multidimensional relevance, and how are they characterized based on their annotated relevance factors, availability, and size?

The field of multidimensional relevance estimation relies significantly on the existence of benchmark collections that facilitate experimentation and evaluation, as these datasets provide the foundation for conducting research and comparing methodologies. This section investigates which benchmark collections have been employed for multidimensional relevance estimation. Specifically, we delve into their specific characteristics, including annotated relevance factors, availability, and size. By exploring and presenting these aspects, we aim to provide a complete overview that aids researchers and practitioners, offering valuable insights into the resources available for advancing this study area.

Table 3.7 provides an overview of various knowledge domains and their associated data collections, most of which are related to initiatives such as TREC, CLEF, and NTCIR, among others. Web search stands out as the domain with the most collections, boasting five distinct datasets. The authors evaluate on custom collections in the fields of educational search, expert finding, local search, math search, and mobile search. That underscores the need to develop new datasets in these areas. A diverse range of collections was utilized across the seventy studies we reviewed. Nineteen of these studies relied on custom collections; three used private datasets, and six used collections crafted by other studies in the literature (outside of traditional evaluation campaigns). We refer to collections crafted by authors of the reviewed or other studies as *Custom Collections*. TREC emerged as the predominant data resource, referenced in 24 studies, followed by five CLEF datasets, mainly for consumer health search. NTCIR and INEX datasets were used in 3 and 2 studies, respectively. For detailed descriptions of each collection, we refer the interested readers to the TREC[1], NTCIR[2], and CLEF[3] official websites. In addition, for those collections that have not originated in these initiatives, we have provided their official names that can be used for search. Regarding the custom collections, the interested readers can identify them based on the studies that investigate web search. In the rest of the section, we briefly overview the most commonly employed collections.

Regarding the TREC collections employed in web search, most originated in the TREC WEB track, which was running between 1999 - 2004 and 2009 - 2014. Up to 1999, collections were based on a 1997 web crawl. Between 2002 and 2004, the *topic distillation* task and the .GOV collection emerged. From 2009 to 2012, the ClueWeb09 collection and *diversity-task* were introduced. In 2013 and 2014, ClueWeb12 launched, shifting focus from diversity to *risk-sensitive* task. Over time, the mentioned collections have expanded to include hundreds of queries and thousands of relevance assessments concerning topical relevance and other factors. These collections can be accessed upon request. The data provided in the TREC Session Track (2010-2014) has also been used across several reviewed studies, such as Li et al. [2017b], Uprety et al. [2018]. While this collection lacks explicit judgments on relevance factors beyond implicit relevance, it has been employed in studies examining the impact of various relevance factors in web search. The Yahoo! Learning to Rank Challenge Dataset, MSLR-WEB10K, and WEB30K collections have been used in studies that mainly explore learning to rank approaches for multidimensional relevance estimation. NTCIR-10 INTENT and NTCIR-12 IMine-2 collections have been exploited to support experiments

[1] https://trec.nist.gov/, accessed on 26/9/2023.
[2] https://research.nii.ac.jp/ntcir/index-en.html, accessed on 26/9/2023.
[3] https://www.clef-initiative.eu/, accessed on 26/9/2023.

Table 3.7: Knowledge domains and associated benchmark collections.

Knowledge Domain	Collections' Source
Web Search	A. TREC: Clueweb09, GOV2, Session Track (Clueweb09, Clueweb12) B. Yahoo! Learning to Rank Challenge Dataset C. MSLR-WEB10K & WEB30K D. NTCIR-10 INTENT2 & NTCIR-12 IMine2 E. Custom Collections (N=8)
Medical Search	A. CLEF eHealth 2015/16, 2018, 2020 B. TREC: Precision Medicine 2017/18/19, Genomics 2006/07, Clinical Decision Support (CDS), Health Misinformation 2020 Collection C. CLIREC Dataset
Social Search	A. CLEF: Microblog Cultural Contextualization 2017 B. TREC: Microblog 2011/12 C. Custom Collection (N=1)
E-commerce Search	A. Amazon Review 5-Core dataset, Amazon product search B. Custom Collections (N=3)
Academic Search	A. NTCIR Corpus MathIR-Wikipedia B. Custom Collections (N=3)
Blog Post Search	A. TREC: BLOG06
Newswire Stories Search	A. Reuters: Corpus Volume 1 (RCV1-v1)
Community Question Answering	A. TREC: ClueWeb09B, GOV2 B. Custom Collection (N=1)
Geographic Information Retrieval	A. TREC: Robust Retrieval Track 2004 B. MIDR_2010
Personalized Contextual Search	A. TREC: Contextual Suggestion 2013
XML Retrieval	A. INEX-2005 Collection

related to topic distillation [Shajalal et al., 2020].

Shifting our focus to medical search, both CLEF and TREC collections have been used. The CLEF eHealth collections allow experimentation in consumer health-related search on the web and evaluations based on topical relevance and readability. The Precision Medicine track uses two primary document collections, one including scientific abstracts and clinical trials. The aim is to retrieve relevant scientific abstracts for specific patient conditions and identify clinical trials for which a patient might qualify. Relevance is determined based on four patient-related dimensions: disease, gene, demographic, and other. The TREC Genomics Track introduced a task in 2006 centered on retrieving biomedical documents. Similarly, the Clinical Decision Support track uses medical case narratives to retrieve biomedical articles. In both tracks, relevance has been assessed based on topicality in a three-scale. However, the reviewed studies that leveraged these collections introduce retrieval approaches tailored to the characteristics of the task or leveraged the different patient-related aspects to estimate different relevance signals [Xu et al., 2016, Qu et al., 2020]. The TREC 2020 Misinformation track, focusing on COVID-19 misinformation, aims to retrieve useful, credible, and correct information. Finally, CLIREC is a test collection for evaluating clinical information retrieval that exploits a set of manually crafted PICO-structured queries to retrieve medical documents [Znaidi et al., 2016]. These collections are accessible and provide hundreds of queries along with thousands of relevance assessments.

In social search, CLEF's Microblog Cultural Contextualization is a multilingual collection that contains millions of event-related micro-blogs; along with TREC's Microblog collections, these collections have been employed in the vast majority of the reviewed studies, besides the work by Tamine et al. [2011]. Most of the e-commerce studies rely on machine learning approaches for ranking. As a result, these studies leverage the collections mentioned in Table 3.7, which mainly comprise thousands of queries, millions of products/documents, and, often, search log data. The academic search domain primarily utilizes custom collections. Similarly, custom collections have been utilized in legal, educational, expert finding, local, math, mobile, and personalized bookmarking search. Regarding the rest of the domains presented in Table 3.7, we refer the interested reader to the reviewed domain studies for further information regarding the collections.

Concluding, our analysis brought to light some notable observations. Firstly, the need for available datasets with annotation based on several relevance factors in several of the identified domains. Secondly, it revealed a pronounced correlation between the amount of research studies in a particular domain and the availability of benchmark collections for that domain. Finally, it highlighted the absence of datasets with annotations based on diverse relevance factors. Indeed, even if several studies created retrieval systems that leverage multiple relevance factors, the authors based their evaluation on labels that solely assess topical relevance. It is acknowledged that generating such annotations demands more time and resources. Nevertheless, the potential to develop multidimensional retrieval systems could make it worthwhile since these systems have been shown to enhance performance across various reviewed search tasks.

3.4 Discussion and Suggestions for Future Research

This section discusses the findings from our thorough literature examination concerning estimating multidimensional relevance. The aim is to synthesize the primary findings and underscore the significant contributions of this review.

Our analysis revealed that relevance is conceptualized and operationalized as a multidimensional notion across various knowledge domains and search tasks. Over the years, this research area has facilitated numerous international collaborations, maintaining a steady volume of publications. Moreover, the domain connects industry and academia, with some domains dominated by industrial contributions (e.g. e-commerce) and others, like the medical domain, by academia. Nonetheless, there are evident synergies between the two. Such collaborations underscore the theoretical interest and the substantial real-world applicability of multidimensional relevance search systems. Although our review included several diverse domains and tasks, we distinguished shared practices regarding the exploited relevance factors and the models employed to estimate multidimensional relevance.

Relevance Factors. Regarding the employed relevance factors, some have a consistent presence across diverse domains. Nevertheless, a significant inconsistency in their definitions and operationalization emerged. Specifically, there were instances where relevance factors, while conceptually similar, were articulated with varying terminology. For example, factors such as credibility, reliability, trustworthiness, genuineness, authority, objectivity, correctness, and factuality. These factors have been employed in the literature to determine, up to a certain degree, whether a user should "trust" a piece of information. However, we noticed that the relationship between them exhibits a form of dynamically changing contextual dominance, meaning that one study might consider reliability to be superior to credibility, i.e. using credibility as a feature of reliability, while others do the opposite. This variability complicates the endeavor of providing formal definitions for the diverse notions, and future research should address this issue. Additionally, we noticed inconsistencies regarding the computation of several factors. For example, some studies estimated reliability with respect to a document's source, i.e. leveraging its attributes or metadata. Other studies measured it by considering the document's contents. Moreover, others are based on the user's perceived trust in a source, giving it a user-specific viewpoint. Similar observations have been made for other relevance factors, such as those related to the temporality of information (e.g. recency, freshness). In this case, some studies calculate it based on the document's metadata, by considering the content, and also with respect to a user's related content. Similar observations can be drawn for many relevance factors in the literature and significantly undermine any effort for homogeneity.

Attempting to address the aforementioned issue, we put forward a structured formulation for defining relevance factors. In this formulation, authors should clearly define a relevance factor and elucidate its operationalization and relationship with other relevance factors from the literature. Specifically, the authors should mention whether the consider relevance factor has been estimated with respect to user [U] (e.g. leveraging a user profile), documents [D] (e.g. leveraging documents' metadata or attributes), task [T] (e.g. follow the relevance process of the search task like Qu et al. [2020, 2021]), content [C] (e.g. text), or other [O] viewpoints. Following this approach, introducing a new term becomes unnecessary if its estimation relies on viewpoints already covered by a another concept. Based on our personal viewpoint, introducing a new concept (i.e. new terminology) requires that the concept encompasses new viewpoints. In any other case, the proposed approach just amplifies the quality of estimating a concept. For example, if a study introduces a neural method to calculate readability using the content of documents, it simply offers a more refined estimate compared to traditional readability formulas that also utilize document content. Given that a new concept has been introduced, the authors should describe its relationship with other concepts in the literature, followed by a justification. Based on definitions provided in the reviewed studies, an identified relationship is $quality\ of\ information \gg reliability \gg credibility$. This relationship implies that the notion of information quality includes both the concepts of reliability and credibility and the concept of reliability also encompasses credibility. It is important to note that reliability is defined as the degree to which users place trust in a source, while credibility is dependent on the source itself. Consequently, reliability is assessed from the perspective of users [U], and credibility is assessed based on the document's metadata [D]. As a result, the estimation of information quality takes into account factors related to both user and document factors [U, D].

Aggregation Approaches. Our distinction between model-driven and data-driven approaches sufficiently allowed us to classify most of the studies in our review. Learning to rank and model-driven approaches exhibit distinct characteristics in their methodologies. Model-driven strategies are rooted in explicitly defined mathematical models. Our analysis showed that while many studies propose intricate methods to calculate a relevance factor's score, they mainly use a simple linear combination to estimate a final relevance score. While alternatives like copulas and MCDM methods have been suggested, they have yet to gain the community's attention, as most recent studies still exploit a linear combination. These approaches have a tendency to prioritize transparency and interpretability, which enhances their ease of understanding. However, this preference for transparency may come at the cost of potentially lower performance and, in some cases, increased computational complexity during inference.

Conversely, data-driven approaches harness a wide range of methods to address the challenge of aggregating information in multidimensional relevance estimation. These methods generally result in improved performance across most tasks; however, this improvement comes at the cost of reduced interpretability. Based on our analysis, label aggregation ranks among the predominant approaches for multidimensional relevance estimation with LtR methods. This method provides a straightforward approach for converting a multidimensional relevance problem into a single relevance estimation problem. LambdaMART and Coordinate Ascent have consistently stood out as top-performing methods throughout the studies we reviewed. Moreover, several researchers explore new directions, like query-dependent ranking models or models that adapt to changing user behaviors. Finally, interpretable multidimensional ranking models represent another avenue that is increasingly capturing research interest, especially in domain-specific search tasks.

Benchmark Collections. Our exploration points to an emerging need for benchmark collections annotated with a variety of factors across domains. That does not necessarily entail diving into exceedingly complex relevance factors. Instead, initial research efforts can be simple, focusing on exploiting relevance signals tied to document attributes. Doing so makes it feasible to delve deeper into how integrating these attributes impacts retrieval performance metrics, such as citations and the quality of venues in academic search. We have noted that structured, multidimensional collections play a pivotal role in shaping the research landscape. This observation is substantiated by initiatives that have created benchmark collections, like TREC, NTCIR, and CLEF, that guide the academic community's focus towards specific topics. Conversely, the industry operates independently from these trends, often addressing unique challenges and producing original datasets. Creating benchmark collections for multidimensional relevance might be a time-consuming and expensive task. However, the emergence of LLMs offers promising potential, primarily as tools to deliver relevance annotations [Thomas et al., 2023].

3.5 Prospects of the Study and Limitations

This section outlines our study's limitations associated with the search strategy's effectiveness, the coding scheme's reliability, and potential biases inherent to our methodology. Nonetheless, we highlight the relevance and significance of this study driven by recent technological advances. The recent advent of large language models and their impressive relevance labeling capabilities [Thomas et al., 2023] highlight the timely significance of this review. As discussed in Section 3.3.4, developing models for multidimensional relevance necessitates new benchmark collections, especially for specific domains. Nevertheless, the creation of these collections is both resource-intensive and time-consuming. The work of Thomas et al. [2023] paves the way for leveraging LLMs for annotation tasks traditionally reserved for human annotators. However, the efficacy of these models in performing such tasks is contingent upon the quality of the prompts. Our review, especially Section 3.3.3.1 detailing the several definitions associated with the identified relevance factors, might be instrumental in crafting these prompts.

Another factor underscoring the significance of our study pertains to the LLMs' proficiency in text comprehension, which potentially allows them to estimate topical relevance scores. This advancement allows the community to transition from developing IR models centered solely on topical relevance to multidimensional relevance models incorporating user, task, and domain characteristics into a retrieval process.

Having pointed out its potential, we now turn our attention to the limitations of our study. At the initial stage of the literature review process, the existing research landscape was ambiguous and difficult to predict. Due to this uncertainty, a more expansive exploration was followed, resulting in a broad scope for the systematic literature review. This relatively broad scope has been refined due to the inclusion/exclusion criteria we have selected and the search strategy we followed in our research. As a result, it is not feasible to claim that this review includes every article that leverages more than one relevance factor (defined in Section 2.1.1) for multidimensional relevance estimation. Nonetheless, our study offers a selection of articles that touch upon diverse knowledge domains and different search tasks to provide a comprehensive summary of research surrounding this topic.

Since a precise number of papers relevant to the studied topic is indeterminate, it is challenging to assess the extent to which the included studies cover the whole population. Despite this limitation, we have endeavored to ensure that our review captures a broad and representative spectrum of the available literature. After securing our final set of included studies and examining a substantial portion of them, we conducted targeted searches on Google Scholar. These searches were focused on specific research domains (for instance, the medical domain) and particular relevance dimensions (such as credibility). We then reviewed the results from these targeted searches. This procedure was replicated across domains and relevance dimensions to verify that we had identified all essential studies for our review study. By doing that, we encountered studies found in our prior searches and were subsequently either included or excluded from our review. We considered that a good indication of coverage and proceeded with our analysis. Nonetheless, future research on specific knowledge domains, particularly those underrepresented in our review, like mobile and geographic search, could uncover additional pertinent studies.

Another limitation is related to the application of the coding schema. While the schema was straightforward to apply for specific attributes of the paper (such as publication year and affiliations), its application became more subjective for other aspects, like those related to the relevance factors. As a result, studies that provided a formal definition were represented more lucidly than those that did not. Additionally, there were instances where papers did not comprehensively detail the tools and methodologies they utilized. This lack of full disclosure posed challenges in interpreting and conveying their findings.

In systematic reviews, it is common to encounter publication bias. Due to the uncertain breadth and depth of our review's outcomes, we limited our search to peer-reviewed publications, amplifying this bias. It is noteworthy that other studies not subjected to this peer review criterion might offer valuable insights. Therefore, we recommend that interested researchers and practitioners consult the tracks listed in Table 3.7 to obtain a broader perspective on the reviewed topic.

3.6 Conclusion

In our systematic review, we analyzed 70 studies to explore the methods scholars have employed in multidimensional relevance estimation within the field of Information Retrieval. The multidimensional nature of relevance is complex and diversely conceptualized across domains. This complexity, coupled with the variety of terminologies and methodologies, has presented challenges in standardizing definitions and operationalizations. To bring clarity, we proposed a structured formulation emphasizing clear definitions and transparent operational relationships between relevance factors. This approach promotes consistent future research. The recent advent of LLMs amplifies the timely significance to our review. With their advanced relevance labeling capabilities, LLMs offer potential solutions to challenges in creating benchmark collections for multidimensional relevance, a task traditionally reliant on human annotators. However, the success of LLMs relies on crafting precise prompts. Our review, especially the detailed definitions of relevance factors, can guide this prompt creation process. Moreover, LLMs' text comprehension proficiency suggests a plausible ability to estimate topical relevance scores in the future. This development signifies a potential shift in Information Retrieval, moving from models focused solely on topical relevance to those embracing multidimensional relevance. By considering user, task, and domain characteristics, such models mark a promising future direction, as they might offer a closer approximation to *user relevance*. In summary, our review sheds light on the complexities of multidimensional relevance, proposes a pathway for future research, and underscores the transformative potential of the domain due to the advancement of LLMs.

Chapter 4

Clinical Trials Retrieval

This chapter introduces the task of clinical trials retrieval, which is the professional search task to which we implement our research. Our work is oriented around two principal directions: processing unstructured patient information from Electronic Health Records (EHRs) and enhancing retrieval performance using DtMRF and Neural-DtMRF. This chapter reviews existing literature in these areas, identifying research limitations our study seeks to address. Additionally, the chapter introduces the benchmark collections employed to assess the efficacy of our proposed approaches.

4.1 Introduction

Clinical trials are the established scientific approach for assessing the effectiveness of new biological agents, drugs, devices, or procedures in preventing or treating diseases in human populations [Fleming and DeMets, 1996]. Recruiting a sufficient number of patients to participate in a clinical trial is one of the main encountered challenges, as it not only causes delays and leads to trials' failure, but also compromises the validity of the conducted studies by limiting their generalizability [Gul and Ali, 2010, Penberthy et al., 2012].

The process of enrolling participants in clinical trials is intricate and comprises various steps, as reported by Jain et al. [2019]. Typically, the process is initiated when a healthcare provider actively searches for an appropriate clinical trial for a certain patient. The healthcare provider searches by utilizing the patient's clinical or genomic data derived from EHRs, which encompasses laboratory reports, radiology reports, or clinical notes [Landolsi et al., 2023]. The search yields a list of clinical trials where the studied patient may meet their eligibility requirements. At this point, human effort is necessary to refine the potential trial list and generate meaningful trial recommendations for the considered patient. Once a suitable trial is identified, the results are shared with the patient's provider, who decides whether to proceed with this trial and requests a detailed patient prescreening to be performed. After completing the prescreening process, the patient is contacted and offered the option to enroll in the trial. Patients who accept undergo a final screening to evaluate their eligibility, followed by a consenting process. If the screening is successful and the patient consents, s/he is officially enrolled in the clinical trial.

EHRs have emerged as the preferred and effective approach for identifying and enrolling participants in clinical trials, complemented by strategies like reaching out to past participants and reviewing upcoming clinic schedules [Hersh, 2007, O'Brien et al., 2021]. Empirical evidence indicates that incorporating EHR-based patient-screening in this task's workflow enhances recruitment rates [Effoe et al., 2016]. However, according to O'Brien et al. [2021], the lack of research-focused EHR-based modules restricts the optimal utilization of EHRs in recruitment efforts. The two systematic reviews by Von Itzstein et al. [2021] and Chow et al. [2023] aim to comprehensively explore end-to-end applications of artificial intelligence in clinical trial enrollment by analyzing various research studies. Their findings highlight the widespread utilization of Natural Language Processing (NLP) for extracting information from unstructured EHRs, the significant time savings achieved compared to manual screening methods, and the critical considerations of maintaining patient confidentiality and data security. Another study further underscores the importance of leveraging NLP in clinical trial recruitment while emphasizing the need to assess their real-world adoption and effectiveness [Idnay et al., 2021]. These insights further reinforce the premise that EHRs, with their abundance of patient-related information, including clinical narratives such as clinical notes, hold great potential for facilitating clinical trial enrollment. However, as their primary purpose is to support clinical care rather than clinical trial enrollment, these narratives can be lengthy, unstructured, or contain several textual peculiarities, such as medical jargon and abbreviations. An example of a synthetic patient's admission note that contains several patient-related information can be seen in Figure 4.1.

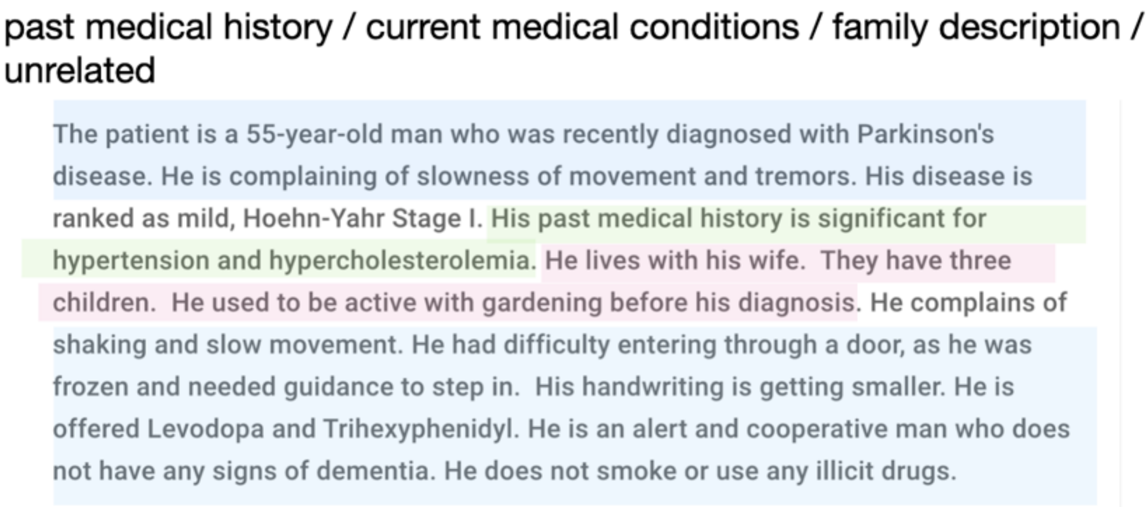

Figure 4.1: A sample admission statement in an EHR. The text above the figure outlines the diverse information available within clinical notes.

To handle these textual characteristics, several approaches in the literature exploit rule-based, hybrid, or neural-based NLP methods to extract valuable information from clinical narratives as highlighted by Landolsi et al. [2023], Link et al. [2022], Hobensack et al. [2023]. In Section 4.2.1, we provide a comprehensive review of relevant research studies.

Our research focuses on the initial step of the workflow presented above, where healthcare providers (i.e. expert users) actively search for eligible clinical trials by leveraging a patient's EHR. Hereafter we refer to this search task as *eligibility screening*. To aid this process, a written description of each clinical trial is being published in dedicated sites, such as the ClinicalTrials.gov website[1], usually following a specific document structure. This structure often contains the title and the description of a trial, a brief summary, and a section dedicated to the desired participant characteristics (gender, age, clinical conditions, etc.), i.e. the eligibility section. Considering an expert user, the process of eligibility screening can be described as follows [Ni et al., 2019]: initially, the process includes reviewing a patient's electronic health record aiming at identifying important aspects, e.g. demographics, clinical conditions, among others. For the determination of a patient's eligibility, the expert user compares the patient's information with a trial's recruitment requirements, mentioned within its eligibility section. Specifically, a patient's eligibility to a trial is determined by the trial's criteria (inclusion and exclusion), which are parts of the content of the document's eligibility section. These eligibility criteria are usually mentioned in a semi-structured format, i.e. an unordered list, as depicted in Figure 4.2. An eligible trial is one for which a patient covers all of its inclusion criteria and, simultaneously, none of its exclusion criteria. However, as both a patient's health record and a trial's requirements are mentioned in an unstructured or semi-structured format the process can not be perceived as a simple text matching task. Consequently, the task aimed at finding patients eligible for clinical trials is a complex retrieval task. As clinical trials retrieval, we refer to the retrieval task in which, given an admission note that contains patient-related information (i.e. query), the search engine aims to retrieve clinical trials (i.e. document collection) in which the patient can participate (i.e. retrieve eligible clinical trials).

Clinical trials retrieval differs from ad-hoc retrieval tasks, as patients' relevance to some document parts, those related to the exclusion criteria, negatively influences its utility. Specifically, treating this task as an ad-hoc task, the end-user might be presented with topically-relevant trials for which the considered patient might not be eligible to participate. As a result, the expert user is still committed to manually reading the appropriate document sections of the top-ranked documents to determine a patient's eligibility. An action that, as outlined by Soboroff [2022],

[1] Database of privately and publicly funded clinical studies conducted around the world, accessed on 24/7/23.

Criteria

Inclusion Criteria:

- Male or female adults ages 18-40 or of 65 and or older at the time of enrollment
- Eligible to receive Fluad® (MF59Flu) or Fluzone® (HDFlu) if age 65 or older
- No history of anaphylactic reaction to gelatin, neomycin, or other vaccine component
- Not pregnant
- No immunosuppression or immunodeficiency
- No acute illness at time of vaccination
- Determined by medical history and clinical judgment to be eligible for the study, by being generally healthy, with no autoimmune or immunosuppressive conditions and having stable current medical conditions (subjects with preexisting stable disease, defined as disease not requiring significant change in therapy or hospitalization for worsening disease 12 weeks before receipt of study vaccine, will be eligible. A change in dose or therapy within a category (e.g., change from one nonsteroidal anti-inflammatory drug to another) is allowed. A change to a new therapy category (e.g., surgery or addition of a new pharmacological class) is only allowed if it is not caused by worsening disease. A change to a new therapy category caused by worsening disease is considered significant and therefore ineligible for enrollment.
- Patients with diabetes mellitus are eligible for inclusion if they have had a hemoglobin A1c measurement of <8.0 within the past 6 months prior to enrollment. These hemoglobin A1c measurements are recommended at least twice yearly by the American Diabetes Association (ADA), and the target levels here are representative of the goals of the ADA. These hemoglobin A1c levels will ensure that these participants have good glycemic control. (American Diabetes Association. American Diabetes Association Position Statement: Standards of Medical Care in Diabetes- 2015. Diabetes Care 2015;38(Suppl. 1): S1-S94)
- Able to follow study procedures in the opinion of the investigator
- Expected to be available for the duration of the study
- Weighs >110 lbs

Exclusion Criteria:

- Known or suspected immunodeficiency or receiving treatment with immunosuppressive therapy including cytotoxic agents or systemic corticosteroids (e.g., for cancer, HIV, or autoimmune disease). If systemic corticosteroids have been administered short term for treatment of an acute illness, subjects will be included if corticosteroid therapy (inhaled, intranasal, and intra-articular corticosteroid therapy is permitted) has been discontinued for at least 30 days.
- Serious chronic medical conditions including metastatic malignancy, severe chronic obstructive pulmonary disease requiring supplemental oxygen, end-stage renal disease with or without dialysis, clinically unstable cardiac disease, or any other disorder that, in the investigator's opinion, precludes the subject from participating in the study. Diabetic patients will be excluded if they do not have a hemoglobin A1c measurement within the past 6 months or if they had a hemoglobin A1c measurement of an A1c >8.0

Figure 4.2: A clinical trial's inclusion and exclusion criteria, mentioned in a semi-structured format.

makes users dissatisfied as they are shown trials that they are explicitly excluded from. Despite the fact that several studies in the literature follow the aforementioned approach, [Koopman and Zuccon, 2016, Agosti et al., 2019, Rybinski et al., 2021], both the proposed DtMRF and the Neural-DtMRF are designed to overcome this problem, as we will explain in Chapter 7.

To conclude, our research proposes solutions that can potentially fully automate the process of eligibility screening for clinical trials or, at least, significantly reduce the required human effort. To achieve that, we focus our endeavors into two research directions: Firstly, we focus on methods that can be used to extract essential information from the unstructured information contained in EHRs. We leverage state-of-the-art transformer-based methods and combine them with well-known rule based approaches, for entity extraction and semantic meaning disambiguation. Additionally, we investigate the usage of a large language model, namely the GPT-3.5 model, commonly referred to as ChatGPT[2], to extract patient-related

[2] Introducing ChatGPT, accessed on 12/4/2023.

information from clinical narratives. In this approach, a clinical note that contains various patient-related information (e.g. underlying medical problem, family history, patient's demographics) is processed through ChatGPT aiming to automatically synthesize queries for searching eligible clinical trials for the considered patient. Secondly, we aim to improve the retrieval process in this task by leveraging the task's characteristics, especially by incorporating the negative influence of relevance to a trials exclusion criteria in the retrieval process. Towards this direction, we leverage DtMRF and Neural-DtMRF model and investigate the degree to which they improve retrieval effectiveness and interpretability. Additionally, we leverage LLMs and explore whether these models can be used to estimate the eligibility of a patient to a given clinical trials. Following this introduction, the following section provides a literature review centered on essential contributions and methodologies in medical information extraction and clinical trials retrieval.

4.2 Literature Review

This section explores two core research areas central to our study: medical information extraction and relevance estimation in clinical trials retrieval. It highlights the progression from conventional rule-based techniques to the recent integration of large language models in medical information extraction. It also emphasizes using LLMs in query generation to improve retrieval efficiency. Regarding Clinical Trials Retrieval, it analyzes studies that leverage retrieval approaches to enhance eligibility screening for clinical trials. It underscores the different formulations of this task in the literature and the state-of-the-art approaches. This review establishes the context and groundwork for our forthcoming discussions and experimentation in Chapters 6 and 7.

4.2.1 Medical Information Extraction

This section reviews research conducted in the field of medical information extraction, highlighting the transition from traditional rule-based approaches to the adoption of LLMs. As our research aims to extract patient-related information to enhance retrieval performance, we also discuss relevant studies that utilize PLMs/LLMs as an intermediary step for query generation across several retrieval tasks in the literature.

4.2.1.1 Medical Information Extraction: From Rule-based Approaches to LLMs

Medical information extraction is a sub-field of natural language processing that focuses on extracting essential information from unstructured clinical text, such as EHRs, clinical notes, medical literature, and patient narratives. The primary goal is to improve medical decision-making, disease surveillance, clinical research, and personalized patient care. To that aim, research interest has been focused on NLP tasks such as medical named entity recognition, relation extraction, event extraction, temporal information extraction, and negation detection, among others [Landolsi et al., 2023, Navarro et al., 2023, Linna and Jr., 2022, Zaikis and Vlahavas, 2021]. The field of medical information extraction has experienced substantial development over the years, transitioning from rule-based approaches to supervised machine learning methods, progressing even further with the adoption of deep neural networks, transformer-based models, and, ultimately, domain-specific LLMs. Early IE systems relied on manually crafted rules to identify and extract relevant clinical information. Despite their limitations in scalability, for many tasks, especially involving extraction of numbers, acceptable performance was often achieved with relatively simple rule-based approaches [Kreimeyer et al., 2017, Wang et al., 2018, Magoc et al., 2023]. A representative approach, the ConText algorithm introduced by Harkema et al. [2009], can be used to identify negated content in clinical notes, among other functionalities. The emergence of machine learning techniques along with the availability of domain-specific data sets accessible to the research community with a data-use agreement (e.g. i2b2[3] and MIMIC II [Saeed et al., 2011]), allowed more accurate and robust extraction of medical entities and their relationships [Jiang et al., 2011]. With the advent of deep learning, models based on word embeddings that leverage Recurrent neural networks (RNNs) enabled more effective representation of complex medical language [Wu et al., 2020]. Nonetheless, the limited availability of datasets has restricted the presence of deep learning approaches in non-English languages, such as French, resulting in a comparatively smaller adoption [Wu et al., 2020]. Transformer architectures, such as BERT [Devlin et al., 2019] achieved state-of-the-art performance on many generic NLP tasks, and following it, many clinical and biomedical variations, like ClinicalBERT [Alsentzer et al., 2019], SciBERT [Beltagy et al., 2019], among others, have been proven effective in domain-specific NLP tasks [Landolsi et al., 2023, Hahn and Oleynik,

[3] National NLP Clinical Challenges, accessed on 12/4/2023.

2020].

The introduction of large language models like GPT-3, PaLM [Chowdhery et al., 2022], and GPT-4 [OpenAI, 2023], among others, has revolutionized the field of natural language processing [Fan et al., 2023]. Their pre-trained knowledge and fine-tuning capabilities have facilitated substantial progress in various NLP tasks, such as information extraction, summarization, and question-answering. Zhao et al. [2023] provide a detailed overview of four aspects of LLMs, namely the pre-training process (i.e. data collection, architectural design, and model training), adaptation tuning (i.e. effectively tune pre-trained LLMs), utilization (i.e. usage of LLMs to solve downstream tasks), and evaluation. They highlight the main issues of LLMs, such as the problem of hallucination generation [Bang et al., 2023], the inability to address tasks that require knowledge beyond the training data (i.e. knowledge recency), and the inconsistency in the provided answers, among others. Finally, they discuss the potential risks and capabilities of LLMs that may arise within the medical domain.

In the medical domain, general- purpose LLMs have been employed to analyze EHRs and clinical notes (i.e. unstructured clinical text) to aid the diagnostic process and offer treatment suggestions, among others [Fan et al., 2023]. In addition, Liu et al. [2023b] investigate the usage of ChatGPT[2] and GPT-4 in another task, i.e. medical text anonymization. Their empirical evaluation showed that both of these models (in a zero-shot setting) are capable of de-identifying medical data compared to ClinicalBERT. Regarding the employed prompts, they found that explicit prompt design that contains a well-written description of the desired output, clearly defines the task, and provides concrete examples, leads to better performance. As highlighted by Zhao et al. [2023], answer inconsistency is a significant issue of generative LLMs. To solve this problem, Chuang et al. [2023] proposed SPeC, a model-agnostic soft prompt-based calibration pipeline that addresses the issue of output variance in clinical note summarization. By employing soft prompts along with discrete prompts, the proposed method effectively mitigates summarization variance while still harnessing the benefits of prompt-based summarization across three LLMs. Since the introduction of LLMs and their adoption to solve specific NLP tasks in the health domain (biomedical and clinical), several studies have investigated whether these general-purpose LLMs are proper tools or if smaller, pre-trained models on domain-specific NLP tasks should be used instead [Lehman et al., 2023, Agrawal et al., 2022, Gutierrez et al., 2022, Moradi et al., 2021, Hu et al., 2023, Tang et al., 2023]. Now, we provide an overview of the aforementioned studies, by focusing on the investigated NLP task, the compared LLMs, and pre-trained/fine-tuned LMs. In addition, we comment on the selected prompting, as it is highly related to the effectiveness of LLMs [Zhao et al., 2023, Perez et al., 2021].

Lehman et al. [2023] investigated whether LLMs can yield better effectiveness across three medical NLP tasks in clinical information extraction; two were related to the multi-label classification of clinical sentences, and one was related to medical question-answering. To that aim, they compared 12 language models, i.e. BioClinRoBERTa [Lewis et al., 2020], GatorTron [Yang et al., 2022b] (an LLM trained on de-identified clinical texts), PubMedGPT (which is now renamed to BioMedLM[4], GPT-3 and T5 [Raffel et al., 2020]. Regarding the prompts used with GPT-3, a single prompt was employed to simultaneously instruct the model to generate predictions for all labels. Their findings suggest that models fine-tuned on all available data, particularly BioClinRoBERTa and GatorTron, significantly outperform any in-context learning approach for the selected NLP tasks. Nonetheless, the authors did not employ ChatGPT in their evaluation as it is unavailable via a HIPAA-certified API. Similar conclusions have been drawn from the empirical evaluation of Gutierrez et al. [2022] in biomedical information extraction. In their work, the authors compare the few-shot performance of GPT-3 in-context learning with fine-tuning smaller PLMs, namely PubMedBERT-base [Gutierrez et al., 2022], BioBERT-large [Lee et al., 2020a] and RoBERTa-large [Liu et al., 2019b]. They investigate two biomedical NLP tasks, i.e. named entity recognition and relation extraction, across eight datasets, aiming to extract diseases, chemicals, medical concepts, and genes, identify drug-to-drug and chemical-to-protein interactions, and associate genes with diseases. The authors paid particular attention to the in-context learning process of GPT-3 by following a systematic and task-agnostic process for constructing the prompts. In detail, they constructed prompts based on the *True Few-Shot* training process introduced by Perez et al. [2021], aiming to avoid plausible biases introduced in the model due to prompt selection on a large validation set. Their evaluation suggested that GPT-3 significantly underperforms compared to the employed fine-tuned PLMs. Moradi et al. [2021] investigated whether GPT-3 following a few-shot in-context learning setting outperforms BioBERT on various biomedical and clinical NLP tasks. The prompts associated with GPT-3 contained a description of the task and a few examples,

[4] BioMedLM, accessed on 12/4/2023.

instructing the model on formulating its response. Their findings highlight GPT-3's inability to compete with BioBERT, especially in tasks that require calculating a semantic similarity score between sentences. BioGPT, introduced by Luo et al. [2022], achieved the highest performance compared to GPT-2 and several other domain-specific PLMs when evaluated on six biomedical NLP tasks such as relation extraction, question answering, document classification and text generation. BioGPT is a domain-specific LLM with the same model architecture as GPT-2, and it is pre-trained on a 15M PubMed abstracts corpus.

The findings presented in the aforementioned studies [Moradi et al., 2021, Gutierrez et al., 2022, Lehman et al., 2023, Luo et al., 2022] suggest that BioGPT achieved state-of-the-art performance compared to pre-trained and fine-tuned PLMs in biomedical NLP tasks, while GPT-3 based on in-context learning did not yield performance improvements.

Regarding clinical IE, the work of Hu et al. [2023] explored the potential of using ChatGPT for clinical named entity recognition in a zero-shot setting. Their results showed that ChatGPT surpassed GPT-3 in terms of F1 scores for both exact- and relaxed-matching on an annotated subset of transcribed medical reports, i.e. MTSamples[5]. However, ChatGPT under performed compared to BioClinicalBERT[6] fine-tuned on the i2b2 2010 dataset [Uzuner et al., 2011]. The authors employed two types of prompts; for example, to extract medical problems, the first prompt was: *"Extract without rephrasing all medical problem entities from the following note in a list format:"*; the second, which led to better performance, was: *"Extract without rephrasing all medical treatment, medical procedure, medical intervention, medication, drug entities from the following note in a list format:"*. The performed error analysis revealed that ChatGPT might attempt to infer or summarize information or rephrase terms, even though it has been explicitly instructed not to. Lastly, the authors note that ChatGPT's performance might have been underestimated due to minor changes in its response. Agrawal et al. [2022] investigated GPT-3's and InstructGPT's [Ouyang et al., 2022] ability to perform zero- and few-shot information extraction from clinical text. To that aim, they compared their performance to various LM models. They showed that GPT-3 performs well in clinical NLP over diverse tasks, namely abbreviation expansion, coreference resolution, extraction of biomedical evidence, medication status, and medication

[5] MTSamples, accessed on 14/4/2023.
[6] Bio+Clinical BERT model, accessed on 14/4/2023.

attribute. Based on the employed prompts, findings show that a *guided prompt design* leads to performance improvements. Similarly to Hu et al. [2023], they found that GPT-3's outputs did not always match the annotated text (required output) at the token level, suggesting that its performance could have been higher. In addition, they highlighted GPT-3's bias towards responding to a question, i.e. extracting a piece of information, even though the requested entity does not exist in the given text. All in all, the previous studies showed that ChatGPT and GPT-3, two general-purpose LLMs, have the potential to perform accurate IE in the clinical domain (occasionally even outperforming domain-specific PLMs) and highlighted some potential issues related to models' response behavior (strong dependence on the created prompts) and their evaluation (mainly due to token-level mismatch).

Yang et al. [2022a] developed a large clinical language model from scratch, namely GatorTron. The model has adopted the BERT architecture with three different settings varying from the base model with 345M parameters to the large model with 8.9B parameters. The model has been evaluated across five clinical NLP tasks: clinical concept extraction, medical relation extraction, semantic textual similarity, natural language inference, and medical question answering. Empirical findings show that GatorTron outperforms previous transformer models, such as BioBERT and ClinicalBERT, across all NLP tasks. However, as the authors mention, GatorTron achieved remarkable improvements for complex NLP tasks such as natural language inference and medical question answering, but shows only marginal improvements in simpler tasks such as clinical concept extraction and medical relation extraction.

To conclude, LLMs have shown great potential for medical information extraction. As the empirical evidence suggests, in the biomedical domain, general-purpose LLMs, like GPT-2, GPT-3, and ChatGPT, fail to reach the effectiveness of PLMs in essential NLP tasks. LLMs trained from scratch on domain-specific data, such as BioGPT, performed better than previous state-of-the-art approaches. In contrast, even general-purpose LLMs have improved performance for clinical information extraction over the previous SoA models like ClinicalBERT. However, in both domains, the LLMs performance is highly related to prompt formulation and the models have been found to be very sensitive to that. In addition, it has been reported that it might be the case that the performance of LLMs might have been underestimated due to their tendency to rephrase extracted tokens in their responses. A limitation that has been identified that may play a crucial role is that in the majority of the works, the prompts followed a task-agnostic approach. Moreover, other model parameters, like ChatGPT's parameter related to the system's role, have yet to be fully investigated in the studies mentioned above. Lastly, in the clinical domain, it has been found that LLMs like Gatortron and LMs like BioBERT perform similarly in simple entity extraction tasks.

4.2.1.2 LLMs for Query Generation to Enhance Retrieval

The intersection between Information Retrieval and Natural Language Processing has become more prominent in recent years. Some applications of LLMs in the field of IR involve the creation of synthetic datasets tailored to specific domains and tasks. Specifically, a recent study has leveraged LLMs to generate synthetic training datasets for IR tasks [Bonifacio et al., 2022]. The reported findings suggest that models, fine-tuned exclusively on synthetic datasets, surpass standard approaches, including BM25, as well as recent self-supervised dense retrieval approaches. Similarly, Saad-Falcon et al. [2023] proposed UDAPDR, a strategy that uses synthetic queries created using generative models, such as GPT-3, to train multiple passage re-rankers on queries for target domain passages. The reported evaluation on three datasets showed that UDAPDR could improve zero-shot retrieval accuracy on new domains without using labeled training examples.

Another example of the synergy of NLP and IR is query generation or expansion, which is one of our research focuses. Specifically, due to their vast accumulated knowledge, LLMs might be capable of paraphrasing or expanding queries and improving search quality, especially for standard retrieval models that rely on bag-of-words and are commonly used as first-stage retrievers. In this setting, a query can be input into an LLM as a prompt, accompanied by task-specific instructions, allowing the model to generate contextually relevant and accurate responses (i.e. reformulated query) tailored to both the information needed and the task to be performed. This research direction has been investigated by Claveau [2021]; the proposed approach improves information retrieval using GPT-2 to generate multiple texts based on a given query. The generated texts are concatenated to create an expanded query, providing broad coverage of vocabulary that captures synonyms, hypernyms, and other linguistic relations. Then, the expanded query is used as an input in an IR system. In the described process, the only online task is text generation, while model training and fine-tuning are performed offline.

The experiments conducted on several datasets showed the effectiveness of this approach over other query expansion methods, such as RM3. However, only GPT-2 has been employed in this work, although GPT-3 achieves higher performance in various tasks. As stated in the paper, the main reason is GPT-3's difficulty in engineering prompts to perform the expected generation task. Wang et al. [2023a], proposed query2doc, a query expansion approach that can improve sparse and dense retrieval systems. It leverages *text-davinci-003* to generate pseudo-documents using few-shot prompting and expands the query with the generated pseudo-documents, similarly to [Claveau, 2021]. Experimental results show that query2doc improves the performance of BM25 by 3% to 15% on ad-hoc IR datasets, such as MS-MARCO, without any model fine-tuning. Additionally, the method benefits state-of-the-art dense retrievers in terms of both in-domain and out-of-domain results. Prieto-Chavana et al. [2023] analyze various conditional text generation techniques and compare their performance to rule-based baselines, aiming to understand whether one can automatically formulate search queries based on factual statements that are similar to those formulated by human experts. To that aim, they introduce a dataset for fact-checking and evidence collection. They establish that similarity to human-created search queries is a valuable indicator of the effectiveness of automatically generated queries in retrieving the same evidence. However, they also note that there can be cases where seemingly different search queries may result in collecting the same evidence.

Lee et al. [2023] utilize an LLM and the text from titles and abstracts of research papers to generate keywords for a research paper. Their analysis, suggests that an LLM has the capability to automatically generate keywords, showcasing its potential in this task. Lastly, the work by Wang et al. [2023b] has several commonalities with the approach we propose in this paper. In detail, the authors also leverage ChatGPT and instruct it to create Boolean queries that enhance retrieval effectiveness. They focus on the task of systematic literature review aiming at retrieving studies related to the review topic. Similar to one of our approaches, they also develop various prompts with increasing complexity, including prompts containing example Boolean queries and guided prompts. Their prompts are designed for two tasks, i.e. query generation and refinement. Their findings showed that when ChatGPT was instructed to include MeSH (Medical Subject Headings) terms for some queries, those MeSH terms were incorrect. MeSH is the National Library of Medicine's controlled vocabulary thesaurus, while MeSH terms are biomedical- and health-related terms, including any variant spellings and plurals[7]. Besides this limitation, their empirical evaluation suggests that the generated queries result in high search precision, although at the expense of recall. Also, this study highlighted the ability of ChatGPT to comprehend detailed instructions and create queries with a high level of accuracy, particularly in cases where time is limited and a compromise between precision and recall is acceptable, which is exactly the case in the task studied in our research, i.e. clinical trials retrieval. Therefore, ChatGPT has the potential to serve as a standalone solution or as a complementary component in conjunction with existing semantic-driven approaches for boolean query formalization [Pourreza and Ensan, 2023].

4.2.2 Relevance Estimation in Clinical Trials Retrieval

In the literature, the task of eligibility screening for clinical trials has been mainly explored from two distinct retrieval perspectives. Specifically, in the TREC Med-Track [Voorhees and Hersh, 2012] the inclusion criteria of a trial were used as a query to retrieve eligible patients from a collection of patient health records (trial-to-patient retrieval perspective), while in the TREC Precision Medicine (PM) Track [Roberts et al., 2017, 2018, 2019], the problem formulation was the exact opposite (patient-to-trial retrieval perspective). Moreover, the TREC Clinical Trials 2021[8] and 2022[9] tracks follow the patient-to-trial retrieval formulation by introducing a verbose query representation in the form of a patient's admission note, which constitutes the main difference from the PM track [Soboroff, 2022].

Additional retrieval perspectives have been explored in other research works. For instance, Koopman and Zuccon [2021] explore this task from a cohort-based retrieval perspective, while Liu et al. [2019a] created an IR system that initiates a question-answering interactive session with its end-user to eliminate those trials for which the considered patient is explicitly excluded. Even in this interactive IR system, an initial retrieval step is necessary to reduce the number of the considered clinical trials. Finally, Rybinski et al. [2021] design an end-to-end retrieval system based on an standard retrieval model and a BERT-based neural re-ranker.

Within the literature, several retrieval approaches have been put forth to tackle

[7] National Library of Medicine, accessed on 22/04/2023.
[8] TREC Clinical Trials 2021 Track, accessed on 20/04/2023.
[9] TREC Clinical Trials 2022 Track, accessed on 20/04/2023.

eligibility screening. These either take on a patient-to-trial direction [Koopman and Zuccon, 2016, Agosti et al., 2019, Rybinski et al., 2020] or a trial-to-patient viewpoint [Limsopatham et al., 2014]. It is worth mentioning that Limsopatham et al. [2014] explicitly consider the importance of a trial's inclusion criteria, by introducing a retrieval approach for modeling the mixture of the relevance probability towards the query (trial's inclusion criteria) and the likelihood that a patient's EHR is relevant to these criteria. In contrast, studies that follow the patient-to-trial retrieval perspective, do not explicitly consider a trial's eligibility criteria during retrieval, but only employ some sort of filtering based on a trial's demographic and gender requirements.

Lastly, a considerable amount of literature has been published within the TREC Clinical Trials 2021 Track[10] and the 2022[11]. Our detailed analysis of TREC's publications has highlighted some common practices among the participating teams. It has been found that most of the submitted works filter out (i.e. remove from the final ranking) those clinical trials for which the patient does not meet the required demographic constraints (gender and age). However, two additional recruitment conditions, i.e. recruitment status (clinical trials recruitment phase has a specified time window) and location (many trials enroll patients at specific locations), have not been considered in the TREC initiative [Soboroff, 2022]. As a result, the submitted works did not consider these aspects, although there are important in real-world scenarios. Other studies employ some unsupervised query pre-processing or expansion techniques, such as KeyBERT [Grootendorst, 2020]. Also, the proposed systems in several works extract conditions, medical procedures or drugs related to a patient and expand them using, for instance, the Unified Medical Language System[12].

The top-performing retrieval approach in TREC 2021 [Pradeep et al., 2022] relies on a multi-stage retrieval setting that consists of an initial neural query synthesis step that leads to forty distinct query representations. Those representations are used for retrieval, and the obtained document rankings are fused. These initial retrieval runs leverage the BM25 model. Finally, a two-stage neural re-ranking pipeline trained on clinical trial matching is exploited to create the final ranking. The most successful approach in the TREC 2022 Clinical Trials track, namely

[10]The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings, accessed on 24/7/23.
[11]The Thirty-first Text REtrieval Conference (TREC 2022) Proceedings, accessed on 24/7/23.
[12]UMLS Metathesaurus Browser, accessed on 31/04/2023.

frocchio_monot5_e by team h2oloo, employs the Mono-T5 model to re-rank an initially retrieved set of clinical trials[13]. However, detailed information about the experimental details of this approach has not been provided. Notably, the TREC 2021 top-performing approach relies on an initial retrieval step that employs the BM25 model with neural query generation and query expansion; it is also plausible that this is the case for the TREC 2022 approach as it has been submitted by the same team and exploits the same model. Although these approaches exhibit strong retrieval performance, they encounter challenges in terms of cost—specifically, fine-tuning and maintaining the Mono-T5 model—as well as interpretability concerning the resulting document ranking.

Over the years, several scholars have conducted systematic literature reviews related to medical and clinical IR [Tamine and Goeuriot, 2022, Hersh et al., 2020, Himani and Vaidehi, 2018], including clinical trials retrieval [Sivarajkumar et al., 2023]. Here, we briefly mention their scope, starting from works that offer a broad overview of health informatics, and concluding with those that investigate the specific research area. To begin with, William Hersh, in the fourth edition of his book entitled "Information Retrieval: A Biomedical and Health Perspective," overviews IR systems under the scope of bio-medicine and health domains [Hersh et al., 2020]. Another work by Himani and Vaidehi [2018] analyzes publications and tools by focusing on the diversity of possible medical users and common issues, such as the diverse user vocabulary. Tamine and Goeuriot [2022] conduct a literature review of semantic IR in the medical domain. After introducing the medical domain and the available data sources, the authors present an overview of the employed models and techniques before concluding their work by presenting open challenges and future research directions. Lastly, Sivarajkumar et al. [2023] focus on clinical IR, particularly methods, tools, and techniques that leverage free-text electronic health records. The authors included 184 research works published from 2012 to 2023 in their analysis. Their findings show that despite recent technological advancements, a significant amount of clinical IR systems rely on the BM25 model due to its efficient retrieval capability.

The main limitation of the majority of the proposed approaches is that they are based on traditional text matching, i.e. treat this task as an ad-hoc retrieval task. As a result, these approaches disregard the constraints imposed by a trial's exclusion criteria, i.e not to be present in the patient's clinical record. In addition, even

[13] Overview of the TREC 2022 Clinical Trials Track, accessed on 31/03/2023.

those that consider the exclusion criteria, rely on simple aggregation operators that do not fully capture the task characteristics. Both DtMRF and Neural-DtMRF account for this particular aspect by incorporating negative signals (patient's relevance to a trial's exclusion criteria) into the relevance estimation. Lastly, while numerous studies have delved into query reformulation techniques for enhanced retrieval performance, our experiments provide a comprehensive comparison of several methods, ranging from rule-based approaches to LLMs.

4.3 Benchmark Collections for Clinical Trials Retrieval

This section elaborates on the benchmark collections employed in our experiments presented in Chapters 6 and 7. Our empirical evaluations are performed on three publicly available benchmark collections. For the majority of our experiments we use the collections introduced in the TREC 2021 [Soboroff, 2022] and 2022[13] Clinical Trials tracks. These collections consists of 375,580 clinical trials originally published in the ClinicalTrials.gov website[1]. Hereafter we refer to them as *TREC 2021* and *TREC 2022*. The TREC 2021 collection has a total of 75 queries and the TREC 2022 has 50 queries, that have been created by individuals with medical training. Both the documents and the queries resemble these presented in Figures 4.2 and 4.1. The third collection we use is created by Koopman and Zuccon [2016], but it has a limited number of relevance assessments, which, as outlined by the authors, may lead to unreliable evaluations for new systems that greatly differ from those used to form the original pool. We refer to this collection as *Clinical*.

Regarding the relevance assessments in the collections, a clinical trial has been evaluated as *eligible*, *excludes*, and *not relevant* to a given clinical note, where *eligible* means that the patient can participate in it, *excludes* means that the patient is explicitly excluded, and *not relevant* which means that the patient does not have sufficient information to qualify for the trial. The following section presents a detailed analysis based on the TREC 2021 collection and queries.

4.3.1 Analysis based on Relevance Judgments

During the TREC 2021 clinical trials track, almost thirty-six thousand documents retrieved by 113 retrieval systems have been judged using shallow pooling [Soboroff, 2022]. From these, a considerable amount are judged as irrelevant (67.7%), 16.8% of them as excluded, and 15.5% as eligible. Regarding the number of documents that are judged as eligible across all of the 75 provided queries, the min, max, and average values are 6, 203, and 74, respectively. Regarding those judged as excluded, the min, max, average are 1, 226, and 80. Also, our analysis has shown that 11 queries have less than 25 eligible trials in the collection. These queries are presented in Table 4.1. Observing the variation of the relevance judgments across queries, one may conclude that in general, the task of eligibility screening is a complex search task, as the majority of the judged documents are irrelevant. However, for some queries, i.e patients, the task of finding eligible trials can be relatively easy, e.g. query 33 with 203 eligible trials; while for others can be hard, e.g. query 6 with only 6 eligible trials. The investigation of the underlying reasons

Table 4.1: Number of relevance judgments for the 11 queries with the fewer identified eligible clinical trials in the collection.

Query:	9	10	18	40	44	50	55	66	72	73	74
not relevant	234	478	224	468	262	374	350	439	320	375	436
excluded	201	36	192	2	189	9	16	37	76	46	2
eligible	13	11	14	11	6	14	23	11	15	20	11

related to the observed variation in the relevance judgments across the queries is out of the scope of our research. However, we mention here a few reasons that, in our opinion, may have lead to these variations.

To begin with, the observed variations in the relevance judgments can be related to the format and structure of the provided patient information in the corresponding queries. For instance, it might be hard to identify the correct patient's medical condition, so a retrieval system might retrieve irrelevant clinical trials. That can be the case for queries 10, 40, 50, 66, 72 and 74, for which the number of irrelevant trials is significantly higher than the number of trials judged as excluded and eligible, suggesting that only a few of the proposed retrieval systems were capable of capturing the primary patient condition.

Another reason can be related to the employed retrieval approaches in TREC 2021 that created the initial pool of judged documents. In particular, as seen in Table 4.1, queries 9, 18, and 44 have a relatively high number of irrelevant and excluded trials compared to eligible trials. For these topics, it seems that the proposed retrieval approaches could capture the primary medical condition mentioned in the queries; however, they were not capable of retrieving eligible trials. Of course, we can not rule out that the identified eligible trials may be the only ones in the collection. However, this is probably unlikely, as only a small portion of the proposed retrieval approaches in TREC[10] incorporate the negative influence of the exclusion criteria into their relevance estimation. Therefore, it is possible that for these queries, the pool of judged documents is biased towards irrelevant and "excluded" documents.

Finally, it is possible that these relevance judgments fully reflect the complexity of the studied search task, meaning that particular medical conditions are studied in many clinical trials while other conditions are in a few. Therefore, some patients can be quickly assigned to a clinical trial, while others cannot.

Based on our analysis, we conclude that more eligible trials may exist for some topics that, because of either of the reasons mentioned above, have not been identified. Moreover, the proposed retrieval methods explicitly consider the negative influence imposed by a trial's exclusion criteria, while the majority of the retrieval approaches used to create the document pool do not. Based on the aforementioned reasons, in several of our experiments we evaluate the retrieval performance based on condensed measures as proposed by Sakai [2007], as a way to deal with retrieved but unjudged documents.

Part II Conceptualizing the Decision Theoretic Framework

Chapter 5

A Decision Theoretic Framework for Multidimensional Relevance Estimation

This chapter introduces the proposed Decision-theoretic Multidimensional Relevance Framework (DtMRF); it describes its components and provides illustrative examples illustrating its IR application. DtMRF is designed as a framework in the sense that it may rely on distinct MADM methods; for this reason, we consider four instantiations of the framework, corresponding to the MADM methods presented in Section 2.2.2. The chapter concludes with the presentation of Neural-DtMRF, the neural extension of DtMRF aiming at predicting appropriate weights for relevance factors, and by drawing the main conclusions of our investigation.

5.1 Introduction

Information Retrieval is a process where a user motivated by a specific task and a related information need aims to identify, among a huge amount of information items, those that fulfill the information need. IR can be then considered a decision-making process in which a user (the decision-maker) assesses the utility of information items based on both objective and subjective factors such as topicality, domain knowledge, expertise, and timeliness (the relevance factors). Based on this interpretation, an IR system plays the role of an intermediary of the decision-maker, with the goal of assessing the utility of an information item to a user's need, by quantifying and aggregating the various objective or subjective relevance factors.

As highlighted in Chapter 3, topicality is the core relevance factor determining the utility of an information item (document) to a specific information need [Saracevic, 2016a, Li et al., 2017a], whereas it is also well recognized that several additional factors may be identified, which contribute to the utility (overall relevance) of an information item to a user, also depending on the considered topical domain and on the considered search task [Oroszlányová et al., 2017, van Opijnen and Santos, 2017, Wiggers et al., 2018]. Additionally, in a same domain, the importance of each relevance factor can be affected by the considered search task and by the user's role and context [Xu, 2007]. As a consequence, the above-mentioned relevance factors as well as their interactions, should be modeled by analyzing their possible trade-off in estimating the overall relevance of an information item with respect to the situation at hand.

Without loss of generality, we introduce two examples to illustrate what is outlined above. Let us assume to have a user engaging in a search task aiming to identify publications to be included in the "literature review" section of a paper. One could assess a document's utility by considering, for instance, three relevance factors, i.e topicality, scope, and understandability. Under this task, a returned document must be related to the query's topic, understandable by the user, while its scope can be relatively broad, meaning that the requested topic can be only partially discussed. Assume that the same user aims to identify publications for supporting the "methodology" section of the paper, which, for instance, is related to applying a specific algorithm. In this case, a retrieved document must be again related to the query's topic and understandable by the user; yet, it is preferred that a document has a narrow scope, meaning that the requested topic be fully discussed in the document.

Based on the previous observations, it is desirable that an IR system accounts for the following requirements. First, the system should estimate the degree to which a document satisfies the considered relevance criteria (factors), i.e be equipped with functions that, for each assessed document, produce a satisfaction (performance) score for each criterion. Second, the system should be able to manage the importance weights possibly associated with the considered criteria; in particular, it would be appreciable that the system be able to automatically compute such weights, based on preferences expressed by the user over the considered criteria or based on the characteristics of a considered search task. In the former example, in the relevance assessment process, a broad scope is more important than in the second example. Third, the system should account for the contribution and for the importance of each relevance factor in estimating the overall document's relevance, by aggregating the information carried by each relevance factor. Last but not least, in the aggregation process, it would be desirable that the system be able to account for the effect, either negative or positive, that a criterion may have in estimating the document's relevance. So, in the first example, the broader a document's scope is, the more useful it is; therefore, it should be ranked in the top positions (positive effect). On the other hand, in the second task, the broader a document's scope, the lower its ranking should be (negative effect). Summarizing, a retrieval process in complex search tasks can be modeled as a decision problem that accounts for several relevance criteria, associates them with importance weights, and further considers their positive or negative effect in determining a document's utility value.

As we have thoughtfully discussed in Chapter 3, several related studies introduce models to estimate multidimensional relevance with reference to a specific domain and task. Commonly, these models allow for an importance weight to be associated with each relevance criterion [da Costa Pereira et al., 2012a, Eickhoff et al., 2013b, Moulahi et al., 2014c, Pasi et al., 2019]. Nevertheless, as shown in the aforementioned examples, some relevance criteria may negatively affect a document's overall relevance under specific situations. However, none of the current approaches explicitly incorporates this aspect into a retrieval process, as they typically account only for criteria that positively impact relevance. Aim of our research is to explore the impact of incorporating all the above characteristics in the process of assessing relevance in complex search tasks; to this purpose we propose the Decision-theoretic Multidimensional Relevance Framework (DtMRF) that: (1) allows to associate importance weights with each relevance criterion and (2) incorporates the effect, either positive or negative, that each criterion has on the utility of an information item. DtMRF exploits a category of Multi-Criteria Decision-Making methods that incorporate the points mentioned above directly in the retrieval process, leading to four distinct instantiations. The mathematical properties and functionalities of these decision-theoretic methods will be explained in detail in Section 2.2.2.

In this chapter we delve into the theoretical implications of the DtMRF and its application in IR, aiming to provide answers to the following research questions:

(RQ1) How can multidimensional relevance estimation in IR be formulated as a decision-making problem that incorporates both positive and negative relevance factors?

(RQ2) How can the considered MCDM methods be leveraged for multidimensional relevance estimation?

(RQ3) In which ways does the proposed Decision-theoretic Multidimensional Relevance Framework lead to transparent and interpretable document rankings?

(RQ4) How does the inclusion of positive and negative relevance factors in the retrieval process affect retrieval behavior?

In addition to addressing the stated research questions, this chapter introduces the neural-based extension of our model, referred to as Neural-DtMRF.

5.2 Formulating IR as a Decision Theoretic Problem

To model multidimensional relevance estimation using DtMRF, the retrieval process must be formulated as a MADM problem, as presented in Section 2.2.1. In this formulation, the IR system is considered a decision-maker that aims to evaluate a finite set of alternatives, i.e documents, by considering several criteria, e.g. domain, user, and task-dependent relevance factors that influence a document's utility (overall relevance) in a particular situation. In addition, each criterion is associated with an importance weight and an objective, and an evaluation function is defined for each criterion to determine the degree to which a document satisfies that criterion. Ultimately, the documents are ranked based on their global performance score estimated by one of the four MADM methods introduced in Section 2.2.2. These components are formally denoted as follows:

- A collection of documents $D = \{d_1, d_2, \cdots, d_m\}$, where each document is considered as an alternative.
- A set of N pre-defined criteria each associated with a relevance factor that influences the utility of a document under a studied search task, denoted as $C = \{c_1, c_2, \cdots, c_N\}$.
- N importance weights $W = \{w_1, w_2, \cdots, w_N\}$ associated with the criteria, where $w_i \in [0, 1]$ and $\sum_{i=1}^{n} w_i = 1$.
- N objectives $O = \{o_1, o_2, \cdots, o_N\}$ associated with the criteria, where o_i corresponds to either a beneficial ($o_i = 1$) or a non-beneficial ($o_i = 0$) criterion.
- N evaluation functions F, one per criterion, that estimate a performance score that measures the degree to which each document satisfies that criterion.
- A MADM method M, employed to aggregate the criteria-related performance scores (assessed by the previously introduced evaluation functions) into a global performance score. This score corresponds to the document's estimated utility under the studied search task.

The following section presents further details related to the DtMRF components.

5.2.1 DtMRF Components

Selecting a set of appropriate criteria to estimate the utility of an information item is crucial, as those criteria should be representative of and appropriately describe a considered situation. A search task can be associated with a set of users (e.g. professionals in a given domain), and domain-dependent factors that affect a user's decision-making process under that search task. Over the years, numerous studies have investigated how users, or user groups, assess a document's utility under a particular search task. Those studies have identified various relevance factors such as topicality, understandability, reliability, scope, novelty, and also interest and habit that are user-related [Xu and Chen, 2006b, Li et al., 2017a]. Furthermore, the importance of the relevance factors mentioned above may differ in distinct knowledge domains, and also other relevance factors, e.g. credibility, may be considered [van Opijnen and Santos, 2017, Oroszlányová et al., 2017].

In addition, the notion of relevance is not only multidimensional but also dynamic, meaning that what users assess as useful to their situation changes through time and across tasks [Saracevic, 2016a]. To clarify, even though one may have identified which relevance factors have to be considered by the system's retrieval process, their importance and objective may change over time and by the user task. That is why the objectives and weights associated with the selected criteria play an essential role, and DtMRF allows their incorporation in the retrieval process.

One can exploit three possible directions to obtain the importance weights associated with each criterion. First, the importance weights can be estimated by employing a particular weighting method; for instance, the CRiteria Importance Through Inter-criteria Correlation (CRITIC) method [Diakoulaki et al., 1995], the Variability and Interdependencies of Criteria (VIC) method proposed by Akestoridis and Papapetrou [2019], or the Entropy Measure (EM) introduced by Deng et al. [2000] to mention a few. Another alternative involves a portion of the available benchmark collection created to tackle the particular search task, i.e. a training data set. In this case, the importance weights associated with the considered criteria can be obtained by optimizing specific evaluation measures on the considered training set. Lastly, the criteria weights can be modified during the search activity by the system's end-user through an appropriate user interface. This approach leads to an interpretable system that allows its end-users to completely control the retrieval process. All in all, selecting the most appropriate weighting approach depends on the context of the studied problem.

As outlined in the introductory example, the objectives of the criteria are highly affected by the undertaken task. As a result, the system should be designed to tackle various similar search tasks in a domain, i.e tasks for which the criteria objectives are pre-defined and constant during the search. Moreover, a system could be equipped with methods to predict the user's undertaken search task and adjust the criteria objectives accordingly. Especially in professional search, as end-users are often experienced, these objectives can be fully controlled during the retrieval process by using an appropriate interface.

Regarding evaluation functions, these estimate the degree (performance score) to which a document satisfies a considered criterion, independently from the other criteria. In the literature, the main relevance factors mentioned above, have been quantified using various functions [Li et al., 2017a]. For instance, the estimation of topicality is usually achieved using BM25 or other IR models, such as neural models. In addition, other studies in particular domains, such as the health domain, introduce custom functions to quantify the selected relevance factors [Grandis et al., 2019]. As long as the employed evaluation functions are strictly monotonic, selecting the most appropriate one is up to the system designer. Nonetheless, it is essential to outline here that these functions are strongly associated with the system's efficiency; i.e an accurate performance score estimation may be computationally expensive but lead to better retrieval effectiveness.

To conclude, the system designer, having studied how users assess a document's utility in a specific search task and defined all the necessary DtMRF components, employs a MADM method to estimate a global performance score corresponding to a document's utility. In this study, we introduce four MADM methods as instantiations of the proposed DtMRF, namely $\text{DtMRF}_{\text{WSM}}$, $\text{DtMRF}_{\text{COPRAS}}$, $\text{DtMRF}_{\text{TOPSIS}}$, $\text{DtMRF}_{\text{VIKOR}}$. The underlying assumption behind them is identical; however, the obtained document rankings may be different due to distinct mathematical properties of each underlying MADM method out of the four presented in Section 2.2.2. In the following section, we provide a few examples that illustrate the usage of DtMRF in IR, whereas we also comment on some aspects that require particular attention.

5.2.2 DtMRF: Proof of Concept

As shown in Section 2.2.1, the information around a MADM problem is organized in a decision matrix. Following this formulation, Table 5.1 introduces an example of a retrieval process in which DtMRF is applied for document ranking. In this example, the IR system aims at providing its end-users with topically relevant documents that are also familiar to them, understandable, and credible. To this aim, the system estimates the utility of five documents $D = \{d_1, d_2, d_3, d_4, d_5\}$, that are evaluated on the basis of four criteria, i.e topicality, familiarity, understandability, and credibility, denoted as $C = \{\text{top}, \text{fam}, \text{cred}, \text{und}\}$. In this example, we assume that familiarity measures the degree to which the content of a document is familiar to a user, which can be, for example, assessed as the similarity of the document content to a user's profile, where high values correspond to high familiarity and low values correspond to novel document content. Regarding the other criteria, these are dependent on the retrieval model (topicality), the user (understandability), and the document collection (credibility). Moreover, let us assume that four evaluation functions have been defined to quantify the satisfaction of these criteria, where the higher the value the higher the degree to which a document satisfies a criterion. Now, the system can provide different document rankings by acting on the weights and on the objectives associated with the considered criteria. In the

Table 5.1: Formulating the retrieval process using DtMRF and an $M_{5\times4}$ decision matrix to organize the problem's information.

	top	fam	cred	und
d_1	25.5	19.3	10.0	1
d_2	23.6	25.0	9.5	0
d_3	12.4	10.0	1.0	1
d_4	32.0	6.8	5.0	0
d_5	5.0	13.2	0.5	1

following, we denote by W the importance weights associated with the four criteria and by O their corresponding objectives. To begin with, setting $W = \{1, 0, 0, 0\}$ and $O = \{1, -, -, -\}$, where "$-$" indicates indifference (a criterion can be either beneficial or non-beneficial), all DtMRF instantiations produce rankings identical to a standard topicality-based retrieval approach ($d_4 > d_1 > d_2 > d_3 > d_5$), as the utility estimation is independent from the other criteria. For $O = \{0, -, -, -\}$, the ranking is reversed. Similarly, the system can be used for documents' ranking based on a single criterion.

Let us consider now the case in which we aim to create a system that provides its end-user with topically relevant and understandable documents. This can be achieved by setting $W = \{.5, 0, 0, .5\}$, when considering the criteria equally important, and $O = \{1, -, -, 1\}$, because both of the considered criteria positively affect a document's utility. All DtMRF instantiations (except from $\text{DtMRF}_{\text{VIKOR}}$[1])

[1]The $\text{DtMRF}_{\text{VIKOR}}$ instantiation with $\nu = .5$ ranks the documents in the following order,

rank the documents in the following order $d_1 > d_3 > d_5 > d_4 > d_2$.

Furthermore, let us consider the case in which the system has to provide documents that are topically relevant, familiar to the user, and which are also credible and understandable. That means the importance weights of the criteria should be set so that $W(\text{top}) > W(\text{fam}) > W(\text{cred}) > W(\text{und})$, for instance $W = \{.4, .3, .2, .1\}$ and with positive objectives, i.e $O = \{1, 1, 1, 1\}$. That leads to the following ranking, $d_1 > d_2 > d_4 > d_3 > d_5$, which is a reasonable ranking based on the requirements of the considered task. Considering now the situation in which the system applies the same importance weights, but it is aimed at providing novel information to its end-user, i.e the objective of the familiarity criterion is non-beneficial; therefore, $W = \{.4, .3, .2, .1\}$ and $O = \{1, 0, 1, 1\}$. This setting leads to the $d_4 > d_1 > d_2 > d_3 > d_5$ ranking for all instantiations, excepts for $\text{DtMRF}_{\text{VIKOR}}$ that creates the $d_4 > d_1 > d_3 > d_2 > d_5$ ranking. Here, the system ranks in the first position the document that satisfies the task characteristics to the highest degree, i.e it provides a highly topical relevant, understandable, and credible document that is not familiar to the user.

Lastly, the following example aims at showing the difference between assigning a negative objective to a criterion and associating a zero importance weight to it. A negative objective indicates that the lower the performance score of a criterion, the higher its contribution to a document's utility. While by assigning a zero weight to it, we neglect its contribution to a document's utility estimation. Having said that, in the previous example we saw that when $W = \{.4, .3, .2, .1\}$ and $O = \{1, 0, 1, 1\}$ the following ranking $d_4 > d_1 > d_2 > d_3 > d_5$ is produced. However, if we set $w_2 = 0$ and evenly distribute this weight among the other criteria ($W = \{.5, 0, .3, .2\}$), the obtained document ranking is $d_1 > d_4 > d_2 > d_3 > d_5$, that corresponds to a system in which $W(\text{top}) > W(\text{cred}) > W(\text{und})$. The first system provides in the first position a document that is topically relevant, novel (i.e not familiar), credible and understandable, while the second system provides a topically relevant document that is credible and understandable. In the latter system it is not feasible to interpret the contribution of a document's novelty to the final ranking, as we have neglected its effect.

The MADM methods exploited at the basis of DtMRF are usually employed in small-scale decision-making problems that involve few alternatives (e.g. selecting

$d_1 > d_3 > d_4 > d_5 > d_2$, as topical relevance of $top(d_4) \gg top(d_5)$.

appropriate automobile seats to be placed in a new vehicle based on a consumer's preferences [Behzadian et al., 2012]), while the performance scores are also estimated accurately, corresponding to an alternative's attributes (e.g. seat's weight, dimensions, among others). That is not the case of an IR approach that involves hundreds of thousands of documents, uncertainty in the estimation of the individual performance scores, and possible missing performance scores for some criteria. Moreover, due to the required matrix normalization step, it is possible that in the unlikely case that all the performance scores for a given criterion are zero, the usage of DtMRF is not computationally feasible. Therefore, when employing DtMRF, one should be aware of these possible situations and perform the necessary actions (e.g. handle missing values) to resolve them. Lastly, as outlined in the previous section, the selection of an appropriate evaluation function influences the system's efficiency. In particular, when the evaluation functions are computationally expensive, DtMRF can be applied on a smaller set of documents, for example as a re-ranker. In this case, an initial ranking can be performed using a specific subset of the selected criteria, and re-ranking on the remaining criteria (relevance factors). In the following section, we have undertaken an in-depth analysis of the ranking behavior of the proposed DtMRF instantiations, by expanding the example presented here.

5.2.2.1 Evaluating Retrieval Behavior Based on Criteria Weighting

This section investigates the impact of varying the weights of specific relevance factors—topicality, familiarity, credibility, and understandability—on document ranking. To conduct this investigation, we follow a retrieval simulation approach and analyze how weight changes affect the rankings obtained by each of the four DtMRF instantiations.

In the simulation, we generate a total of 1,000 documents. Each document is associated with four performance scores corresponding to the four relevance factors. These scores, as illustrated in Figure 5.1, are confined to a range of [0,1]. For topicality, familiarity, and credibility, scores are continuous real numbers generated using right-skewed, normal, and left-skewed distributions, respectively. In contrast, the scores for the understandability factor are binary, generated with a 0.7 probability for 0 and a 0.3 probability for 1. The generated $M_{1000\times4}$ decision matrix simulates the scenario in which a user initiates a search process, the system estimates performance scores for the considered relevance factors, and then applies DtMRF for ranking. In our investigation, all factors are considered as beneficial $O = \{1, 1, 1, 1\}$, i.e. they positively affect a document's overall relevance.

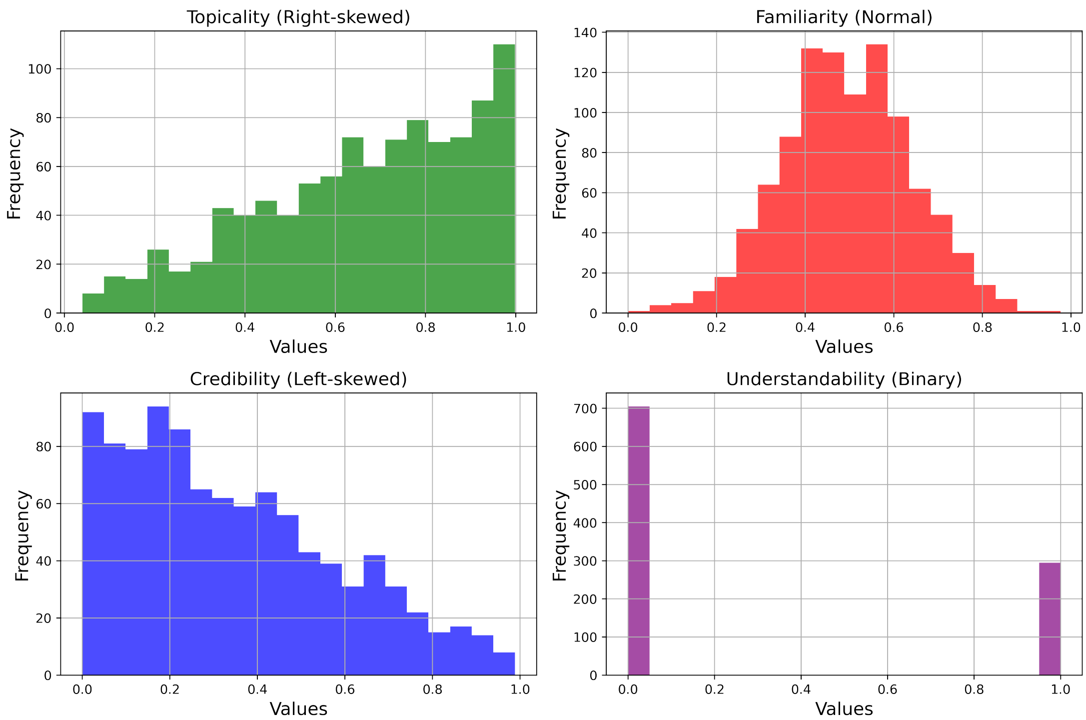

Figure 5.1: Generated distributions for the considered relevance factors.

To assess the impact of varying importance weights assigned to relevance factors on the resultant document ranking, we conduct a weight sensitivity analysis, following the methodology proposed by Alinezhad and Amini [2011]. The weight adjustment for sensitivity analysis is conducted by modifying a single weight w_p in the weight vector W based on a change factor D. Specifically, as the importance weights $W = \{w_1, w_2, w_3, w_4\}$ associated with the relevance factors must be $w_i \in [0, 1]$ and $\sum_{i=1}^{4} w_i = 1$, the following steps ensure that the remaining weights are adjusted in a manner that ensures their sum remains 1. The modified weight is given by $w'_p = w_p + D$. To ensure that the sum of weights remains 1, a correction factor c is calculated as $c = \frac{1-w'_p}{1-w_p}$. The other weights w_j in W, where $j \neq p$, are then adjusted according to $w'_j = c \times w_j$. For example, considering an initial weight vector $W = [.4, .4, .2]$. If $D = .1$ and $p = 0$, then $w'_p = .5$, $c = \frac{1-.5}{1-.4} = \frac{.5}{.6} \approx .8333$, and the new weights become $W' = [.5, .3333, .1667]$.

In our analysis we start with equal importance weights for our factors $W = [.25, .25, .25, .25]$. Then, we alter the weight of one factor, i.e. topicality, by an incremental factor of $D \in [0, .75]$ based on the aforementioned process. As our simulated document collection has one thousand documents, we focus our attention on six documents. In detail, we present how the ranking position of these six documents changes, as the importance weights of the criteria is changing. We have chosen three documents that have the highest scores with respect to the topicality, familiarity, and credibility factors, and two random documents. As understandability is estimated in binary values, we have chosen a random document with score of 1. Table 5.2 presents the selected documents, along with their performance scores across the four relevance factors.

Table 5.2: Performance scores of the selected documents, with respect to the four relevance factors.

Document	Performance Score [Topicality, Familiarity, Credibility, Understandability]
ID 572 [Highest Topicality]	[.998, .421, .108, 1]
ID 578 [Highest Familiarity]	[.464, .977, .810, 0]
ID 3 [Highest Credibility]	[.515, .604, .989, 0]
ID 649 [Highest Understandability]	[.693, .583, .268, 1]
ID 923 [Randomly Selected]	[.996, .591, .373, 1]
ID 900 [Randomly Selected]	[.963, .439, .046, 0]

Figure 5.2 shows how the ranking positions (y-axis) of the six selected documents changes for different DtMRF instantiations when the topicality weight is increased by D (x-axis). Upon reviewing all four figures, it's evident that the four DtMRF instantiations behave similarly in retrieving documents, as the ranking order of the selected documents is mostly identical[2]. However, some differences do arise due to the underlying aggregation mechanisms of each instantiation, i.e. the documents obtain different ranking positions. Each symbol (e.g. $\circ, \Delta$) in the plotted lines indicates a change in the composition of the top-10 ranked documents (i.e inclusion of a new document) relative to the previous model executions, i.e. using different weights. Observing the distance between them one can see the sensitivity of each instantiation to small variations in the weights. In the figures, the red vertical lines

[2] $\text{DtMRF}_{\text{COPRAS}}$ and $\text{DtMRF}_{\text{WSM}}$ have identical behavior because we have assumed positive objectives for all the relevance factors.

serve to pinpoint the specific weight value associated with the topicality factor. This provides a visual marker to easily discern how this particular weight influences the ranking of documents.

The starting retrieval system attributes equal importance to all relevance factors, i.e $D = 0, W = [.25, .25, .25, .25]$. This represents a balanced system where each factor contributes equally to the document ranking. From left to right, the retrieval systems weight more topicality, i.e. documents with higher topical relevance score get higher ranking positions. This behavior becomes evident by observing document 572 that has the highest topicality score in the collection, following the blue line. The final system that weighs solely on topicality $D = .75, W = [1.0, .0, .0, .0]$ retrieves this document in the first position. Close examination of the upper-right sections of each figure reveals specific insights about the rankings of documents 923 and 572. Document 923 has performance scores of $[0.9964, 0.5914, 0.3728, 1]$ for topicality, familiarity, credibility, and understandability, respectively. In contrast, document 572 has scores of $[0.9982, 0.4214, 0.1084, 1]$. In a search context where all selected relevance factors align with user preferences, document 923 is generally preferred over document 572. The instantiations' ability to capture this preference is evident in the figures, where document 923 consistently ranks above document 572 under various weight settings for topicality (i.e. $w_{\text{top}} \geq 0.95$), except when $w_{\text{top}} = 1.0$. The $\text{DtMRF}_{\text{TOPSIS}}$ instantiations captures this behavior more effectively, as the differences in performance scores across the remaining three relevance factors provide enough differentiation to rank the documents.

In terms of the factors of familiarity and credibility, our results align with those observed when increasing the weight of topicality. This suggests that the performance score distribution does not influence the ranking behavior of the DtMRF instantiations. To conserve space, we have opted not to include the corresponding figures. Nonetheless, it is crucial to discuss the models' ranking behavior in relation to the understandability factor, which has binary values. This evaluation holds significance as it models a situation where a search system incorporates a classification score into its ranking mechanism. Additionally, the study highlights a recognized limitation of scoring-based MADM methods.

Scoring-based methods like WSM and COPRAS are sensitive to the scale of the criteria, potentially diminishing the impact of binary-valued criteria in the final ranking. These methods typically lack a built-in normalization mechanism, ampli-

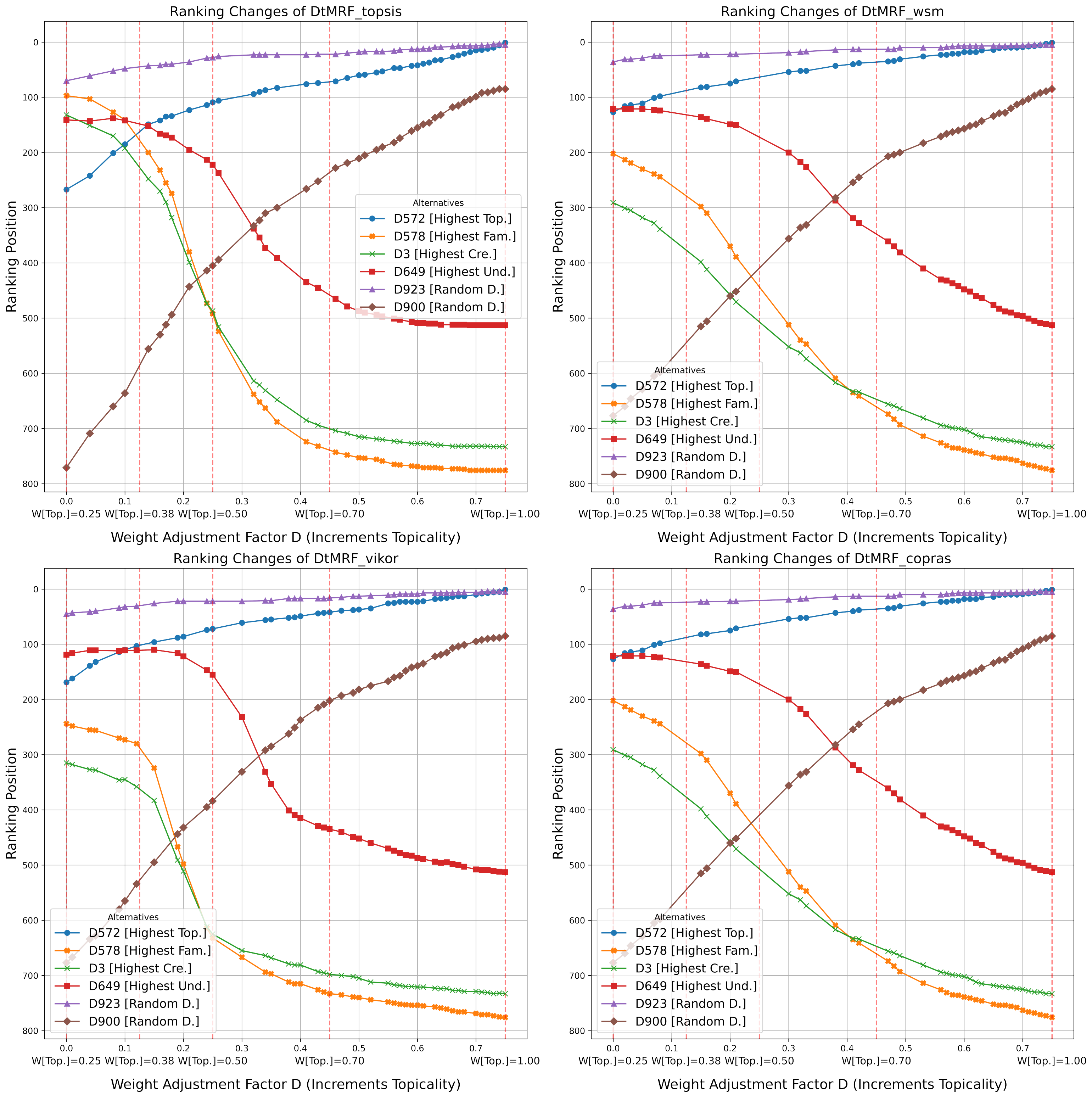

Figure 5.2: Ranking sensitivity to weight changes.

fying the scale mismatch between binary and continuous criteria. Consequently, the decision matrix might require manual normalization to ensure the binary criterion is not underrepresented. Similarly, VIKOR's method of aggregating individual regrets and lack of inherent normalization can make it less suited for handling binary features effectively. In contrast, TOPSIS automatically normalizes the decision matrix, making it more robust to mixed data types, including binary and continuous scores. Furthermore, TOPSIS employs both ideal and negative-ideal solutions in its ranking mechanism, allowing for a more nuanced differentiation between closely ranked alternatives, which can be particularly beneficial when one of the criteria is binary.

We note that binary scores can be used to filter out several documents. However, leveraging DtMRF instantiations like $\text{DtMRF}_{\text{TOPSIS}}$ allows for a more comprehensive relevance estimation. Utilizing soft or hard filtering based on binary relevance scores could eliminate documents that might otherwise be viable when considering the full spectrum of relevance factors. DtMRF enables us to integrate all criteria, binary and continuous, into a single decision framework, thereby facilitating a more balanced and holistic relevance estimation for the documents.

The aforementioned distinctions in handling mixed criteria, particularly binary ones, become evident when we perform the same process as before for the understandability relevance factor. Specifically, $\text{DtMRF}_{\text{TOPSIS}}$ emerges as the more robust method for accurately ranking alternatives in scenarios where the criteria are a mix of binary and continuous values. In the generated collection we have 300 understandable documents, with performance score equal to 1. Three of the selected documents have score of one, namely documents 572, 649, and 923. Examining the sub-figures in Figure 5.3, it is apparent that all DtMRF instantiations consistently rank the discussed documents within the top 300 positions, which is the expected ranking behavior. What stands out is the absence of symbols (e.g. $\circ, \Delta$) along the plot lines for three DtMRF instantiations. We remind that these symbols indicate a change in the composition of the top-10 ranked documents relative to the previous model executions. Due to its score normalization and aggregation techniques, only the $\text{DtMRF}_{\text{TOPSIS}}$ instantiation can modify the rankings within the top-10 positions.

The examination presented in this section highlights the importance of selecting appropriate importance weights when aggregating performance scores. While

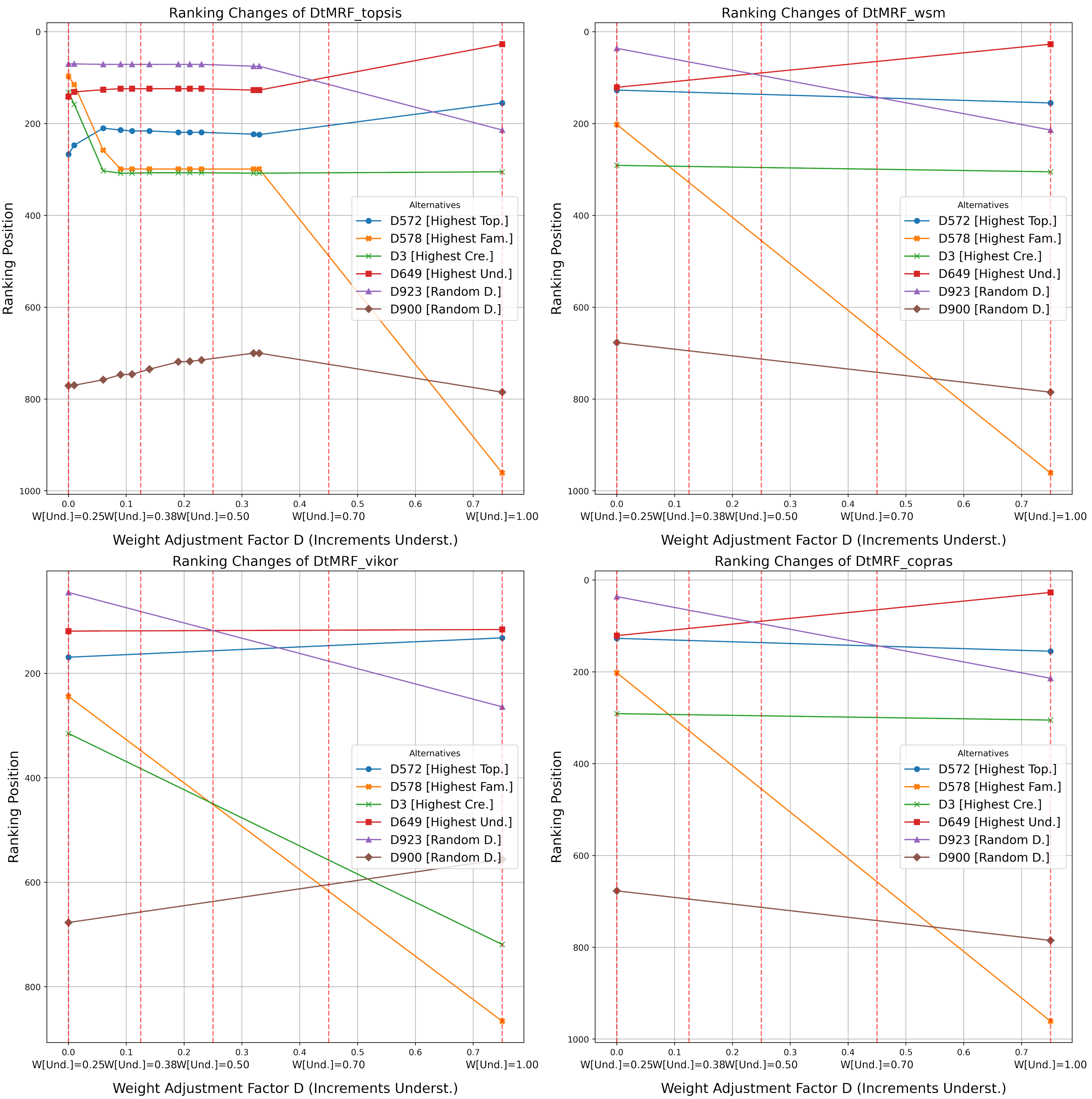

Figure 5.3: Ranking sensitivity to weight changes, optimizing a relevance factor with binary performance scores.

our analysis could be extended to include non-beneficial criteria, doing so would complicate the interpretation of the ranking without necessarily changing the conclusions already reached. Nonetheless, it highlights some limitations related to the commonly employed weighted sum method, and underscores how other methods like DtMRF$_{\text{TOPSIS}}$, can overcome them. The selection of importance weights for criteria is a thoroughly researched area in MCDM. Despite this, the estimation of weights in MCDM is inherently subjective, frequently depending on the decision-maker's judgment. In our framework, an IR system is the decision-maker, allowing for a more objective estimation of importance weights. In the subsequent section, we outline our methodology, which is incorporated into the DtMRF framework for predicting a set of importance weights corresponding to the relevance factors of a given search task.

5.3 DtMRF: Weight Prediction

The section describes how DtMRF instantiations can be combined with a component that estimates importance weights to be associated with a predefined set of relevance factors. These estimated weights are specifically tailored to a distinct search context, an information need (query), and a DtMRF instantiation. In our research, we have formulated this problem as a multi-output regression problem. Following and slightly adapting the formulation presented in Section 2.4, the problem of predicting importance weights is formally defined by the following equation:

$$W = F(x; \Theta) + \epsilon \tag{5.1}$$

In this formulation, x represents the input variables associated with a specific search context, whereas W indicates the importance weights. $F(x; \Theta)$ aims to estimate the weights that would yield an optimal document ranking for a given search context and query, when a DtMRF instantiation is used. While $F(x; \Theta)$ can be any model, in our research we use deep neural networks. In the following section, we elaborate on the essential prerequisites for training a weight prediction model.

5.3.1 Neural-DtMRF Components

Neural-DtMRF maintains the foundational components of DtMRF and introduces further elements that necessitate definition, as described below.

Training Dataset. To utilize a neural model for weight prediction, a specialized training dataset that meets the requirements of a given search task must be constructed. The procedure entails the selection of a document corpus, a query set, and relevance assessments, all associated with a studied search task. Relevance assessments must extend beyond topical relevance and be grounded in multiple relevance factors. Alternatively, each document within the collection could be evaluated for its utility in fulfilling the user's task, essentially incorporating a "usefulness" (utility) judgment, as proposed by Belkin et al. [2009].

Optimal Weights (Model's Outputs). The determination of optimal weights for each query involves several steps. Initially, a DtMRF instantiation is selected for ranking. Then, DtMRF is run for all combinations of weights; this is feasible given that the weights are constrained to sum to one, and a reasonable step size, such as 0.05, can be chosen. Even though the selected optimal weights are influenced by the chosen DtMRF instantiation, as shown in Section 5.2.2.1, the proposed DtMRF instantiations have similar retrieval behavior; it is expected that the weights can generalize. Task-specific requirements dictate the appropriate evaluation measure to be maximized. If the dataset includes multiple relevance factors, an evaluation measure that considers all of them should be employed. Such measures have been introduced by Palotti et al. [2018] and Maistro et al. [2021]. Alternatively, label aggregation techniques can be employed to combine the labels associated with various relevance factors into a single label [Kang et al., 2012, Carmel et al., 2020]. Lastly, if the collection features usefulness judgments, the optimal weights are those that yield the highest usefulness scores.

Model's Inputs and Neural Model. The neural model aims to predict a set of weights for a specific information need within a search context. Consequently, the model's input vector x, can be inherently related to instances of this search context, given their potential influence on the relevance factors under consideration. Although x may differ based on the characteristics of the task and the data at hand, it could be associated with the context of query (q), such as an embedding representation of the query itself. Additionally, as W may be affected by additional search aspects, x could contain features related the user (u), the user-system interactions (i), or other (o). As a result, $F(x; \Theta)$ can be more precisely expressed as $F(q; u; i; o; \Theta)$. Subsequently, the neural network-based regression model could employ a Mean Squared Error loss function in conjunction with a conventional optimization algorithm for training. Nonetheless, the architecture and characteristics of the model should be tailored to the specific tasks under study. During inference, the model processes the input features x and estimates the importance weights for an expressed information need (i.e. query).

While we have not elaborated on the alternative methods for multi-output regression (e.g. chaining), adapting this approach to fit within those frameworks is straightforward, following the formulations presented in Section 2.4.

5.4 Discussion

This chapter presents a generalizable formal framework for multidimensional relevance estimation, featuring four specialized instantiations for document ranking, namely DtMRF.

In addressing the research questions (RQ1), (RQ2), and (RQ3), the examples presented in Section 5.2.2 explain how one can formulate IR as a decision-making problem using the formal setting proposed in this paper. Our comprehensive analysis reveals the expected retrieval behaviour exhibited by the four DtMRF instantiations and emphasizes the advantages of $\text{DtMRF}_{\text{TOPSIS}}$ over specific existing methods in the literature. Moreover, the examples make it evident how DtMRF provides fully interpretable document rankings and how the choice of objectives and weights influences retrieval behaviour, providing evidence that answer the fourth research question. Integrating a neural model for weight prediction allows all DtMRF instantiations to be transformed into hybrid models. These hybrid approaches harness the transparency of decision-theoretic models to aggregate performance scores while benefiting from the predictive capabilities of neural networks to estimate query-specific importance weights.

DtMRF decomposes the relevance estimation in IR into multiple distinct ranking systems, each responsible for estimating a specific relevance factor. The systems' outputs are then combined through a single aggregation mechanism that incorporates the characteristics of the task in the relevance estimation, making the obtained rankings interpretable. Neural-DtMRF extends this framework by learning to fine-tune the score aggregation process according to search and user related information. Consequently, DtMRF is well-suited to meet the specialized requirements of professional search, particularly interpretability and user control. Its transparency enables users to understand the underlying logic that guides the ranking, thereby increasing the reliability and validity of the search results. Additionally, DtMRF offers users a higher degree of control over the search process, as the user can alter the retrieval behaviour by altering the weights associated with the relevance factors. Given that the DtMRF system comprises distinct components, each responsible for specific aspects of relevance estimation, it becomes considerably easier to maintain and update the system. A performance issue in one component can be isolated and addressed without affecting the entire system, simplifying maintenance and incremental updates.

For future work, we plan to enhance Neural-DtMRF by incorporating the capability to estimate both importance weights and objectives. That would allow the system to better adapt to a search task's requirements, hence offering an additional layer of flexibility. The subsequent sections, explore the real-world utility of DtMRF and Neural-DtMRF within the context of a complex search task in the medical domain.

Part III Putting Theory to the Test: Experimental Insights

Chapter 6

Extracting Information from Electronic Health Records

Building upon the foundational theories presented in earlier chapters, this chapter examines the methodologies we have employed for information extraction from Electronic Health Records, explicitly targeting the optimization of clinical trials retrieval. The chapter investigates the effectiveness of various approaches: rule-based methods such as the ConText algorithm, transformer architectures like BioBERT, and large language models including GPT-3.5. These are leveraged for their capabilities in medical entities' extraction and semantic disambiguation, aiming to evaluate whether their outcomes improve retrieval. Besides, the chapter comprehensively analyses our methodological methods, explicates findings using the aforementioned benchmark collections, and underscores pivotal outcomes.

6.1 Investigating Rule-based and Transformer-based Methods for Clinical Trials Retrieval

This section extends the foundational work of Koopman and Zuccon [2014] and Agosti et al. [2019] by introducing an integrated approach that combines novel transformer-based methods with traditional rule-based techniques. Specifically, the focus is on enhancing the extraction of medical entities and the disambiguation of their semantic meanings. Furthermore, the section aims to contribute new empirical insights into different query representations' efficacy in retrieving clinical trials. That is achieved by incorporating additional patient information, including patients' historical information and life-style habits, into the query formulation process.

We specifically aim to address the following research questions:

(RQ1) Does the presence of various medical entities of a clinical note have an impact on the overall retrieval effectiveness?

(RQ2) How does the presence of negated content affect retrieval performance?

(RQ3) How does the presence of sentences with non-identified medical entities impact retrieval performance?

(RQ4) How does the presence of family history and/or patient's historical information affect retrieval performance?

(RQ5) What is the impact of medical entity expansion, using a knowledge base, on retrieval performance?

To address these research questions, we utilize a combination of state-of-the-art methods for entity extraction and semantic disambiguation, including rule-based algorithms and transformer-based models. We formulate a variety of query representations and assess their efficacy through a comparative analysis. The performance is evaluated using two established benchmark collections: TREC 2021 and the Clinical dataset, described in Section 4.3.

6.1.1 Methodology

Figure 6.1 provides a comprehensive overview of the proposed methodology, encompassing information extraction, semantic disambiguation of entity meanings, and entity expansion.

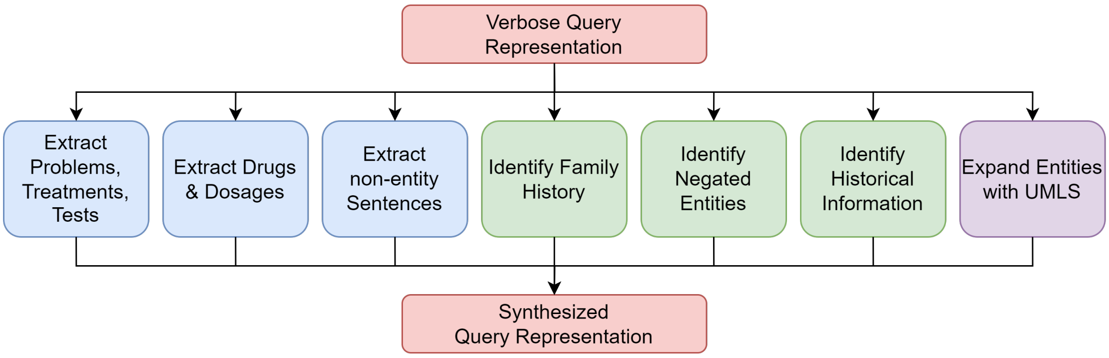

Figure 6.1: Overview of our methodology, where information extraction methods are delineated in blue, semantic disambiguation methods in green and entity expansion methods in purple.

Starting with a patient's clinical note, which serves as a verbose query representation, we generate all of the possible query formulations by employing the methods depicted in Figure 6.1. These synthesized queries are then utilized to retrieve relevant clinical trials. For medical entity extraction—specifically, problems, treatments, and tests—we deploy a pre-trained transformer-based NER model[1], which has been trained on the n2c2 dataset and introduced by Uzuner et al. [2011].

In biomedical literature, various libraries and models are available for drug, dosage, and disease extraction. Following the methodologies of prior studies by Leaman et al. [2021] and Zhang et al. [2021], we evaluated the performance of SciSpacy [Neumann et al., 2019], Stanza [Zhang et al., 2021], and BioBERT [Lee et al., 2020b], finding Stanza to exhibit the most robust performance. Clinical notes often include sentences that do not explicitly mention medical entities but may refer to patient behaviors such as smoking or physical activity. In our experiments, we investigated the effect of these sentences on the retrieval performance to investigate whether these sentences contain noise or valuable information.

Upon extracting the relevant medical entities, we identify those that are negated,

[1]Bert-base-uncased clinical NER, accessed 12/10/2022

pertain to family history, or refer to historical patient data. For negation identification, we contrast the widely-used ConText algorithm [Chapman et al., 2001b, Harkema et al., 2009] with a pre-trained transformer-based model by van Aken et al. [2021], which is fine-tuned on negation assertion in clinical notes. We employ the ConText algorithm as implemented in medSpacy [Eyre et al., 2021] for discerning family history and historical information. That implementation allows for multi-token regular expressions to be used for case-specific semantic meaning disambiguation. Therefore, it enable us to effectively disambiguate the semantic meanings of all extracted entities. Lastly, we utilize the UMLS [Bodenreider, 2004] for entity expansion, retaining the original entities while augmenting them with aliases, UMLS concepts, and definitions, as inspired by Agosti et al. [2019]. The source code that implements our methodology is publicly available online[2].

6.1.2 Experimental Design and Results

This section offers a detailed analysis of the experimental framework, encompassing the evaluation metrics employed, the acquisition and indexing of data collections, and the retrieval configurations implemented.

Evaluation Metrics. To assess the efficacy of our experimental setups, we targeted both early and late precision by measuring P@5 and P@25. Furthermore, we include Bpref in our evaluation metrics, which omits retrieved documents that have not undergone human annotation in its calculation. All metrics consider only *eligible* trials as relevant, following the official guidelines in this task. Readers are referred to Chapter 2.1.3 for an in-depth presentation of the selected measures.

Data Collection and Indexing. The benchmark collections (TREC 2021 and Clinical) for our experiments were obtained from the *ir-datasets* [MacAvaney et al., 2021] and repository[3], respectively. We utilized PyTerrier [Macdonald and Tonellotto, 2020] for indexing, adhering to its default parameters such as porter-stemming and stopword removal. Each document is indexed in its entirety, i.e. including all available sections.

Experiments. In alignment with the leading methodologies in the TREC 2021 and 2022 tracks, as cited in Section 4.2.2, our retrieval approaches leverage the

[2] Source Code for our experimental procedures, accessed 12/10/2022.
[3] A Test Collection for Matching Patient to Clinical Trials, accessed on 12/10/2022.

BM25 model and its default PyTerrier's for all retrieval experiments. We utilize the original query with the BM25 model as our baseline for the empirical evaluation, referred to as *Verbose query*. Against this baseline, we evaluate the efficacy of alternative query representations that have demonstrated superior retrieval performance. The outcomes of these comparisons are tabulated in Table 6.1. The statistical significance is tested against the effectiveness achieved by the verbose query representation according to a paired t-test[4] with Bonferroni multiple testing correction[5], at significance levels 0.05(°).

Table 6.1: Retrieval effectiveness achieved by the top-performing synthesized queries.

Query Repr.	TREC 2021			Clinical		
	Bpref	P@5	P@25	Bpref	P@5	P@25
Verbose query	.184	.291	.211	.065	.050	.023
$Q1_{prob\_treat\_test}$	.211°	.323	.218	.077	.032	.021
$Q2_{drug\_dis}$	.196	.192	.167	.073	.046	.016
$Q3_{un\_comb\_Q1\_Q2}$	**.214°**	.299	**.227**	.084	**.054**	.025
$Q4_{non\_neg\_Q1\_trans}$	**.214°**	.323	.218	.082	.029	.023
$Q5_{non\_neg\_Q1\_con}$	.205	.291	.201	.074	.036	.020
$Q6_{comb\_Q4\_no\_ent}$	.206	.304	.220	.083	.036	.021
$Q7_{comb\_Q3\_no\_ent}$	.212°	.304	.225	**.090**	.050	**.026**
$Q8_{Q4\_rem\_fam\_hist}$	.207	.312	.206	.087	.014	.017
$Q9_{Q4\_rem\_fam}$	.212°	**.331**	.216	.084	.025	.023
$Q10_{Q4\_rem\_hist}$	.205	.304	.202	.083	.021	.018
$Q11_{Q9\_exp\_def}$	.183	.213	.143	.089	**.054**	.017
$Q12_{Q9\_exp\_alia}$	.182	.208	.143	.089	.050	.016
Human ad-hoc	-	-	-	.094	.071	.034

Having analyzed the empirical data, we now focus on addressing the research

[4] Paired Two-Sample Student's t-Test, accessed on 12/10/23.
[5] Bonferroni multiple testing correction, accessed on 12/10/23.

questions that have served as the foundation for this investigation.

RQ1: Does the presence of various medical entities of a clinical note have an impact on the overall retrieval effectiveness? Our empirical analysis indicates that the query representation $Q1_{\text{prob\_treat\_test}}$, comprising concatenated text from a patient's identified problems, treatments, and tests (extracted via the transformer-based model[1]), outperforms $Q2_{\text{drug\_dis}}$, which consolidates the patient's identified diseases and drugs as extracted by Stanza. Notably, an enhancement in retrieval effectiveness is observed when these two query representations are combined by taking their union and retaining the unique terms ($Q3_{\text{un\_comb\_Q1\_Q2}}$). These findings suggest that a multi-faceted query representation, which incorporates various types of medical entities, can significantly improve the effectiveness of clinical trial retrieval. Moreover, the synergistic combination of different query representations ($Q3_{\text{un\_comb\_Q1\_Q2}}$) further substantiates the notion that a more comprehensive query, capturing multiple dimensions of patient information, yields superior retrieval performance.

RQ2: How does the presence of negated content affect retrieval performance? The query representations $Q4_{\text{non\_neg\_Q1\_trans}}$ and $Q5_{\text{non\_neg\_Q1\_con}}$ consist of the non-negated entities extracted from $Q1_{\text{prob\_treat\_test}}$. Our analysis of the bpref metric indicates that query formulations derived from the pre-trained transformer model [van Aken et al., 2021] ($Q4_{\text{non\_neg\_Q1\_trans}}$) exhibit superior effectiveness compared to those generated using the ConText algorithm ($Q5_{\text{non\_neg\_Q1\_con}}$). Overall, excluding negated entities enhances retrieval effectiveness, as evidenced by the increase in the bpref measure across both benchmark collections. These findings yield two key outcomes. First, they underscore the efficacy of pre-trained transformer models, specifically as referenced by van Aken et al. [2021], in generating more effective query representations for clinical trial retrieval than traditional rule-based methods like the ConText algorithm. Second, they corroborate the hypothesis that excluding negated medical entities from query representations can lead to a measurable improvement in retrieval effectiveness.

RQ3: How does the presence of sentences with non-identified medical entities impact retrieval performance? Our analysis reveals that the query representation $Q7_{\text{comb\_Q3\_no\_ent}}$ demonstrates improvements over the baseline across the two collections. This representation combines the non-negated entities from $Q3_{\text{un\_comb\_Q1\_Q2}}$ with sentences devoid of identified medical entities. Similarly,

$Q6_{\text{comb\_Q4\_no\_ent}}$ merges the non-negated entities from $Q4_{\text{non\_neg\_Q1\_trans}}$ with such sentences with non-identified medical entities. Contrary to the notion that these sentences may introduce noise into the query, our findings suggest that they contribute essential information that enhances retrieval performance. However, in the context of the TREC 2021 collection, our observations indicate that the inclusion of sentences with non-identified medical entities does not yield a significant improvement in retrieval performance, as measured by P@5 when compared to $Q4_{\text{non\_neg\_Q1\_trans}}$. That suggests a need for a more nuanced semantic analysis of these sentences to capture patient lifestyle factors accurately. Regrettably, as of our current understanding, no pre-trained model specifically designed to extract lifestyle factors or patient habits from clinical notes exists.

RQ4: How does the presence of family history and/or patient's historical information affect retrieval performance? This analysis investigates the consequences of selectively omitting entities related to family history and patient's historical information from the query representation $Q4_{\text{non\_neg\_Q1\_trans}}$. Specifically, we examine three modified query representations: $Q8_{\text{Q4\_rem\_fam\_hist}}$, which excludes both family history and historical information; $Q9_{\text{Q4\_rem\_fam}}$, which omits only family history; and $Q10_{\text{Q4\_rem\_hist}}$, which removes only historical information. Our results indicate that excluding family history-related entities tends to enhance retrieval precision. However, in alignment with the findings of Koopman and Zuccon [2014], we also observe that removing historical information identified by the ConText algorithm can introduce errors. That is particularly evident when a clinical note predominantly contains historical medical information, in the sense that the clinician wrote it in the past tense.

RQ5: What is the impact of medical entity expansion, using a knowledge base, on retrieval performance? Our empirical investigation reveals that the general application of query expansion techniques, incorporating aliases, medical concepts, and concept definitions, only sometimes universally enhanced retrieval performance. However, among the evaluated query representations, two specific instances, namely $Q11$_Q9_exp_def and $Q12$_Q9_exp_alia, demonstrated superior performance. The former expands $Q9$_Q4_rem_fam by incorporating aliases, medical concepts, and definitions, while the latter includes only aliases and medical concepts. One plausible explanation for this observed behaviour could be the introduction of query topic drift. The expanded queries may incorporate too generic terms, thereby diluting the specificity of the original query and adversely affecting retrieval performance.

An overall observation from Table 6.1 is that none of the algorithmically synthesized queries managed to surpass the performance of human-generated ad-hoc queries, where available. These ad-hoc queries were meticulously crafted by a panel of four medical assessors, as delineated by Koopman and Zuccon [2016]. For specific topics, these assessors generated multiple short queries. In our study, we concatenate these short queries to formulate a singular query representation, retaining only the unique terms. The empirical results corroborate the notion that the query representations evaluated in this study enhance retrieval effectiveness in clinical trials.

6.1.3 A Qualitative Example

Figure 6.2 provides a qualitative example of an EHR (i.e. verbose query representation) from the TREC 2021 Clinical Trials collection to illustrate the information extracted.

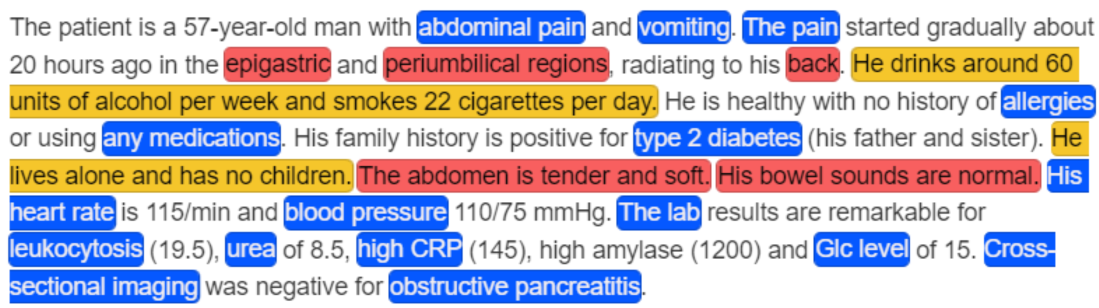
The patient is a 57-year-old man with abdominal pain and vomiting. The pain started gradually about 20 hours ago in the epigastric and periumbilical regions, radiating to his back. He drinks around 60 units of alcohol per week and smokes 22 cigarettes per day. He is healthy with no history of allergies or using any medications. His family history is positive for type 2 diabetes (his father and sister). He lives alone and has no children. The abdomen is tender and soft. His bowel sounds are normal. His heart rate is 115/min and blood pressure 110/75 mmHg. The lab results are remarkable for leukocytosis (19.5), urea of 8.5, high CRP (145), high amylase (1200) and Glc level of 15. Cross-sectional imaging was negative for obstructive pancreatitis.

Figure 6.2: Verbose query representation for topic 21 in the TREC 2021 collection.

Specifically, terms sketched in blue constitute the $Q1_{\text{prob\_treat\_test}}$ representation. Sentences without medical entities are highlighted in yellow. In contrast, those containing medical entities not identified by the deployed methodologies are marked in red. These overlooked entities are noteworthy as they pertain to the patient's condition. The $Q4_{\text{non\_neg\_Q1\_trans}}$ representation encompasses all terms highlighted in blue, except terms such as *allergies*, *any medications*, *cross-sectional imaging*, and *obstructive pancreatitis*, which have been classified as negations. By concatenating the yellow-highlighted sentences with the $Q4_{\text{non\_neg\_Q1\_trans}}$ representation, one obtains the $Q6_{\text{comb\_Q4\_no\_ent}}$ representation for this specific topic. Further, the removal of the identified medical entity *type 2 diabetes* yields the $Q8_{\text{Q4\_rem\_fam\_hist}}$ and $Q9_{\text{Q4\_rem\_fam}}$ representations; it is noteworthy that no terms were identified as historical information for this topic. Finally, the $Q9_{\text{Q4\_rem\_fam}}$ representation can be expanded using UMLS aliases or definitions to generate the terminal query representations for this topic, as delineated in Table 6.1.

6.1.4 Conclusions and Directions for Future Research

Several key conclusions can be drawn in light of the preliminary obtained results. Using an apt query representation enriched with extracted medical entities enhances retrieval performance in patient allocation for clinical trials. Transformer-based models, fine-tuned on domain-specific data for negation identification, exhibit superior performance compared to conventional rule-based approaches. We show that existing transformer-based models are limited in identifying crucial patient information, such as lifestyle habits, when such information is not covered in medical terminology. Nevertheless, these sentences are replete with indispensable information. Removing family-related information augments early precision in clinical trial retrieval. Conversely, the excision of historical information proves less effective, a limitation of the identification methodology employed. Finally, our findings indicate that entity expansion via the UMLS fails to enhance retrieval effectiveness in the tasks under consideration.

This section underscores the significance of information extraction techniques in enhancing clinical trials retrieval performance while illuminating specific limitations. The subsequent section will investigate the capabilities of LLMs to fill the research gaps highlighted in these initial findings. LLMs benefit from training on comprehensive and diverse datasets, which equips them to discern intricate relationships among medical entities. Their ability to understand context and semantics makes them particularly adept at extracting relevant information, even when it is not explicitly expressed in medical terminology.

6.2 Utilizing ChatGPT to Enhance Clinical Trials Retrieval

The extraordinary generative capabilities exhibited by various LLMs have led to extensive discussions regarding their exploration and adoption in the medical domain. One promising area of application for LLMs, such as ChatGPT, is the generation of discharge summaries [Patel and Lam, 2023]. Due to their standardized format, discharge summaries could benefit from the utilization of ChatGPT, which has the potential to enhance the quality of these summaries. An empirical investigation on the usage of ChatGPT has been conducted by medical experts [Cascella et al., 2023]. Their findings support the idea that ChatGPT can be utilized for generating medical notes, given an adequate amount of patient-specific information. Furthermore, they outline that ChatGPT has the potential to effectively handle complex data and extract valuable information from various medical texts, including EHRs, clinical notes, and research papers. However, the primary constraint of ChatGPT lies in its inability to address causal relationships among conditions and comprehend the intricate connections between different conditions and treatments. Another possible application of an LLM model, specifically the GPT-3 model, in a medical context is presented by Sezgin et al. [2022]. The authors discuss the implementation and operationalization of GPT-3 in clinical practice, focusing on factors such as integration with existing hospital networks, ensuring secure connectivity, incorporating text summarization services, and storing generated information in patients' EHRs. Although LLMs exhibit promising capabilities, there are limitations that give rise to concerns regarding their adoption in the medical domain as highlighted by several scholars [Harrer, 2023, Albrecht et al., 2022, Patel and Lam, 2023, Cascella et al., 2023]. Besides their well-known limitations such as lack of understanding and reasoning, hallucination generation [Ji et al., 2023], inconsistent responses, outdated knowledge, their application in the medical domain also raises reliability and safety concerns [Albrecht et al., 2022]. For instance, the ability of ChatGPT to generate answers that may seem plausible but are incorrect raises concerns about the possibility of providing incorrect information regarding medication changes, which can significantly impact patient care [Patel and Lam, 2023]. Having said that, in our study, ChatGPT is employed as an intermediate step in a retrieval pipeline, while its role is to extract patient-related information from unstructured clinical notes rather than generate new pieces of information or make any decisions. Hence, our approach has a reduced exposure to the aforementioned limitations typically associated with LLMs.

Prior to the advent of LLMs, extracting information from clinical notes predominantly relied on domain-specific pre-trained models, akin to those discussed in the preceding section. These models were fine-tuned to discern critical patient-related details [Landolsi et al., 2023]. Nevertheless, tasks such as extracting lifestyle factors of a patient pose challenges, primarily due to the scarcity of suitable models and data, as previously underscored.

To address that, we propose an approach that automates the initial step of the enrollment process while maintaining the quality of medical service and minimizing direct risks to patients. We leverage ChatGPT to extract patient-related information from unstructured clinical notes and generate search queries for retrieving potentially eligible clinical trials. This study aims to address the following research questions:

(RQ1) Is the patient-related information extracted by ChatGPT sufficient to improve retrieval performance?

(RQ2) Which of the employed prompting approach yields the highest retrieval performance in the studied search task?

(RQ3) Does the utilization of ChatGPT enhance retrieval performance compared to existing state-of-the-art approaches in the literature?

(RQ4) What is the achieved retrieval performance of queries generated by ChatGPT and those generated by humans?

(RQ5) Can the integration of ChatGPT into the clinical trial enrollment pipeline be beneficial?

There are several reasons behind our decision to employ this model instead of other domain-specific LLMs such as GatorTron or Med-PaLM [Singhal et al., 2022]. To begin with, due to the popularity of ChatGPT compared to other domain-specific LLMs, a vast amount of information about prompt engineering in this model is publicly available[6, 7]. The availability of these resources, along with the significant

[6] Awesome ChatGPT Prompts, accessed on 21/04/2023.
[7] Techniques to improve LLM's reliability, accessed on 21/04/2023.

amount of research works related to ChatGPT, provide essential insights that aid our investigation. In addition, we mainly focus on the simple task of medical IE and not on more complex NLP tasks, such as medical question answering or medical reasoning. As we have already mentioned, the GatorTron model has been found to perform marginally better than previous state-of-the-art models in medical IE tasks [Yang et al., 2022a]. Lastly, our study explores some prompting approaches that mimic a user-system conversation; we refer to these as two-step processes. Therefore, ChatGPT is more suitable to be employed in this context. The aforementioned ChatGPT's ability, i.e. user-system conversations, makes it also more suitable than *text-davinci-003* model. In addition to that, it is also cost-effective while it provides comparable performance to *text-davinci-003*, as it has a lower price per token[8].

The following section outlines the methodological framework, emphasizing the prompts used to generate query representations. Subsequently, we present a comprehensive analysis of the experimental results, shedding light on the retrieval effectiveness of various query representations. Finally, we compare the capabilities of ChatGPT in formulating queries for unstructured clinical notes against human-generated queries and provide valuable insights into its applicability in the domain of clinical trials retrieval.

6.2.1 Methodology

In this section, we expound upon the methodological framework that underpins our research. We outline the strategies employed, justify their selection, and describe how they are operationalized to address our research questions.

In the approaches we propose to address the task of information extraction with ChatGPT, we acknowledge the inherent limitations of LLMs and actively incorporate measures to mitigate these in our implementation. Some limitations of LLMs pertain to issues such as response consistency and the generation of hallucinated content. The hallucination effect holds minimal applicability in our study, given that the model is primarily tasked with information extraction, modification, and structuring rather than autonomous decision-making or content generation. Additionally, by integrating the proposed approach as an intermediary component in a

[8]OpenAI Guides, accessed on 21/04/2023.

retrieval pipeline, any system failures would affect only the efficiency of trial enrollment rather than posing direct risks to patient safety. To manage the consistency of the generated responses, we leverage specific ChatGPT parameters, which are elaborated upon subsequently.

ChatGPT offers a range of parameters that can be tuned to influence the characteristics of its generated responses[9], such as *temperature*, *top_p*, *n*, *stream*, *stop*, *max_tokens*, *presence_penalty*, *frequency_penalty*, and *logit_bias*. Also, using the API, it is possible to alter the *system's role* parameter; this possibility has been exploited by a limited number of works in the literature [Deshpande et al., 2023, Qiu et al., 2023]. In our experiments, we have carefully set the system's role either using the specific variable that can be set through the API or by specifying it in the prompt text. The definition of the system's role has a direct impact on the provided responses, influencing both the content generated and the level of expertise reflected in those responses. It also imposes ethical boundaries for the generated responses. In some cases, the generated content can be significantly altered due to the selected system role, as recent empirical findings by Deshpande et al. [2023] suggest. Considering an example, a system role assigned as "friendly assistant" might use more casual language in the response. In contrast, a role defined as a "professional consultant" might lead the system to generate responses in a more formal language. The study we performed also aims to investigate a high-level distinction, as we set the system's role to either a *general assistant* or a *medical assistant*. According to the American Association of Medical Assistants (AAMA), medical assistants are involved in various administrative duties, among which is that of updating and filing patient medical records[10]. Therefore, in our experiments, the system's roles have been set as *medical assistant*; Qiu et al. [2023] in their study, also set the system's role as *medical assistant*.

Among the remaining parameters of ChatGPT, in our experiments, we modified *temperature*, *presence_penalty*, and *frequency_penalty* parameters according to the needs of the considered experiment, as we will explicitly comment in the following sections. Based on the documentation, *temperature* and *top_p* parameters should not be altered together, as they control the robustness of the model's response. Higher *temperature* values like 0.8 make the output more random, while lower values like 0.2 will make it more focused and deterministic. However, even a *temperature*

[9] OpenAI API, accessed on 21/04/2023.

[10] American Association of Medical Assistants, accessed on 18/04/2023.

value of zero may lead to small variability in the generated responses. Regarding the *presence_penalty*, and *frequency_penalty* these have a range between $[-2, 2]$. Positive *presence_penalty* values weight new tokens based on whether they appear in the text so far. As a result, the model is more likely to add new topics in the generated response. Positive *frequency_penalty* values decrease the model's ability to repeat the same tokens. To obtain comprehensive information regarding these and additional ChatGPT parameters, we direct the reader to the official API documentation[11].

In our approach we design prompts that can be classified into three categories based on their purpose. Those that guide ChatGPT to create queries for clinical trials retrieval (presented in Sections 6.2.1.1 and 6.2.1.2), those that extract specific information from clinical notes (Sections 6.2.1.3, 6.2.1.4 and 6.2.1.6), and those that identify medical entities and disambiguate their meaning (Section 6.2.1.5). Based on the assigned system role through the dedicated system's variable, the employed prompts are divided into those in which the system acts like a medical professional (Sections 6.2.1.1, 6.2.1.4 and 6.2.1.5, 6.2.1.6) and those in which the system has a generic role (Sections 6.2.1.2, 6.2.1.3). In one prompt, described in Section 6.2.1.3, the system does not have a domain-specific role nor knowledge about the general task to be performed, i.e. clinical trials retrieval. All employed prompts are zero-shot except the one presented in Section 6.2.1.6, which is a two-shot prompt. In most prompts (except the one presented in Section 6.2.1.3), the model has been discouraged from elaborating or reasoning upon its response. Lastly, in all prompts, the model is instructed to provide its answer in a specific format, i.e. list of terms or json-like format.

The following sections, particularly Tables 6.2 to 6.8, comprehensively describe the prompts used in our study. In each section, we comment on the motivation behind the usage of the particular prompt and the selected ChatGPT parameters, and we acknowledge the identified limitations and issues encountered in the obtained ChatGPT generated responses.

Moreover, to show the behavior of ChatGPT, we present as a qualitative example the models reply to the employed prompts when the following clinical note is used as input replacing the *Clinical Note* token when mentioned:

[11] OpenAI API frequency and presence penalties, accessed on 21/04/2023.

> *"Patient is a 45-year-old man with a history of anaplastic astrocytoma of the spine complicated by severe lower extremity weakness and urinary retention s/p Foley catheter, high-dose steroids, hypertension, and chronic pain. The tumor is located in the T-L spine, unresectable anaplastic astrocytoma s/p radiation. Complicated by progressive lower extremity weakness and urinary retention. The patient initially presented with RLE weakness where his right knee gave out with difficulty walking and right anterior thigh numbness. MRI showed a spinal cord conus mass which was biopsied and found to be anaplastic astrocytoma. Therapy included field radiation t10-l1 followed by 11 cycles of temozolomide 7 days on and 7 days off. This was followed by CPT-11 Weekly x4 with Avastin Q2 weeks/ 2 weeks rest and repeat cycle."*

6.2.1.1 Query Generation with Domain-Specific System Role and Task Description

The methodology outlined in this section integrates domain-specificity by configuring the system's role and incorporating a detailed task description in both the system's role and prompt texts. Specifically, as illustrated in Table 6.2, ChatGPT is instructed to function as a specialized medical assistant whose primary responsibility is to identify suitable clinical trials for a patient based on the supplied medical note. The specific objective of the task—namely, the retrieval of eligible clinical trials—is explicitly articulated in the prompt text and within the parameters defining the system's role. In this configuration, the model is designed to produce a single keyword-based query, an exemplar of which, the selected topic, is presented in the last row of Table 6.2. In this configuration, our objective is to give the model ample contextual information concerning its designated role and the specific task it is expected to accomplish. Additionally, we aim to afford the model a degree of freedom by incorporating phrases such as *"You can be synthetic and include all possible topics that you believe this patient can participate in, based on the provided medical note."*

This approach enables the model to generate more comprehensive and relevant queries. In addition, we set the *temperature* parameter equal to 0 to ensure that the generated response can be as deterministic as possible. The *frequency_penalty* is equal to 1.5 and the model is instructed through the prompt text to avoid having

Table 6.2: Prompt designed to generate queries for clinical trials retrieval. The table presents the assigned system's role, input prompt, and an example response. In this prompt, the system's role contains a detailed description of the user and the task.

System's Role (set through the API)
You are a helpful medical assistant that needs to retrieve eligible clinical trials for your medical patient. You can be synthetic and include all possible topics that you believe this patient can participate based on the provided medical note.
Prompt Text
From the following medical note, use all available information and also related medical concepts, symptoms, diseases, synonyms, or other information to create a keyword-based query that can be used to retrieve clinical trials. Do not include the same keywords twice. Respond with a comma-separated list of keywords that will be used for search. Do not explain or elaborate. Medical note: *{Clinical Note}*
Model's output (*Clinical Note* in Section 6.2.1)
"anaplastic astrocytoma", astrocytoma, pilocytic, spinal, "cord tumor", "urinary retention", "bladder dysfunction", "lower extremity weakness", "leg weakness", paraesthesia, "radiation therapy", surgery, temozolomide, hypertension, "chronic pain"

multiple occurrences of the same terms. Finally, we set the *presence_penalty* equal to 1, so that the model have more flexibility to include new terms. Due to the simplicity of the requested output, i.e. a comma-separated list of keywords, the generated responses were consistent and there was no need to curate them. As it can be seen in Table 6.2, some of the extracted information is in the form of bi-grams or even tri-grams. Some of them are explicitly mentioned in the provided clinical note (e.g. *anaplastic astrocytoma*), while others have been probably generated based on the content (e.g. *cord tumor*). Also, we observed that the generated response contains some new terms like *pilocytic* or *surgery*, probably generated due to the selected model parameters and the prompt instructions. However, in this qualitative example, one can observe that various patient-related information that might be important for the retrieval task has not been extracted (e.g. *high-dose steroids, CPT-11, Avastin*).

6.2.1.2 Query Generation with Generic System Role and Specified Task Description

In this approach, the system assumes a generic role, implying that the model lacks domain-specific expertise in the medical field. In the prompt presented in Table 6.3, the system functions as a general-purpose assistant, but its specific role is elaborated in the input prompt. The task is explicitly defined and includes details regarding the collection from which clinical trials are to be sourced. Initially, the model is directed to formulate a keyword-based query utilizing the information in the provided clinical note. Subsequently, within the same interaction, it is tasked with refining this query by incorporating synonyms or related medical concepts. Both of the requested outputs are created in a single interaction with the system. The parameters governing the model's behavior remain consistent with those employed in the preceding prompt, serving an analogous purpose (*temperature* is 0, *frequency_penalty* is 1.5, and *presence_penalty* is 1). Probably due to the prompt's complexity, this setting has two empirically identified shortcomings. First, the system fails to follow the instructions to avoid the term *clinical trial*, and the usage of abbreviations in the generated response. Second, for some topics, the *"[query_keywords]"* or *"[query_keywords_expanded]"* tokens were missing from the end of the generated responses.

Regarding the quality of the generated response in the demonstrated example, one can observe that the first query contains the patient's medical condition, and then the patient's medication and therapy. When the model is asked to refine the query, it successfully retained the original terms and expanded them with extra terms, including the term *neuropathic*, that was not present in the original clinical note. This extraction pattern and query creation is observed for the vast majority of the queries in the used collections. However, as one can observe based on the example model's output, various patient-related information that might be essential for this task is missing.

6.2.1.3 Two-step Information Extraction and Expansion with Generic System Role

In this process, the system operates under a generic role, without information related to the nature of the information to be extracted or the specific task this information is used. In this process, the system operates under a generic role, with

Table 6.3: Prompt aiming at query generation, with an explicit mention about the task to be performed and the system's role in the prompt text. The system's Role is generic, i.e. "You are a helpful assistant."

Prompt Text
Act as a medical assistant. Your task is to retrieve clinical trials from a registry of clinical trials in the United States. To achieve that, you have access to a medical clinical note of a patient. Follow my instructions precisely to extract the requested information from a patient's medical clinical note. Do not explain or elaborate. Respond with exactly what I request, and reply in the requested format. 1. Write a keyword-based query that can be used in a search engine to search for clinical trials in which this patient can participate. [query_keywords] Answers' format: [query_keywords] "query_text" [query_keywords] 2. Refine the query based on further details, such as synonyms or related medical concepts [query_keywords_expanded]. Answers' format: [query_keywords_expanded] "query_text" [query_keywords_expanded] Do not mention the terms clinical trial in the created queries as we search in a collection of clinical trials. Do not use abbreviations, use the resolved abbreviation format. Medical note: *{Clinical Note}*
Model's output (*Clinical Note* in Section 6.2.1)
[query_keywords] "clinical trial anaplastic astrocytoma spine radiation temozolomide Avastin CPT-11" [query_keywords] [query_keywords_expanded] "clinical trial anaplastic astrocytoma spine radiation temozolomide Avastin CPT-11 neuropathic pain lower extremity weakness urinary retention hypertension" [query_keywords_expanded]

neither the nature of the information to be extracted nor the specific task outlined. We employ a bi-phase process involving two separate interactions with the system, aiming first to extract a comprehensive list of keywords encapsulating the content of the provided text (Phase 1) and subsequently augment this list with semantically related keywords (Phase 2). As a result, the model generates two discrete sets of keywords.

The model's *temperature* is equal to zero, the *frequency_penalty* is set to 2 (maximum value) and the *presence_penalty* to 1. We have set the *frequency_penalty* to its maximum value due to the absence of instructions within the prompt to prevent the repetition of identical terms. Here, as one can observe in Table 6.4, the model outputs two sets of keywords (one per system interaction), the first contains extracted terms (in form of uni-grams, bi-grams, etc.) and the second that expands them with 20 additional related terms. The selection of the number of expansion terms is motivated by traditional pseudo-relevance feedback approaches that request a specific number of related terms to be added in the original query. By combining the model's outputs we run three distinct retrieval searches; one using the extracted terms, one only with the expanded, and one with their combination. Regarding the generated responses, as it can be seen under the model's

Table 6.4: Two-step process (one input-output per row) for information extraction from a clinical note text. In the first step, we instruct ChatGPT to extract the information, while in the second-step, we instruct it to expand these keywords. System's role:"You are a helpful assistant."

Prompt Text (1/2)
From the following text, extract a list of keywords that fully describe its content. Text:*{Clinical Note}*
Model's 1st output (*Clinical Note* in Section 6.2.1)
"45-year-old man", "anaplastic astrocytoma", spine, "lower extremity weakness", "urinary retention", "Foley catheter", "high-dose steroids", hypertension, "chronic pain", "T-L spine", radiation, "progressive lower extremity weakness", "RLE weakness", "right knee", "right anterior thigh numbness", "spinal cord conus mass", biopsied, "field radiation", t10-l1, temozolomide, CPT-11, Avastin
Prompt Text (2/2)
Expand the extracted list of keywords with 20 additional related keywords.
Model's 2nd output (*Clinical Note* in Section 6.2.1)
"progressive weakness", unresectable, cycles, tumor, chemotherapy, "neurologic examination", neurosurgery, oncology, "palliative care", "imaging studies", radiotherapy, corticosteroids, "nerve compression", "muscle weakness", "bladder dysfunction", "motor deficits", paraplegia, "spinal cord","cancer treatment"

1st output, Table 6.4, the extracted keywords and phrases indeed describe the information of the clinical note almost completely. Specifically, for the example query, only the bi-gram *"difficulty walking"* has not been extracted. This behavior has been observed for the majority of the queries in the employed collections. In the second interaction with the system, ChatGPT is instructed to expand the list of extracted keyword with related keywords. By observing the generated keywords (e.g. *"palliative care"* or *"imaging studies"*) one can conclude that some of the selected terms are describing broader medical concepts. As a result, it is expected that using these terms for retrieval might lead to decrease in precision-oriented measures.

6.2.1.4 Single-step Information Extraction and Expansion with Domain-Specific System Role

In this setting, domain specificity has been achieved by setting the system's role and by mentioning the search task (refer to Table 6.5). Specifically, ChatGPT's role has been set as a medical assistant, focusing on identifying medical conditions, treatments, and related terminology. This approach demands knowledge of medical terminology and abbreviations. Also, in this approach, information extraction and expansion have been instructed in a single step, and the purpose (i.e. search for clinical trials) is mentioned in the prompt ("identify clinical trials of interest").

The model's *temperature* is equal to zero, the *frequency_penalty* is set to 2 (maximum value) and the *presence_penalty* to 1. These values have been selected to reduce generation randomness (*temperature*), avoid the extraction of the same terms multiple times (*frequency_penalty*), and give the model some flexibility to generate new terms when perform query expansion. However, in this setting ChatGPT is instructed to extract only information related to the patient's medical condition and treatments.

Table 6.5: Single-step setting in which ChatGPT is instructed to extract only information related to the patient's medical condition and treatments. System's role:"You are a helpful medical assistant."

Prompt Text
Please identify the patient's medical condition and current treatments, including any alternative names, abbreviations, or synonyms for these terms, as well as any additional criteria that may be important for identifying clinical trials of interest. Respond with a comma-separated list of keywords that will be used for search. Do not elaborate or explain. Patient's medical note: *{Clinical Note}*
Model's output (*Clinical Note* in Section 6.2.1)
"anaplastic astrocytoma", spine, "lower extremity weakness", "urinary retention", "Foley catheter", "high-dose steroids", hypertension, "chronic pain", "T-L spine", radiation, "RLE weakness", temozolomide, "CPT-11 Weekly", Avastin

The selection of these particular entities to be extracted is motivated by our results presented in Section 6.1, in which the findings suggested that using a patient's medical problem, treatments and tests as query improves performance in clinical trials retrieval. The generated responses were consistent in terms of their format. However, although the model has been instructed to include alternative names, abbreviations, or synonyms, for the vast majority of queries, its responses only contain extracted keywords.

6.2.1.5 Directing ChatGPT for Information Extraction and Entity Meaning Disambiguation

A common limitation across the previously discussed prompt configurations is their ineffectiveness in disambiguating the semantics of the extracted keywords and phrases. In our research, we conceptualize clinical trials retrieval as a search task executed by a professional user. As we mentioned in Chapter 2, a distinctive characteristic of professional search is the necessity for end-users to have control over the search process. Attaining for that, we investigated ChatGPT's ability to summarize the patient's medical information by resolving abbreviations, recognizing key information, and adding pertinent MeSH terms (i.e. biomedical- and health-related terms that are synonyms to those present in the query). As a result, we have designed the prompt presented in Tables 6.6, that aims to extract information that is most commonly found within a patient's clinical note. The model's response to this prompts is presented in Table 6.7, due to space constraints. By employing this prompt, one can possibly extract a patient's lifestyle factors, a piece of information for which, to the best of our knowledge, there is not a publicly available model.

The prompt text clearly outlines the system's role and the search task in this configuration. The model is directed to extract patient-related information from the provided clinical note without further elaboration. Explicit instructions are also given to avoid responding if the requested information is not mentioned in the text. This preventive measure is informed by prior research indicating that ChatGPT may engage in inferential reasoning [Uzuner et al., 2011, Hu et al., 2023]. The model's *temperature* is equal to zero, and both the *frequency_penalty* and the *presence_penalty* are set to 0. These values have been selected to allow the model to use the same terms if needed, for instance when a patient's diagnosis and medical problem are identical. In this setting, we aim to leverage the diverse information extracted to formulate queries that provide end-users with explicit

Table 6.6: Medical entity extraction and meaning disambiguation (Part 1). System's Role: "You are a helpful medical assistant that needs to retrieve eligible clinical trials for your medical patient."

Prompt Text
Follow my instructions precisely to extract the requested information from a patient's medical clinical note. Do not explain or elaborate. Respond with exactly what I request, and reply in the requested format. From the following clinical note, resolve all the abbreviation mentioned in the text, and then extract: the patient's, age, gender, medical problem, diagnosis, diseases, symptoms, medications, drugs, dosages, treatments, medical history, family history, lifestyle factors, lab examinations, lab results, vital signs. Add MeSH terms that are relevant to the patient's medical problem, diagnosis or disease. If you are unable to extract the information, write 'N-A'. Answer in JSON format: {"answer":{ "abbreviations": "resolved abbreviations", "diagnosis":"", "medical_problem":"", "diseases":"", "drug":"", "dosages":"", "symptoms":"", "treatments":"", "medications":"", "medical_history":"", "family_history":"", "lifestyle_factors":"", "lab_examinations":"", "lab_results":"", "vital_signs":"", "gender":"", "age":"", "MeSH_terms":"" }} Medical note: *{Clinical Note}*

knowledge of the included entities for clinical trials retrieval. The rationale behind this query generation approach is identical to the one presented in Section 6.1 The primary distinction lies in utilizing a single LLM instead of multiple methods.

This prompt setting underscores ChatGPT's proficiency in extracting and disambiguating patient-related information. Nonetheless, our empirical analysis revealed several minor inconsistencies. One such issue pertains to formatting the generated output, particularly in the JSON structure, where elements like commas were occasionally missing. That is an insignificant issue that can be automatically rectified. Nonetheless, future versions of the GPT model, such as GPT-4 [OpenAI, 2023], attains more to the given instructions related to the model's output.

Additionally, the model exhibited limitations in its response characteristics. For instance, despite seemingly "acknowledging" its inability to provide certain infor-

Table 6.7: Medical entity extraction and meaning disambiguation, model's response (Part 2).

Model's output based on prompt 6.6 (*Clinical Note* in Section 6.2.1)
{ "answer": { "abbreviations": "s-p: status post, RLE: right lower extremity, MRI: magnetic resonance imaging, t10-l1: thoracic vertebrae 10 to lumbar vertebrae 1, CPT-11: irinotecan, Q2 weeks: every two weeks", "diagnosis": "anaplastic astrocytoma", "medical_problem": "anaplastic astrocytoma of the spine, severe lower extremity weakness and urinary retention", "diseases": "hypertension, chronic pain", "drug": "steroids, temozolomide, irinotecan, Avastin", "dosages": "high-dose steroids, temozolomide 7 days on and 7 days off, CPT-11 Weekly x4 with Avastin Q2 weeks- 2 weeks rest and repeat cycle", "symptoms": "lower extremity weakness, right anterior thigh numbness, urinary retention", "treatments": "field radiation, temozolomide, CPT-11 with Avastin", "medications": "steroids, temozolomide, irinotecan, Avastin", "medical_history": "anaplastic astrocytoma of the spine", "family_history": "N-A", "lifestyle_factors": "N-A", "lab_examinations": "N-A", "lab_results": "N-A", "vital_signs": "N-A", "gender": "male", "age": "45", "MeSH_terms": "Astrocytoma, Spinal Cord Neoplasms, Lower Extremity, Urinary Retention, Steroids, Temozolomide, Irinotecan, Bevacizumab, Radiation, Magnetic Resonance Imaging" }}

mation, the model failed to use the "N-A" token consistently. Instead, it resorted to verbose phrases like "this information is not explicitly mentioned in the text" or "lifestyle_factors: 'Not applicable'." Furthermore, the model occasionally included negated content in its responses, which leads to sub-optimal retrieval performance, as our empirical evaluation outlined in the previous section. To give some examples, the model's output for "lifestyle_factors" included terms such as "non-smoker, non-alcoholic, non-illicit drug user, menopausal." However, for information retrieval, the ideal output would solely consist of the term "menopausal," indicating a patient's current medical situation. Regarding the quality of the generated MeSH terms, prior research has indicated that these terms usually do not align with the official MeSH thesaurus [Wang et al., 2023b]. In our experiments, we have not investigated further towards this aspect. Another observation was the model's incomplete extraction of available information in a clinical note during the initial interaction.

Consequently, a multi-step interaction may be more effective for comprehensive information extraction, albeit at a higher operational cost. In our study, we utilized the prompt setting outlined in Table 6.6 to assess the efficacy of a cost-effective prompting approach for information extraction.

In summary, the findings from this prompting setting indicate that ChatGPT is capable of extracting and categorizing diverse medical information embedded within a clinical note. Furthermore, the model can be explicitly guided to refrain from inferring absent information, thereby mitigating the issue of hallucinatory output. As a result, ChatGPT can serve as a valuable tool for structuring clinical notes, thereby facilitating healthcare professionals by minimizing the time and effort needed for this task.

6.2.1.6 Combine ChatGPT with a Transformer-based Model for Negation Removal

Drawing upon the insights gained from previous prompting strategies, we have observed that ChatGPT has the ability to identify medical entities with minimal inaccuracies while also effectively disambiguating their meanings. However, as the model is trained to generate responses to questions, it occasionally provides negated content in its answers, especially in the setting presented in the previous section. That, as mentioned in previous studies and shown in our results, negatively impacts retrieval effectiveness [Chapman et al., 2001a, Koopman and Zuccon, 2014].

Therefore, this section presents an approach that combines ChatGPT with the pre-trained transformer model based on ClinicalBERT [Alsentzer et al., 2019] we leveraged in Section 6.1 [van Aken et al., 2021]. By doing that, we aim to leverage the power of ChatGPT for medical entity identification and the ability of the pre-trained transformer-based model to identify and remove negations. Our approach consist of four steps as shown in Figure 6.3.

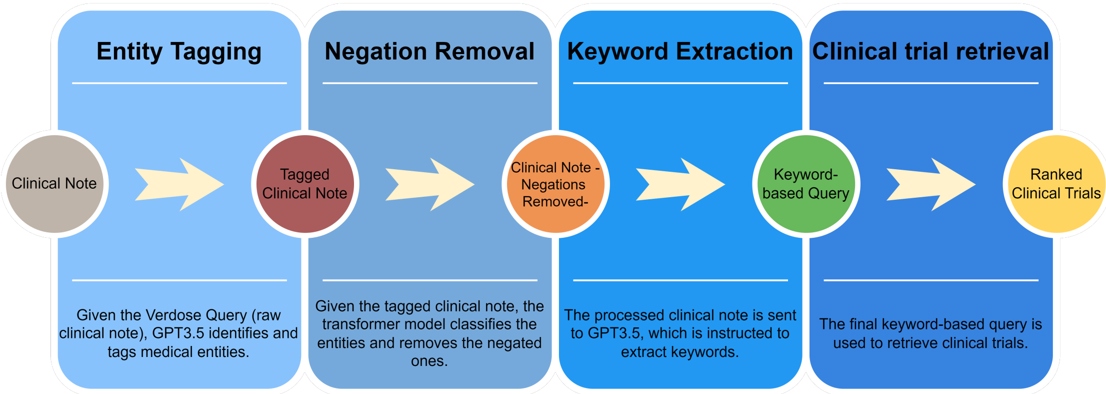

Figure 6.3: Combining ChatGPT with a negation classification model.

In detail, in the first step a clinical note is given as input to ChatGPT using the prompt text presented in the second row in Table 6.8. Here, ChatGPT is instructed to rewrite the clinical note by enclosing medical problems, treatments, tests, symptoms, and lifestyle factors within [entity] tokens. The *Temperature*, *frequency_penalty*, and *presence_penalty* variables are set to zero, to keep the answer as consistent as possible, and discourage the model to use different terms than those originally present in the clinical note. However, as it can be seen in the example response, there are entities that the model fails to tag, for example the *lower extremity weakness* entity. To further aid ChatGPT in this task, two examples of the requested annotation are also provided, motivated by the chain-of-though [Wei et al., 2022] prompting approach that uses a few examples in the prompt. By doing that, the clinical note is ready to be used as input to the fine-tuned model introduced by van Aken et al. [2021] in step two. The model classifies all tagged entities, and those found as *absent* (i.e. negated content) are removed. In the third step, a non-negated version of the clinical note is given as input to ChatGPT through the prompt text presented in the fourth row in Table 6.8. This step leverages the keyword extraction approach presented in Table 6.5, with the same model parameters (*temperature* is equal to zero, *frequency_penalty* is two and the *presence_penalty* is 1.). Although the same prompt is utilized in both cases, we observed small variations in the model's output. These differences are likely attributable to the absence of certain content, which could adversely affect textual coherence. In this example, the terms "therapy" and "weekly," are missing, while "T-L spine" has been transformed to "t10-l1 temozolomide."

Table 6.8: Approach that combines ChatGPT with the pre-trained model for negation detection. Each row corresponds to one ChatGPT input-output. System's role:"You are a helpful medical assistant."

Prompt Text (1/2)
Follow my instructions precisely to tag the requested information in a patient's medical clinical note. Do not explain or elaborate. Respond with exactly what I request, and reply in the requested format. Rewrite the clinical note by adding the token [entity] at the begging and the token [entity] at the end of the mentioned medical problems, symptoms, lifestyle factors, medical treatments and medical tests. For example, Input: The patient recovered during the night and now denies any shortness of breath. Output: The patient recovered during the night and now denies any [entity] shortness of breath [entity]. Input: An 8-year-old male presents in March to the ER with fever up to 39 C, dyspnea and cough for 2 days Output: An 8-year-old male presents in March to the ER with [entity] fever [entity] up to 39 C, [entity] dyspnea [entity] and [entity] cough [entity] for 2 days. Do the same for the clinical note: *{Clinical Note}*
Model's 1st output (*Clinical Note* in Section 6.2.1)
[entity] Patient [entity] is a 45-year-old man with a history of [entity] anaplastic astrocytoma [entity] of the spine complicated by severe lower extremity weakness and urinary retention [entity] s/p Foley catheter [entity], [entity] high-dose steroids [entity], [entity] hypertension [entity], and chronic pain. The tumor is located in the [entity] T-L spine [entity], [entity] unresectable anaplastic astrocytoma [entity] s/p radiation. Complicated by progressive lower extremity weakness and urinary retention. Patient initially presented with [entity] RLE weakness [entity] where his right knee gave out with difficulty walking and right anterior thigh numbness. MRI showed a spinal cord conus mass which was biopsied and found to be [entity] anaplastic astrocytoma [entity]. Therapy included field radiation [entity] t10-l1 [entity] followed by 11 cycles of temozolomide 7 days on and 7 days off. This was followed by [entity] CPT-11 Weekly x4 [entity] with [entity] Avastin [entity] Q2 weeks/2 weeks rest and repeat cycle.7
Prompt Text (2/2)
Please identify the patient's medical condition and current treatments, including any alternative names, abbreviations, or synonyms for these terms, as well as any additional criteria that may be important for identifying clinical trials of interest. Respond with a comma-separated list of keywords that will be used for search. Do not elaborate or explain. Patient's medical note: *{Clinical Note}*
Model's 2nd output (*Clinical Note* in Section 6.2.1)
"anaplastic astrocytoma", spine, 'lower extremity weakness", "urinary retention", "Foley catheter", "high-dose steroids", hypertension, "chronic pain", "radiation therapy", "t10-l1 temozolomide", CPT-11, Avastin

Finally, the extracted keywords are used as a query for retrieval, leading to a ranked list of clinical trials.

Prior to our decision to employ this setting to remove negated content from a clinical note, we tried to directly instruct ChatGPT to perform this task. However, the obtain results did not meet our expectations. In detail, we experimented with prompts like *"Remove the negated content of the following medical note. "* or *"re-write the following medical note by removing the negated medical problems, symptoms..."*. The models responses, were totally altering the meaning of the original clinical note, for instance: Given the original text *"He is healthy with no history of allergies..."*, the model identified the negation in the sentence, but its output was *"He is healthy with history of allergies..."*; therefore, we proceeded with the presented approach.

6.2.2 Experimental Design and Results

This section provides a thorough overview of the experimental framework, encapsulating the evaluation metrics, the rationale behind their selection, and the data acquisition and indexing methods. It concludes with a synopsis of the range of experiments executed in this study, as summarized in Table 6.9 and presented in the subsequent sections.

Evaluation Metrics. We employ a variety of evaluation measures, including nDCG@5, nDCG@10, P@5, P@10, R-Precision (Rprec), and Mean Reciprocal Rank (MRR). We adhere to the official guidelines and report Bpref and P@25 to ensure a comprehensive evaluation. It is worth noting that, except for nDCG, all other metrics consider only *eligible* trials as relevant, following the official guidelines. Readers are referred to Chapter 2.1.3 for an in-depth presentation of the selected measures.

Data Collection and Indexing. The benchmark collections (TREC 2021 and 2022) for our experiments were obtained from the *ir-datasets* [MacAvaney et al., 2021]. We utilized PyTerrier [Macdonald and Tonellotto, 2020] for indexing, adhering to its default parameters such as porter-stemming and stopword removal. Each document is indexed in its entirety, i.e. including all available sections.

Experiments (Baselines). For clinical trials retrieval, our approach relies on the BM25 model, in line with top-performing approaches in TREC 2021 and 2022 tracks, as mentioned in Section 4.2.2. Specifically, we leverage PyTerrier's implementation of the BM25 model, with its default parameters, in all our retrieval experiments. As baseline in our empirical evaluation we leverage the original query text with the BM25 model. We refer to this experiment as *BM25*. Additionally, we utilize KeyBERT for keyword extraction from the original queries and use them for retrieval. This experimental approach is referred to as *KeyBERT* and also functions as a baseline. In this experiment we used the original KeyBERT implementation and set its parameters so that it extracts twenty uni-grams or bi-grams (*keyphrase_ngram_range=(1, 2)*, *top_n=20*).

Experiments (Ours). Regarding our experiments that incorporate ChatGPT, they are organized into four specific categories based on the intended purpose of the employed prompts: query generation (*QGMT* and *QGGT*), information extraction (*IEG* and *IEMT*), targeted extraction of patient-related information (*IEMDMT*), and lastly, our hybrid methodology for negation removal (*NRIEMT*). Each original query is initially processed through ChatGPT based on a prompt discussed in the previous sections. Each query is handled in an isolated system interaction through the API, establishing a distinct conversational context. ChatGPT leverages the gpt-3.5-turbo model, while the parameters associated with each experiment have been elaborated upon in earlier sections. Subsequently, these generated queries are subjected to PyTerrier's standard preprocessing steps, including porter-stemming and stopword removal, before being employed for clinical trial retrieval. In our experiments, the queries generated by both ChatGPT and KeyBERT incorporated n-grams. Specifically, we restricted our focus to uni-grams and bi-grams, effectively treating tri-grams or larger n-grams as uni-grams. Finally, the RM3 model is employed for pseudo-relevance feedback-based query expansion and is integrated into all previously mentioned retrieval pipelines. We utilize its PyTerrier implementation, specifying the number of feedback documents as ten and the number of expansion terms as twenty.

Each section in the following evaluates the retrieval performance achieved by queries generated under specific instructions given to ChatGPT. Section 6.2.2.1 evaluates the queries it generates when asked to create queries for retrieval. Section 6.2.2.2 those obtained when asked to extract essential patient information from clinical notes. Section 6.2.2.3 evaluates retrieval performance based on the combinations of several medical entities in distinct query representations. Finally, Section 6.2.2.4

Table 6.9: Summary of the conducted experiments.

Experiment Details (IDs)	Query	Details
	Baseline Experiments	
BM25 (**BM25**)	Original query with the default PyTerrier's pre-processing.	-
KeyBERT+BM25 (**KeyBERT**)	Keywords extracted from the original query using KeyBERT.	-
	ChatGPT Experiments (Ours)	
Query **G**eneration, **M**edical Role & **T**ask Description (**QGMT**)	A single keyword-based query.	Section 6.2.1.1
Query **G**eneration, **G**eneric Role & **T**ask Description (**QGGT**)	This prompt outputs two sets of queries. A single keyword-based query, and a refined query that contains novel terms.	Section 6.2.1.2
Information **E**xtraction, **G**eneric Role (**IEG**)	This prompt outputs two sets of queries. One with all extracted keywords and phrases and one with novel expansion terms.	Section 6.2.1.3
Information **E**xtraction, **M**edical Role & **T**ask Description (**IEMT**)	A single keyword-based query.	Section 6.2.1.4
Information **E**xtraction & Entity **M**eaning **D**isambiguation, **M**edical Role & **T**ask Description (**IEMDMT**)	This prompt outputs various patient-related information. Therefore, many queries variations are constructed and used for retrieval.	Section 6.2.1.5
Negation **R**emoval and ChatGPT **I**nformation **E**xtraction, **M**edical Role & **T**ask Description (**NRIEMT**)	A single keyword-based query without negated terms.	Section 6.2.1.6

reports the impact of refining queries by removing potential negated entities.

6.2.2.1 Retrieval Effectiveness of ChatGPT-Generated Queries

The results presented in Table 6.10 show the retrieval performance achieved by prompts that instruct ChatGPT to generate a single query (per clinical note) to be used for clinical trials retrieval (presented in Sections 6.2.1.1 and 6.2.1.2).

First we comment on the retrieval performance achieved by the BM25 and KeyBERT retrieval pipelines (i.e. the baselines), with and without the incorporation of the RM3 model for query expansion. One can notice that the performance improvements are not consistent between the two collections. In the TREC 2021 collection, KeyBERT outperforms BM25 only in terms of Bpref, while in TREC 2022 KeyBERT underperforms only for MRR. This performance variations might occur due to the retrieval of unjudged documents in the top ranking positions, which are considered as not relevant during our evaluation. Another reason might be the semantic characteristics of the provided queries in the two collections, for instance the number of hard and easy queries. As hard queries we assume queries for which the patient might be an infant, or the considered disease might be very rare. However, we have not investigated further towards these directions in our study. Combining BM25 with the RM3 model (i.e. BM25 + RM3), improves or has a slightly lower performance compared to BM25 across all measures, except MRR where the observed decrease is greater but not statistically significant. All in all, no single baseline approach consistently outperforms the others across all metrics and collections. Also, incorporating the RM3 model has mixed effects on the performance achieved by BM25 and KeyBERT, as it is (probably) affected by the number of relevant documents in the top ten ranking positions.

Table 6.10: Retrieval performance across the two benchmark collections, by the prompts related to query generation for clinical trials retrieval.

	TREC 2021						TREC 2022					
	Rprec	Bpref	P@10	P@25	MRR	nDCG@10	Rprec	Bpref	P@10	P@25	MRR	nDCG@10
BM25	.162	.184	.264	.211	**.471**	.469	.180	.172	.272	.235	.507	.394
BM25 + RM3	**.184**	.241°	**.285**	**.231**	.463	**.484**	.208	.212	.270	.253	.427	.389
KeyBERT	.150	.209°	.217	.183	.405	.386°	.192	.199	.274	.246	.449	.397
KeyBERT + RM3	.151	.220°	.203	.175	.385	.371°	.218	.228°	.262	.254	.401	.384
QGMT	.170	.238°	.236	.197	.470	.393	.232°	.265°	.346	.273	.547	.471
QGMT + RM3	.181	**.257**°	.265	.221	.414	.409	**.262**°	.301°	**.368**	**.323**	.534	**.498**
QGGT (Initial)	.151	.227	.211	.167	.394	.326°	.234	.269°	.340	.303	.526	.452
QGGT (Initial) + RM3	.166	.244°	.217	.188	.405	.342°	.260°	.297°	.364	.318	.499	.477
QGGT (Initial & Refined)	.132	.220	.204	.165	.403	.307°	.221	.281°	.314	.278	**.566**	.417
QGGT (Initial & Refined) + RM3	.146	.245°	.229	.191	.444	.348°	.243°	**.307**°	.344	.296	.506	.441

The remaining rows of Table 6.10 show the retrieval performance achieved by the usage of the ChatGPT generated queries with BM25 and the standard PyTerrier query processing steps. We remind that the acronym *QGMT* stands for **Q**uery **G**eneration, **M**edical Role & **T**ask Description, as described in Section 6.2.1.1. In this approach, thoroughly described in Section 6.2.1.1, the prompt provides contextual information to ChatGPT and instructs it to generate a single keyword-based query. The acronym *QGGT* refers to **Q**uery **G**eneration, **G**eneric Role & **T**ask Description prompts, which has been presented in Section 6.2.1.2. Here, a single prompt is employed to generate two distinct keyword-based queries namely, *Initial* and *Refined*. We run a total of four experiments; two exploit the *Initial* query, and two exploit the concatenation of the *Initial* and the *Refined* queries.

In our experiments, we found that using solely the *Refined* query, which mostly consists of novel terms (not included in the original query), lead to great decreases in performance across all measures and collections. Specifically, P@10 for the TREC 2021 collection was .195 and .212 when RM3 is used for query expansion; similarly for the TREC 2022 collection. This finding suggests that the new terms added to the *Refined* query have probably a broader semantic meaning and lead to topic drift.

The QGMT experiment outperformed all of the QGGT related experiments in the TREC 2021 collection and the QGMT + RM3 outperformed all of the QGGT related experiments in both collections. Also, QGMT + RM3 shows statistically significant improvements in Bpref for both TREC 2021 and TREC 2022 collections. In general, this finding suggests that exploiting the ChatGPT's variable that assigns a particular role to the AI system leads to different responses. This finding is inline with related studies [Deshpande et al., 2023]. To further support the validity of the above finding, we run the QGMT experiment by setting the system's role to generic, i.e. "You are a helpful assistant." The obtained results are presented in Table 6.11. It can be seen that when the system's role is generic, the retrieval performance decreases. By observing the generated queries, we noticed that the model's behavior has also changed; specifically, with the generic role, the first extracted terms for the majority of the queries are the patient's age and gender. When the role is the domain-specific, the first extracted terms are related to the patient's medical problem.

Table 6.11: Retrieval performance across the two benchmark collections, using the prompt presented in Section 6.2.1.1 with different system's role.

TREC 2021						TREC 2022					
Employed Prompt (refer to Section 6.2.1.1) (System's role: as in Table 6.2)											
Rprec	Bpref	P@10	P@25	MRR	nDCG@10	Rprec	Bpref	P@10	P@25	MRR	nDCG@10
.170	**.238**	**.236**	**.197**	**.470**	.393	**.232**	**.265**	**.346**	.273	**.547**	**.471**
Employed Prompt (refer to Section 6.2.1.1) (System's role: "You are a helpful assistant")											
Rprec	Bpref	P@10	P@25	MRR	nDCG@10	Rprec	Bpref	P@10	P@25	MRR	nDCG@10
.160	.195	.215	.189	.454	**.443**	.216	.211	.324	**.281**	.518	.468

To conclude, our findings suggest that instructing ChatGPT to directly generate queries for clinical trials retrieval has some potential, although the generated queries lead to inconsistent retrieval results across the different collections. Also, by experimenting with the assigned role to the AI system, we observed that it is better if the role is related to the considered domain.

6.2.2.2 Retrieval Effectiveness of Synthesized Queries Based on Information Extracted by ChatGPT

This section presents the results obtained by the prompts presented in Sections 6.2.1.3 and 6.2.1.4. Both of these approaches guide ChatGPT to perform information extraction from a given text, and their main differences rely on the number of system interactions and the level of the provided contextual information (i.e. related to the task and to the system's role).

Table 6.12: Retrieval performance across the two benchmark collections, using prompts that guide ChatGPT to extract keywords.

	TREC 2021						TREC 2022					
	Rprec	Bpref	P@10	P@25	MRR	nDCG@10	Rprec	Bpref	P@10	P@25	MRR	nDCG@10
BM25	.162	.184	.264	.211	.471	.469	.180	.172	.272	.235	.507	.394
BM25 + RM3	.184	.241°	.285	.231	.463	.484	.208	.212	.270	.253	.427	.389
KeyBERT	.150	.209	.217	.183	.405	.386	.192	.199	.274	.246	.449	.397
KeyBERT + RM3	.151	.220	.203	.175	.385	.371	.218	.228	.262	.254	.401	.384
IEG (Extracted)	.172	.214°	.260	.219	.486	.442	.234°	.229°	.360°	.300°	.561	.495°
IEG (Extracted) + RM3	.196°	.258°	.299	.243	.498	.463	.252°	.278°	.360	.322°	**.618**	.495°
IEG (Extracted & Expanded)	.163	.224°	.252	.205	.500	.424	.211	.242°	.350	.294	.570	.476
IEG (Ext. & Exp.) + RM3	.165	.258°	.248	.203	.470	.414	.240	.282°	.346	.309°	.538	.472
IEMT	.195°	.250°	.273	.240	.502	.470	.250°	.270°	.358°	.308°	.609	.505°
IEMT + RM3	**.212°**	**.275°**	**.323°**	**.261°**	**.541**	**.512**	**.276°**	**.298°**	**.372°**	**.338°**	.576	**.517°**

We remind that the first four rows of Table 6.12 are are the same in all the presented results tables, as they concern our baselines. The *IEG* abbreviation stands for **I**nformation **E**xtraction, **G**eneric **R**ole and refers to the prompt presented in Section 6.2.1.3; the *IEMT* (**I**nformation **E**xtraction, **M**edical Role & **T**ask Description) refers to the prompt presented in Section 6.2.1.4.

Inline with the previous findings, providing the AI system with clear domain and task information leads to better responses, and, as a result, to better retrieval performance. In detail, the IEMT experiment outperforms both the IEG (Extracted) and the IEG (Extracted & Expanded) across all measures and both collections. In addition, IEMT performs equally or better to the BM25 + RM3 baseline in TREC 2021 collection and outperforms it in TREC 2022. From these results,

Table 6.13: Retrieval performance across the two benchmark collections, by a query that is formulated by combing the extracted 'diagnosis', 'medical_problem', 'diseases', 'drug', 'symptoms', 'treatments', 'medications', 'lab_examinations', identified by ChatGPT and expanded by suggested 'MeSH_terms'.

	TREC 2021						TREC 2022					
	Rprec	Bpref	P@10	P@25	MRR	nDCG@10	Rprec	Bpref	P@10	P@25	MRR	nDCG@10
BM25	.162	.184	.264	.211	.471	.469	.180	.172	.272	.235	.507	.394
BM25 + RM3	.184	.241°	**.285**	**.231**	.463	**.484**	.208	.212	.270	.253	.427	.389
KeyBERT	.150	.209	.217	.183	.405	.386	.192	.199	.274	.246	.449	.397
KeyBERT + RM3	.151	.220	.203	.175	.385	.371	.218	.228	.262	.254	.401	.384
IEMDMT	.174	.238°	.237	.211	**.485**	.397°	.247°	.260°	**.348°**	.285	.528	.458
IEMDMT + RM3	**.185**	**.268°**	.264	.222	**.485**	.428	**.258°**	**.285°**	.328	**.311°**	**.560**	**.461**

the observed increase in Bpref is statistical significant in both collections. With the incorporation of the RM3 model for query expansion, we observe that the performance of IEMT (experiment IEMT + RM3) shows a statistical significant increase over the BM25 baseline for the majority of the presented measures and across both of the employed collections.

Focusing on the IEG (Extracted) and the IEG (Extracted & Expanded) experiments, one can observe that the latter decreases the retrieval performance across most of the presented measures. This finding further supports our observation that when ChatGPT is instructed to expand or refine its response, it tends to add terms with a broader semantic meaning leading to topical drift and performance decrease.

6.2.2.3 Retrieval Effectiveness of Queries Containing Medical Entities Extracted and Annotated by ChatGPT

By leveraging the prompt presented in Section 6.2.1.5 we extracted and disambiguate the meaning of various patient-related information from a clinical note. Then, based on our previous findings (refer to Section 6.1), we synthesized queries by combining the extracted information in various possible combinations. To give an overview, we estimated the retrieval performance achieved by combining the identified diseases and medical problems with symptoms, or combining diagnosis, problems and diseases with treatments we also tried many other combinations. In Table 6.13 we present the results obtained by the query that achieved the highest Rprec in both collections (experiment IEMDMT). We remind that *IEMDMT* stands for **I**nformation **E**xtraction & Entity **M**eaning **D**isambiguation, **M**edical Role & **T**ask Description. The query contains the unique terms obtained by concatenating the extracted medical diagnosis, problem, diseases, drugs, symptoms, treatments, medications, examinations, along with the generated MeSH terms. Although Wang et al. [2023b] noticed that ChatGPT suggests non-existing or poor quality MeSH Terms, our experiments suggest that including the suggested MeSH terms in the query leads to small percentage improvements in performance (around 6% in TREC 2021 and 9% for TREC 2022). In contrast, resolving the abbreviations hurt retrieval effectiveness. However, we leave a more detailed analysis related to the quality of the suggested MeSH terms and the accuracy of the resolved abbreviations for future work. Including information related to a patient's past medical history, family history, lab results, drug dosages, and lifestyle factors in the synthesized query, lead to decreases in performance.

Based on previous observations reported in Section 6.2.1.5, we suspect that the reasons underlying the observed performance decreases might be related to the following reasons. First, the decreases might be due to the model's responses when a clinical note did not contain information related to the past medical history, family history, e.g. *"not applicable."* In fact, ChatGPT did not identify information about a patient's family or medical history for several clinical notes used in our study. Another reason might be related to the inclusion of negated answers, that have been observed during the extraction of a patient's lifestyle factors. To conclude, the reported retrieval performance using this prompt, might have been underestimated due to the format of the model's responses. Therefore, further research could improve the information extraction process, by perhaps exploiting another prompting technique or extracting the information with more interactions with the system.

6.2.2.4 Retrieval Effectiveness of Queries Containing Non-Negated Medical Entities

The results presented in this section correspond to the hybrid approach, presented in Section 6.2.1.6, that exploits a pre-trained language model for negation identification and ChatGPT for information extraction. We remind that the presented experiment *NRIEMT*, leverages the same prompt with the IEMT experiment presented in Section 6.2.2.2. This prompt instructs the model to extract information by providing all contextual information related to the task and the system's role. Therefore, Table 6.14 repeats the retrieval performance achieved by IEMT, for comparison

Table 6.14: Retrieval performance across the two benchmark collections, by the prompts related to keyword extraction applied after negated content has been removed from the original queries. For comparison purposes we present also the IEMT prompt setting.

	TREC 2021						TREC 2022					
	Rprec	Bpref	P@10	P@25	MRR	nDCG@10	Rprec	Bpref	P@10	P@25	MRR	nDCG@10
BM25	.162	.184	.264	.211	.471	.469	.180	.172	.272	.235	.507	.394
BM25 + RM3	.184	.241°	.285	.231	.463	.484	.208	.212	.270	.253	.427	.389
KeyBERT	.150	.209	.217	.183	.405	.386	.192	.199	.274	.246	.449	.397
KeyBERT + RM3	.151	.220	.203	.175	.385	.371	.218	.228	.262	.254	.401	.384
IEMT	.195°	.250°	.273	.240	.502	.470	.250°	.270°	.358°	.308°	.609	.505°
IEMT + RM3	**.212°**	**.275°**	**.323°**	**.261°**	**.541**	**.512**	**.276°**	**.298°**	.372°	**.338°**	.576	**.517°**
NRIEMT	.185	.229°	.260	.221	.508	.449	.247°	.249°	**.374°**	.315°	.598	.496°
NRIEMT + RM3	.191	.255°	.277	.241	.491	.466	.272°	.285°	.372°	**.338°**	**.623**	.509°

purposes.

The NRIEMT experiment aims to investigate whether removing negated information prior to performing IE for a clinical note leads to better retrieval performance based on the premise that the generated queries will be focused on the patient's existing conditions. An effectiveness indicator would have been an increase in all measures, but specifically in MRR, P@10 and P@25. However, this is not the case in our experiments. Firstly, both NRIEMT and NRIEMT + RM3 experiments, show statistically significant improvements over the BM25 baseline in both collections, suggesting that this prompting method is stable. By comparing the IEMT and NRIEMT experiments, one can observe that the results are not consistent across collections. For TREC 2021, precision drops, while in TREC 2022 we have small and not statistically significant improvements. A plausible reason might be the characteristics of the queries in each collection, i.e. for TREC 2022 there might be more negated medical entities and their removal might positively impact retrieval performance for those queries. Another reason, might be that ChatGPT generated slightly different responses (do not extract the same terms as IEMT), especially for queries from which various information have been removed. In any case, the presented findings encourage further exploration towards this research direction, i.e. how one can exploit LLMs for negation handling in clinical text.

6.2.3 Summary of Findings and Discussion

In this section we summarize the main contributions and findings of the investigation presented in this section. Firstly, to summarize the findings of our work, we rank the conducted experiments based on their achieved Rprec measure (refer to Table 6.15). We select Rprec as it has been empirically found that it is highly correlated with MAP [Manning et al., 2008] and therefore provides a better understanding of the quality of the obtained document ranking. Then, we compare our best performing experiment with the best-performing information extraction approach presented in Section 6.1, and with the best performing approaches proposed in TREC 2021 and 2022. Finally, we conduct additional experiments to investigate if ChatGPT generated queries can reach the retrieval performance achieved when human-generated queries are used for clinical trials retrieval.

By ordering our experiments based on their achieved performance, one notices that for TREC 2022, all of the conducted experiments outperform both of the selected baselines (BM25 and KeyBERT). For TREC 2021 it can be seen that only the information extraction prompts lead to better Rprec values. Across collections, the prompts that instruct ChatGPT to extract information from the clinical notes (those whose abbreviation starts with "IE" and "NRIE,") lead to greater Rprec values than those that instruct it to generate queries (abbr. starts with "QG"). The only exception is the IEG (Extracted & Expanded) experiment, which in TREC 2021 underperforms compared to BM25 and in TREC 2022 it has the lower Rprec value among our experiments. Thus, ChatGPT can be used to process unstructured clinical notes, but it is preferred to be instructed to extract patient-related information rather than generate queries for clinical trials retrieval.

Table 6.15: Ranking of experiments based on their achieved Rprec across the two collections. Refer to Table 6.9 for more information about the used abbreviations. The value in brackets is the achieved Rprec.

Rprec - TREC 2021	Rprec - TREC 2022
QGGT (Initial & Refined) [.132]	**BM25 [.180]**
KeyBERT [.150]	KeyBERT [.192]
QGGT (Initial) [.151]	IEG (Extracted & Expanded) [.211]
IEG (Extracted & Expanded) [.162]	QGGT (Initial & Refined) [.221]
BM25 [.162]	QGMT [.232]
QGMT [.170]	QGGT (Initial) [.234]
IEG(Extracted) [.172]	IEG(Extracted) [.234]
IEMDMT [.174]	IEMDMT [.247]
NRIEMT [.185]	NRIEMT [.247]
IEMT [.195]	IEMT [.250]

Another finding is highlighted by the Rprec achieved by the IEG (Extracted &

Table 6.16: Comparison of ChatGPT to transformer-based and rule-based approaches.

	TREC 2021		
	Bpref	P@5	P@25
$Q9_{Q4\_rem\_fam}$ (Section 6.1)	.212	**.331**	.216
IEMDMT	.238	.269	.211
IEMT	**.250**	**.331**	**.240**

Expanded) and the QGGT (Initial & Refined) experiments. As it can be seen in the table, both of these experiments have poor performances in both collections. We remind that in these prompts, ChatGPT has been instructed to "Expand the extracted list of keywords with 20 additional related keywords" (IEG) and "Refine the query based on further details, such as synonyms or related medical concepts." (QGGT). Through the presented examples we noticed that the selected expansion terms are related to more generic medical concepts than those mentioned in the original clinical note. Therefore, based on the experiments conducted in this study, it is concluded that query expansion with ChatGPT leads to poor performance. Future studies should further investigate towards this direction aiming to investigate whether ChatGPT can be instructed to generate more specific medical concepts.

Regarding the NRIEMT experiment, it did not meet our expectation to improve the performance of the IEMT experiment by removing negated medical entities. In addition, our attempts to remove the negated content of a clinical note using ChatGPT, highlighted some warning behavior. Specifically, as we have briefly mentioned in Section 6.2.1.6, we have observed that ChatGPT altered the meaning of the original text, rather than removing the negated medical entities. In the studied task, this behavior might have a small impact, as it will only lead to a performance decrease during clinical trials retrieval. However, as the employment of LLMs as assistants to conduct medical bureaucratic tasks is supported by various organizations, this behavior might cause serious problems in other applications.

Table 6.16 presents a comparison of the optimal method identified in Section 6.1; i.e.the query that includes medical entities pertinent to a patient's issues, treatments, and tests, while excluding family history and negated entities. The performance achieved by this query is compared to the IEMDMT and the IEMT experiments,i.e. those that guide ChatGPT to do the same, over TREC 2021 collection. As it can be seen in the table, the IEMDMT prompt improves the retrieval performance over the previous study only in terms of Bpref. This finding suggests that ChatGPT, with the prompt used in this study, does not extracts patient-related information as accurately as the previously employed domain-specific state-of-the-art methods. However, in previous sections we have identified various reasons that might have underestimated the performance achieved by IEMDMT and point out various research directions that might improve it. Nonetheless, the best performing prompt, i.e. IEMT, improves or has the same performance as the one reported in Section 6.1. To conclude, it seems that the employed prompt (IEMDMT) is not sufficient to capture all the semantic information that is present in a clinical note, as good as the other employed methods.

6.2.3.1 Comparison with the SoA Approaches

This section compares the retrieval performance achieved by the top performing experiments conducted in our research study with the SoA approaches in TREC 2021 and TREC 2022 (refer to Table 6.17). Unfortunately, our understanding around the TREC 2022 SoA approach is limited as the participating team did not provide a detailed description of their experimental design.

The SoA approach of TREC 2021, conducted by the same team, is described by Pradeep et al. [2022] and it has been briefly analyzed in Section 4.2.1.1. The approach is a multi-stage neural ranking approach. Here, we further analyze the approach by focusing on the initial retrieval phase. Specifically, given a clinical note, the authors leverage a neural query synthesis (NQS) method (i.e. a zero-shot document expansion model) to generate forty sentence-long queries. Each of these queries is used independently, in addition to the original clinical note, as input in a retrieval pipeline that exploits the BM25 and RM3 models, and the obtained results are fused. The aforementioned process is their first-stage retrieval method, whose results are presented in the second row of Table 6.17 (“Neural Query Synthesis”).

Table 6.17: Comparison between our best performing experiments with the state-of-the-art approaches of TREC 2021 and TREC 2022.

	TREC 2021				TREC 2022		
	nDCG@10	P@10	MRR		nDCG@10	P@10	MRR
TREC's Median	.304	.161	.294	TREC's Median	.392	.258	.411
Neural Query Synthesis	.473	.276	.434	*frocchio* run	.463	.324	.537
monoT5 *'CT*	**.712**	**.593**	**.816**	*frocchio_monot5_e* run	**.613**	**.508**	**.726**
IEMT+RM3	.512	.323	.541	NRIEMT + RM3	.509	.372	.623

Then, a neural re-ranker based on the monoT5 model, fine-tuned for the task of clinical trials retrieval is employed to create the final ranking. The achieved performance is presented in the third row of Table 6.17 ("monoT5 *'CT*"). As it can be seen, the IEMT+RM3 experiment leads to better retrieval effectiveness compared to the NQS (first-stage retrieval) approach. We further note that even the IEMT experiment reaches similar performance for nDCG@10 and P@10 compared to the proposed NQS approach. Also, IEMT reaches higher MRR (.502) than NQS. In addition, the IEMT and IEMT+RM3 experiments are single and two-stage retrieval approaches, respectively. In contrast, the NQS approach necessitates forty initial retrieval runs, involving the use of generated query variations. Consequently, our approaches exhibit lower complexity when compared to NQS, while at the same time achieve superior retrieval performance.

The SoA approach in the TREC 2022 Clinical Trials track, also employs the Mono-T5 model and it has been proposed by the same research team. Due to the lack of experimental details, it is not feasible to further comment on this approach. However, based on the submitted experiments, we can observe that our best performing experiment in the TREC 2022 collection outperforms the *frocchio* run, that was probably the employed first-stage retrieval (based on the run name). Nonetheless, also the IEMT experiment outperforms the *frocchio* run in all of the presented retrieval measures.

To conclude, based on the comparisons presented in Table 6.17, the experiments conducted in this work outperformed the first-stage retrieval approaches exploited in both the TREC 2021 and 2022 study. However, our experiments did not yield performance improvements compared to the SoA performance achieved by the monoT5 based neural re-ranker, as we leverage the BM25 model to estimate topical relevance. Nonetheless, when combined with our retrieval approach, these neural re-ranking models may enhance their retrieval performance.

6.2.3.2 Comparison with Human-generated Queries

The objective of this section is to investigate if the retrieval performance achieved with ChatGPT-generated queries can match that of human-generated queries in the context of clinical trials retrieval. To achieve that we exploit the retrieval collection introduced Koopman and Zuccon [2016], i.e. the Clinical collection presented in Section 4.3. The provided collection bears resemblance to the ones used in the TREC 2021 and TREC 2022 clinical trials tracks. In their study, Koopman and Zuccon [2016] asked four medical assessors to provide, for each patient case, several ad-hoc keyword queries that they would issue to a search engine to find clinical trials. As reported, the final collection contains a total of 489 unique queries, with 8.2 keyword queries per patient case created by the four medical assessors. Hereafter we refer to the medical assessors as assessor A, B, C, and D.

To investigate whether the ChatGPT-generated queries can match that of human-generated queries we conduct the following experiments. First, for each patient case in the collection, we concatenate the n ad-hoc keyword queries created by each individual assessor into a single ad-hoc keyword query which contains all the unique keywords. For example, assuming that the assessor A has created two distinct ad-hoc keyword queries $q1 = keyword1$, $keyword2$, $keyword3$ and $q2 = keyword2$, $keyword4$ for a specific patient case, the final query that is used for retrieval will be $qf = keyword1$, $keyword2$, $keyword3$, $keyword4$. The created qf simulates a scenario in which an expert user is presented with a patient's case and is asked to think of all possible clinical trials this patient can participate. Following this process for each assessor we have created four ad-hoc queries for each patient case.

Using the distinct created queries qfA, qfB, qfC, and qfD as inputs, we run four retrieval experiments based on the same experimental set-up presented in Section 6.2.2. However, medical assessors A, B and C have not provided ad-hoc keyword queries for 1, 3, and 1 patient cases respectively. In the conducted experiments, these patient cases have been removed from the evaluation. In addition, we concatenate qfA, qfB, qfC, and qfD into a single query that contains their unique keywords, i.e. $qfall$, and use this as input in another retrieval experiment. We assume that the $qfall$ query accumulates the knowledge of various medical experts in a single representation. Lastly, to generate queries with ChatGPT we employed the IEMT approach presented in Section 6.2.1.4.

To compare the retrieval effectiveness achieved by the human-generated queries with that of those generated by ChatGPT, we report the obtained P@10 and the Bpref measures in Table 6.18. For both measures we assume only the *eligible* clinical trials as relevant, i.e. trials assessed as *excludes* are considered as *not relevant*. As the authors report, in this benchmark collection, the number of assessed documents is limited and therefore the evaluation might be less reliable for new systems. To overcome this issue, we evaluate the retrieval performance by employing the condensed measures approach proposed by Sakai [2007] as a way to deal with retrieved but unjudged documents. The obtained results are reported in Table 6.18.

Table 6.18: Retrieval performance achieved by the human-generated and ChatGPT-generated queries.

Experiment	Bpref	P@10 (Condensed)
qfA	.117	.140
qfB	.093	.120
qfC	.059	.100
qfD	.116	.138
qfall	.090	.110
IEMT	.107	.131

Based on the table, IEMT leads to better retrieval performance compared to the accumulated queries of assessors B, C and all combined (i.e. *qfall*). However, it fails to improve retrieval performance against assessors A and D. It is reported by Koopman and Zuccon [2016] that assessors B and C, i.e. those whose queries underperformed compared to ChatGPT's, had a different behavior than assessor A and D. Specifically, assessors B and C created many small queries, while assessors A and D created fewer but longer ones (i.e. contained more keywords). These findings suggest that ChatGPT might be better for information extraction for this tasks compared to medical experts, under certain circumstances. Nonetheless, more experiments should be conducted to further investigate whether ChatGPT, or another LLM, can generate better queries for clinical trials retrieval than medical experts.

6.2.4 Limitations and Potentials of ChatGPT

In this section, we report the limitations encountered in our experiments. Additionally, we examine the potential positive and negative impacts of utilizing the proposed approach for automating the task of clinical trials retrieval.

6.2.4.1 Usage and Experimental Limitations of ChatGPT

This study uses ChatGPT (specifically the gpt-3.5-turbo model) by leveraging OpenAI's API. Our experiments were conducted over two weeks starting after March 16th 2023. It is important to note that ChatGPT is treated as a black box in this study, as details regarding its architecture and the specific training data used are undisclosed. Moreover, we remind that each query is processed in a separate system interaction via the API, i.e. through a new conversation chat. According to the documentation, ChatGPT has been trained using data available until September 2021. However, it remains uncertain whether the model has specifically been trained on the TREC 2021 topics used in this study, which were published in May 2021, or on similar topics such as those discussed in the previous work by Koopman and Zuccon [2016]. In addition, the model has probably not been exposed to or trained on the TREC 2022 topics, as they were published on June 2, 2022. The behavior of the ChatGPT model is non-deterministic, meaning that it can produce different responses for the same input. To mitigate this, we set the *temperature* parameter to zero in our experiments, intending to make the generated responses as deterministic as possible. However, as mentioned in the documentation, a slight amount of variability may still be present. Additionally, it is essential to note that all of our experiments were performed using a single user account. Therefore, whether the provider retains user-related information from previous system conversations and uses it to personalize the model's responses is uncertain. In order to examine the response variability of the model, we conducted a repeat of the IEMT experiment on April 24 2023, which was one month after the initial experiment. It is important to note that the same user account was utilized for both experiments. The retrieval performance achieved in the two experiments was found to be identical. This result suggests that the IEMT prompting approach exhibits robustness, at least when the same user account is employed. However, future work should further study this aspect to understand ChatGPT's response variability better.

6.2.4.2 Potentials of using ChatGPT for Clinical Trials Retrieval

In this research, ChatGPT is used as an intermediary tool to facilitate information extraction from unstructured clinical notes, a task that traditionally demands substantial human involvement. In this section, we highlight the benefits of employing ChatGPT, as described in our study, to automate or semi-automate this essential stage of the clinical enrollment workflow. The advanced language processing capabilities of ChatGPT enable efficient extraction of the desired information from clinical notes, minimizing the required human effort and mitigating the risk of human errors. Concurrently, our findings indicate that queries generated by ChatGPT result in improved retrieval performance compared to existing approaches in the literature, and, occasionally, those generated by human experts. The capabilities of ChatGPT indicate that it can replace multiple domain-specific language models that are fine-tuned for extracting specific pieces of information. Consequently, ChatGPT will simplify the operation and maintenance of an information extraction system designed to process unstructured clinical notes. Concerning the issue of hallucination generation, the proposed approach mitigates this concern by restricting the model's focus to information extraction rather than attempting to answer medical queries or make decisions about a patient's medical situation. In conclusion, ChatGPT has the potential to automate the information extraction process fully. However, it would be more suitable to incorporate human oversight and validation as an additional step to ensure the extracted information's accuracy, completeness, and reliability.

6.2.4.3 Risks of Using ChatGPT for Clinical Trials Retrieval

One of the main limitations of ChatGPT and other generative language models is their non-deterministic behavior. In this study, we acknowledged this limitation and took explicit actions to address it, although slight variations in model responses may still occur. In addition to non-determinism, other limitations should be considered in medical information extraction and retrieval, such as their lack of explainability and potential concerns related to data privacy. Given the nature of the information extraction task in this study, where the focus is on extracting simple information rather than making decisions about a patient's situation, obtaining explanations from the model is of lower significance. It is worth noting that in the proposed approach ChatGPT does not make decisions about a patient's situation and its outputs do not directly impact patient care. However, these limitations should be carefully considered when the model is expected to answer questions about a patient, such as determining whether the patient's condition is chronic or acute. The process of extracting information from clinical notes using ChatGPT raises valid concerns regarding data privacy and security, as it involves handling sensitive patient information that needs to be protected. To mitigate this risk, it would be essential to employ de-identification techniques on the EHRs prior to extracting information from them, for example by using models based on recurrent neural networks [Ahmed et al., 2020] or even other LLMs [Liu et al., 2023b]. These de-identification methods help anonymize patient data, reducing the likelihood of exposing sensitive information during the extraction process. Additionally, it is crucial to implement the proposed approach in a manner that complies with regulations that ensure the security and confidentiality of patient-related information, such as the Health Insurance Portability and Accountability Act (HIPAA)[12]. By adhering to these measures, healthcare providers can protect patient privacy while leveraging ChatGPT for information extraction purposes.

6.2.5 Conclusions and Directions for Future Research

The primary objective of the research presented in this section is to improve the initial stage of the clinical enrollment workflow by incorporating ChatGPT as an intermediate component within a retrieval pipeline. ChatGPT extracts information from unstructured clinical notes that detail a patient's medical situation, employing specially designed prompts. The generated responses from ChatGPT serve as queries to identify relevant clinical trials in which the patient may be eligible to participate. The proposed approach explicitly addresses several limitations commonly associated with LLMs, including the potential for generating hallucinations and response inconsistency. Our investigation assesses the model's capabilities in query generation, information extraction, and disambiguation of essential patient-related information found in unstructured clinical notes. Based on the empirical evaluation conducted, we provide a summary of the key conclusions of our study in terms of the research questions presented in this study.

Analyzing the results presented in Table 6.15, it becomes evident that within the utilized prompting strategies, the IEMT experiments exhibit noteworthy enhance-

[12] Health Insurance Portability and Accountability Act of 1996 (HIPAA).

ments in retrieval performance for both benchmark collections. These experiments involve assigning a domain-specific role to the model and providing it with a comprehensive task description. The observed improvements are statistically significant, highlighting the effectiveness of this approach. Furthermore, the IEMT and the IEMT+RM3 experiments yield better retrieval performance compared to the state-of-the-art approaches for query generation (refer to Table 6.17) as well as human-generated queries (refer to Table 6.18). These findings strongly support the conclusion that the patient-related information extracted using ChatGPT is sufficient to enhance retrieval performance in the studied search task and it also outperforms the SoA approaches and, occasionally, even human-generated queries. In consideration of the final research question pertaining to the potential advantages of incorporating ChatGPT into the clinical trial enrollment pipeline, we contend that the answer is positive, provided that the employed system effectively addresses the risks outlined in Section 6.2.4.3 and protects patients privacy.

Our future work will address several shortcomings identified in our empirical experiments. To begin with, in our experiments we noticed that when ChatGPT is instructed to expand or refine its response (i.e. generated query or extracted information), it tends to add terms with a broader semantic meaning, leading to topical drift and a performance decrease. We would like to investigate towards this direction aiming to instruct ChatGPT to generate terms with requested semantic meaning. In addition to that, we intend to perform a detailed analysis related to the quality of the suggested MeSH terms and the correctness of the resolved abbreviations. As mentioned in Section 6.2.1.6, the attempt to instruct ChatGPT to identify and remove negated content from a clinical note resulted in unexpected outcomes that completely changed the semantic meaning of a clinical note. Therefore, this model limitation requires further investigation in future studies, as it can potentially negatively impact the interpretation of clinical information. Similarly, more extensive experiments will be conducted to investigate the models response variability over time, across different user profiles and slightly modified prompts. As discussed in Section 6.2.1.5, ChatGPT has demonstrated the ability to disambiguate the meaning of different medical terms within a clinical note. Based on this finding, we intend to leverage ChatGPT, or another domain-specific language model, such as GatorTron, to automate further an additional stage of the clinical enrollment workflow, specifically the eligibility screening process.

6.3 Discussion

The central aim of our research into information extraction from EHRs is to investigate its potential to enhance the initial phase of clinical trial enrollment, i.e. searching for eligible clinical trials for a given patient.

Our study begins by employing established rule-based techniques alongside cutting-edge pre-trained language models to assess their strengths, weaknesses, and potentialities. Subsequently, we delve into the utility of LLMs, specifically ChatGPT, for information extraction within the same experimental framework. Our results endorse the hypothesis that ChatGPT is highly effective in extracting patient-related information. Notably, the model surpasses previous state-of-the-art methods and, in some instances, even outperforms queries generated by human experts. Such advancements hold significant promise for elevating the quality of healthcare services while simultaneously lowering the workload of healthcare practitioners.

Nevertheless, we must underscore the issues surrounding data privacy and patient safety, even though we have outlined potential solutions to these challenges in preceding sections. The practical implementation of a large language model extends beyond mere performance metrics; it also hinges on the model's sustainability and maintenance costs. A comprehensive evaluation of the feasibility of deploying such a model within a healthcare organization exceeds the boundaries of this dissertation. Nonetheless, methods like knowledge distillation [Hinton et al., 2015] and quantization [Jacob et al., 2018] may make the adaptation of such technologies feasible even for small medical organizations.

Chapter 7

DtMRF, Neural-DtMRF, and LLMs for Clinical Trials Retrieval

This chapter offers an exhaustive evaluation of experimental outcomes related to the utilization of the Decision-theoretic Multidimensional Relevance Framework and its Neural extension in the context of clinical trials retrieval. Additionally, it encompasses the empirical results derived from a methodology that employs LLMs to evaluate patients' eligibility for clinical trials. These methods present a contradiction: DtMRF is characterized by high interpretability, whereas LLMs lack such transparency in inference. The chapter concludes with a synthesis of these results, outlining their theoretical and practical applications in the field of clinical trials retrieval.

7.1 Introduction

This chapter introduces our approaches toward enhancing retrieval performance and interpretability in clinical trials retrieval. We employ three distinct methodologies, namely, DtMRF, Neural DtMRF, and a Large Language Model. The latter employs a specifically designed prompt to assess patients' eligibility for participation in clinical trials. As outlined in Chapter 5, both DtMRF and Neural-DtMRF necessitate identifying the following components.

A collection of Documents. In our research, we leverage the TREC 2021 collection to comprehensively examine the retrieval effectiveness and behavior of the DtMRF. Regarding the Neural-DtMRF approach, experiments are conducting utilizing the TREC 2021 and 2022 collections.

Estimating Relevance in Clinical Trials Retrieval. Considering the utility of a clinical trial for end-users who seek to allocate patients, we have identified the factors of relevance that we need to incorporate into our models. As Section 4.1 outlines, patients will be ineligible for participation in a clinical trial if their information aligns with the trial's exclusion criteria, regardless of whether their data matches its inclusion criteria and research objectives. Based on that, we design our models so that they estimate relevance accounting for this task requirements. DtMRF assess a document as relevant if the patient information has high relevance to a trials inclusion and main parts (e.g. title, summary), and as low as possible to the trials exclusion criteria.

Objectives and Importance Weights. Regarding the objectives to be associated with the relevance factors, these are beneficial if the factors are estimated with respect to any document part except its exclusion criteria. Any factor estimated based on a trial's exclusion criteria is a non-beneficial factor. Obtaining the importance weights differs between the DtMRF and Neural-DtMRF methods. We comment on that in the dedicated sections.

Evaluation functions to Estimate the Relevance Factors. Our relevance factors mainly rely on topical similarity. Therefore, we rely on relevance factors such as *topicality*, *coverage* and *coherence*, as we will elaborate in the dedicated sections.

Aggregation and Final Ranking. Our experiments evaluate the performance of all DtMRF instantiations in the TREC 2021. Based on that, we select the best-performing instantiation to be used in our experiments that leverage Neural-DtMRF and the TREC 2022 collection.

This chapter presents three experiments designed to address distinct research questions. Section 7.2 focuses on leveraging the DtMRF model, conducting experiments on the TREC 2021 dataset to understand the model's performance, behavior, and possible improvements. In Section 7.3, we extend our work by using Neural-DtMRF to predict the importance weights for each query. For this set of experiments, we utilize the highest-performing DtMRF instantiation in the domain of clinical trials retrieval. Section 7.4 employs an entirely different approach for comparative analysis. We use an LLM that assesses whether a patient meets the eligibility criteria for a clinical trial. The chapter concludes with a synthesis of the findings, discussing their theoretical and practical implications in the domain of clinical trials retrieval.

7.2 Leveraging DtMRF for Clinical Trials Retrieval

In this section we show how DtMRF incorporates the expert users' task-based decision-making behavior directly into its retrieval process. We propose two retrieval approaches to achieve that, focusing on answering the following research questions:

(RQ1) To what extent does the inclusion of negative relevance factors enhance retrieval effectiveness in clinical trials retrieval?

(RQ2) Which of the DtMRF instantiations considered in the study is better suited for clinical trials retrieval?

(RQ3) How does the retrieval effectiveness of DtMRF compare to alternative approaches proposed in the existing literature?

The first approach is a single run retrieval approach based on one relevance factor, i.e topicality, calculated with respect to three document representations that convey different relevance importance. The second approach is a re-ranking retrieval approach that additionally incorporates two more relevance factors, namely coverage and coherence.

An overview of the proposed single run retrieval approach is presented in Figure 7.1.

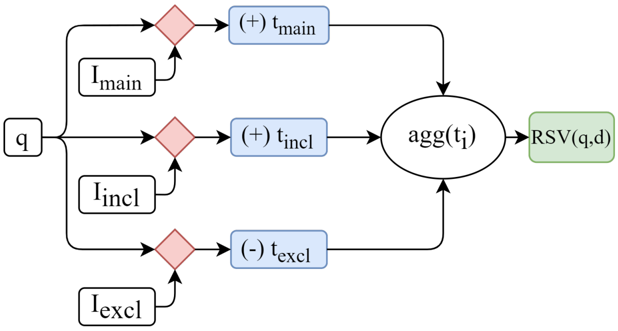

Figure 7.1: Overview of the proposed single run retrieval approach.

In detail, given a document d, we extract its inclusion and exclusion criteria from its eligibility section, and we merge the other main sections, i.e its title, detailed description, and summary, into a single text. Then, we separately index these three textual parts by creating three different indices, i.e I_{main}, I_{incl}, and I_{excl}. By doing that, we have created three distinct representations for each document which, in this task, come with a different importance to its overall relevance assessment.

Following that, given a query q, we estimate the topical relevance (t) of each document part (main, inclusion, exclusion) to this query by a standard retrieval model (i.e its employed evaluation function) and obtain the topical relevance scores (i.e performance scores) associated with the t_{main}, t_{incl} and t_{excl} criteria, for every document in the collection. However, as we have outlined earlier, high topical relevance between a patient's information and a document's exclusion part is not desirable for this task. This requirement is incorporated into this retrieval approach on the last step of our methodology, i.e the aggregation.

The final step is the aggregation, $agg(t_i)$, over the three individual relevance criteria, t_i, by explicitly considering their independent contribution (i.e their objective), either positive or negative, to the document's overall relevance (utility). In particular, the t_{main} and the t_{incl} are considered beneficial criteria, while the t_{excl} is considered a non-beneficial criterion. The aggregation step is conducted by using one of the proposed instantiations of DtMRF and it estimates a Retrieval Status Value (RSV) for each document (i.e, a global performance score). It is important to clarify here that the selection of the beneficial and non-beneficial criteria follows the characteristics of this particular task and the aggregation step is aligned with the task-based decision behavior of a professional user. That means that each DtMRF instantiation we have considered in this work, will penalize a clinical trial for which a patient's information has high topical relevance to the trial's exclusion criteria.

The problem formulation as described above can be summarized in the decision matrix presented in Table 7.1 where, the x_{ij} values represent the topical relevance scores obtained by a state-of-the-art retrieval model that has been used as an evaluation function. Regarding the importance weights to be associated with the criteria t_i, we obtained these weights by means of a parameter sweep with step size 0.1, and explored how these affect the obtained ranking.

Table 7.1: The $M_{m \times 3}$ decision matrix employed in the single run retrieval approach.

	t_{main}	t_{incl}	t_{excl}
d_1	x_{11}	x_{12}	x_{13}
d_2	x_{21}	x_{22}	x_{23}
...	...	...	...
d_m	x_{m1}	x_{m2}	x_{m3}

Given the $M_{m \times 3}$ decision matrix, the $\text{DtMRF}_{\text{TOPSIS}}$, $\text{DtMRF}_{\text{VIKOR}}$, and $\text{DtMRF}_{\text{COPRAS}}$ can be directly employed to estimate a document's utility and rank the documents, by following the steps described in Section 2.2.2. However, the $\text{DtMRF}_{\text{WSM}}$ can not be directly applied due to its requirement that all the criteria should be beneficial criteria (which is not the case in this search task). Therefore, for the $\text{DtMRF}_{\text{WSM}}$ we have formulated the problem as presented in Equation 7.1, to capture the contribution of each criterion to the document's utility and particularly the negative effect of a patient's topical relevance to a trial's exclusion criteria.

$$RSV(q, d) = w_{main} t_{main} + w_{incl} t_{incl} - w_{excl} t_{excl} \tag{7.1}$$

The formulation presented in Equation 7.1 has also been followed by a few approaches in TREC clinical trial 2021 track[10].

Finally, it is important to outline two advantages of the proposed retrieval approach. Specifically, this retrieval approach allows the formulation of distinct queries on distinct document sections, while a standard retrieval approach considers a single query to be used for estimating the topical relevance of a document. This advantage may be of great importance for this task, under the assumption that particular information present in a patient's health record can be significant to determine her/his eligibility, while other information can be significant to identify possible clinical trials. For instance, a patient's current clinical condition or symptom can be used as query against the I_{main} index, while other patient information such as, habits, family history as queries against the I_{incl} and I_{excl} indices.

The second advantage is related to the fact that this single run retrieval approach can be split in two parts; one dedicated to the estimation of the three criteria (t_{main}, t_{incl} and t_{excl}) and the second is dedicated to the employment of the aggregation schema that incorporates the desired decision behavior in the retrieval process. Therefore, it enables the combination of DtMRF with various models that estimate topical relevance, e.g. neural models, that are capable of capturing the semantic similarity between a patient's information and the desired parts of a document.

Similarly to the proposed single run retrieval approach, the re-ranking approach we propose also exploits the inclusion and exclusion document parts. In this approach, we investigated whether more relevance factors (i.e coverage (cov) and coherence (coh), introduced by Li et al. [2017a]), can further improve the retrieval effectiveness. These relevance factors are used to estimate a document's scope, that has been defined by Xu and Chen [2006b] as the "extent to which the topic or content covered in a retrieved document is appropriate to the user's need." Therefore, estimating coverage, coherence, along with topicality, can provide a more holistic estimation of a document's utility. However, the evaluation functions used to estimate these criteria (refer to Section 7.2.1) are computationally expensive and this is the reason why this approach is used for document re-ranking.

In this approach, we associate three relevance factors i.e coherence, coverage, and topicality with a document's inclusion and exclusion parts and then aggregate the obtained performance scores to estimate a document's overall relevance score. Also, we assume that every performance score obtained from a document's inclusion part is beneficial to its overall relevance, while every performance score obtained from a documents exclusion part is non-beneficial. An overview of the proposed re-ranking approach is presented in Figure 7.2.

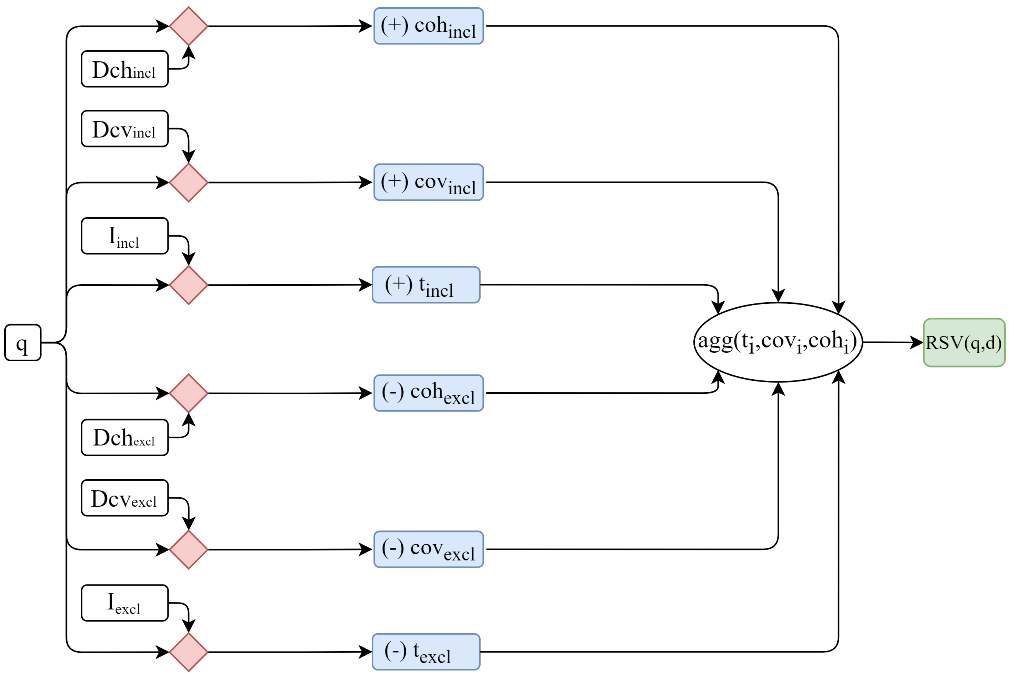

Figure 7.2: Overview of the proposed re-ranking retrieval approach.

During indexing, we create the I_{incl}, and I_{excl} indices, as described in the previous section, which are used to estimate the respective topical relevance scores. In addition, we estimate and store the static embedding representation of each word (excluding stop-words) that is present the inclusion (Dch_{incl}) and exclusion (Dch_{excl}) document parts. These representations are later used to estimate coherence. Moreover, we split the inclusion and exclusion document parts into small chunks of consecutive words (excluding stop-words) using a fixed window size to obtain the Dcv_{incl} and Dcv_{excl} document representations. These sets of consecutive words are used during retrieval time for the estimation of coverage. Further details related to the creation of these representations (Dcv and Dch) are provided in Section 7.2.1.

During retrieval, this approach involves an initial retrieval step from which we obtain the top-1000 retrieved documents for a query q (this step is omitted from Figure 7.2). Then, by exploiting the aforementioned document representations and the evaluation functions that are described in detail in Section 7.2.1, we estimate the performance scores of each document over the six considered criteria. Similarly to the single run retrieval approach, the last step involves their aggregation, $agg(t_i, cov_i, coh_i)$, by a DtMRF instantiation.

This problem formulation can be summarized in the decision matrix presented in

Table 7.2 where, the x_{ij} values represent the performance scores obtained by three employed evaluation functions (refer to Section 7.2.1). Regarding the importance weights, similarly to the single run retrieval, we obtained them by means of a parameter sweep with step size 0.1 using a portion of the data. Again here,

Table 7.2: The $M_{m\times 6}$ decision matrix employed in the re-ranking retrieval approach.

	t_{inc}	t_{exc}	cov_{incl}	cov_{excl}	coh_{incl}	coh_{excl}
d_1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
d_2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}
...	...	...	...	...	...	...
d_m	x_{m1}	x_{m2}	x_{m3}	x_{m4}	x_{m5}	x_{m6}

starting from the $M_{m\times 6}$ decision matrix, the DtMRF$_{\text{TOPSIS}}$, DtMRF$_{\text{VIKOR}}$ and DtMRF$_{\text{COPRAS}}$ can be directly employed, but DtMRF$_{\text{WSM}}$ can not. Therefore, following the same intuition as before, for DtMRF$_{\text{WSM}}$, we have formulated the problem as presented in Equation 7.2.

$$\begin{aligned} RSV(q,d) = w_{tincl}t_{incl} - w_{texcl}t_{excl} + w_{covincl}cov_{incl} - w_{covexcl}cov_{excl} \\ + w_{cohincl}coh_{incl} - w_{cohexcl}coh_{excl} \end{aligned} \tag{7.2}$$

Finally, this retrieval approach shares the same advantages as the single run retrieval approach.

7.2.1 Experimental Design and Results

This section presents the experimental setup employed to answer the research questions outlined in the introduction. It presents the employed evaluation metrics, the dataset and its indexing, and the conducted experiments. In addition, it describes the extraction of a trial's inclusion and exclusion criteria. Also, it presents the estimation of the performance scores associated with the considered relevance criteria, i.e the evaluation functions (Section 7.2.1.1). Subsequently, the results of the experiments are discussed, including performance comparisons and notable observations.

Evaluation Metrics. We employ several precision-oriented effectiveness measures, i.e Bpref, Mean Reciprocal Rank (MRR), Precision at several cut-offs, and NDCG, as the relevance assessment is on a graded relevance scale. Moreover, for the reported effectiveness measures that are based on binary relevance assessment, only eligible trials (label 2) are assumed as relevant.

Data Collection and Indexing. The empirical evaluation is performed using the TREC 2021 collection. As outlined in Section 7.2, both the proposed retrieval approaches depend on the extraction of a trial's inclusion and exclusion criteria along with other sections. Specifically, we would like to remind here that the considered document sections in the proposed implementation are the title, detailed description, and summary, that were concatenated into one text; and the eligibility section, from which we extracted the inclusion and exclusion criteria. Here, we provide further details regarding their extraction and present the related statistics obtained from the employed benchmark collection.

To begin with, for 990 of the available clinical trials, it is not feasible to extract any of their sections. That is because a part of these trials (841) have not been approved by the U.S. Food and Drug Administration (FDA), and therefore the related information was removed. The remaining missing trials contain the text "Please contact site for information." and no other information is available. For our implementation, we remove these documents from the considered collection. In addition, our analysis revealed that a portion of the documents has several empty sections. Therefore, we have further investigated towards that direction to understand the degree to which missing values can influence the obtained experimental results.

Our analysis shows that the detailed description section is missing for 33.0% of the total documents, but the title and summary sections are present for all the clinical trials. As a result, it is feasible to create the concatenated text using these main document sections and obtain the I_{main} document representation for every document in the collection. Regarding the eligibility section, it is present in all of the considered documents. So, we automatically extract the inclusion and exclusion criteria from the eligibility section by developing a set of linguistic rules that leverage their semi-structured format. These linguistic rules mainly exploit the presence of a header, i.e "inclusion criteria:" and "exclusion criteria:", which delimits the two section parts. However, their extraction is not feasible for all documents, as for some of them, these criteria are mentioned in an unstructured format.

The employed extraction method is unable to obtain both the inclusion and exclusion criteria for 2.5% of the total documents, due to their unstructured format. For such documents, the whole eligibility section has been used to obtain the required document representations (I_{incl}, I_{excl}, Dch_{incl}, Dch_{excl}, Dcv_{incl}, and Dcv_{excl}). We remind that, I_{incl} and I_{excl} refer to the indexed documents' textual parts, as described in Section 7.2. Dch_{incl}, Dch_{excl} refer to static word-embedding representations, and Dcv_{incl}, Dcv_{excl} refer to extracted documents' textual chunks (refer to Section 7.2).

Also, 1.7% of the documents with a semi-structured eligibility section contain only inclusion criteria. In these cases, it is not possible to obtain the required I_{excl}, Dch_{excl}, and Dcv_{excl} document representations, and a place holding text is used instead. Consequently, the developed extraction approach has been able to extract both of the necessary document parts for the vast majority (97.5%) of the total documents provided in this benchmark collection. While only a small portion of these documents do not contain exclusion criteria.

Experiments (Baselines). We experiment with two well-known retrieval models, BM25 and ln_expB2, and use them to estimate the retrieval effectiveness on various document representations. The purpose of these experiments is two-folded. First, the conducted experiments contribute to identifying the retrieval model that yields better effectiveness in this collection. Second, comparing the retrieval effectiveness obtained across different document representations allowed us to understand their independent contribution.

The findings of our experimentation are presented in Table 7.3. Here, each row shows the retrieval performance obtained by the two employed retrieval models across different document representations, following the BM25/ln_expB2 format. Specifically, the I_{raw} representation is obtained by indexing the the whole text of the documents contained in the collection, while the creation of the I_{main}, I_{incl} and I_{excl} has already been discussed in detail. Also, to create the $I_{main,inc}$ representation, we combine and index the main and inclusion document parts, i.e we have eliminated a trial's exclusion criteria. That document representation has also been employed by a significant amount of research works in TREC 2021[10]. Lastly, as the considered documents are structured, we also evaluated the retrieval performance of the BM25 field (BM25f) model, that leverages the I_{main}, I_{incl}, I_{excl} indices. The obtained results are presented in the last row of Table 7.3, and it can be seen that the model does not outperform the other approaches.

A general observation is that the ln_expB2 model performs equally or better than the BM25 model, for the majority of the employed evaluation measures. Therefore, this model is selected to estimate the topical relevance scores associated with the t_{main}, t_{incl} and t_{excl} criteria.

Table 7.3: Effectiveness across different document representations for BM25 and ln_expB2 models. Models' effectiveness follows the format: BM25/ln_expB2.

	Bpref	Rprec	MRR	P@10	NDCG@10
I_{raw}	.18/.17	.24/.24	.47/.47	.27/.28	.47/.50
I_{main}	.17/.19	.23/.24	.47/.52	.27/.30	.44/.48
I_{incl}	.18/.17	.22/.22	.48/.47	.29/.28	.47/.46
I_{excl}	.08/.08	.11/.10	.31/.32	.15/.16	.34/.35
$I_{main,inc}$	**.20/.22**	**.25/.27**	**.49/.54**	**.28/.30**	**.51/.52**
$I_{main,inc,excl}$ $(BM25f)$	.19	.17	.43	.26	.46

Another significant observation that can be drawn from Table 7.3, is that even for the I_{excl} index, the MRR value is relatively high; a result that is counter-intuitive to our initial hypotheses that for an eligible trial, relevance to the exclusion criteria should be close to zero. To clarify, the employed model retrieves eligible trials while indexing only a trial's exclusion criteria. First, it is plausible that this is due to the applied extraction method and specifically due to the documents for which we fail to extract their inclusion and exclusion criteria. We remind that for these document we have used a trial's eligibility section to create both I_{incl} and I_{excl} indices. However, that has only influenced a small portion of the total documents (2.5%). Another, possible co-existing, reason may be related to the employed retrieval model and its estimation of topical relevance with respect to the query and the provided index. Nonetheless, the relatively low values of Bpref, Rprec and P@10 measures suggest that our initial intuition still holds, while a more accurate topical relevance estimation would minimize retrieval uncertainty, and therefore this issue will be overcome, at least partially.

All in all, our performance analysis across the two well-known retrieval models and document representations showed that the best document representation is the $I_{main,inc}$ and the best performing model is the ln_expB2. Hereinafter, we refer to this retrieval approach as $\mathrm{DFR_{cin}}$; this is one of the two considered baselines in our research. Due to the fact that the $\mathrm{DFR_{cin}}$ is influenced by the information extraction approach we used to extract a trial's inclusion criteria, we introduce another retrieval baseline, namely $\mathrm{DFR_{bsl}}$, that uses the I_{raw} index and the ln_expB2 retrieval model. Both of these retrieval approaches completely disregard the topical relevance of a patient's information to a trial's exclusion criteria. Therefore, they are sufficient baselines to investigate whether the DtMRF, that considers the negative effect of a trial's exclusion criteria, can further improve the retrieval effectiveness.

Experiments (Ours). We compare the four DtMRF instantiations and the previously mentioned baseline models. Additionally, we explore the impact of the importance weights on the retrieval performance of DtMRF variations.

The following sections provide details regarding the estimation of the estimations of the score associated with the considered relevance factors and our experiments.

7.2.1.1 Estimation of the Considered Relevance Factors

This section provides further details related to the evaluation functions used for the estimation of the individual performance scores for each of the considered relevance criteria, i.e topicality, coverage and coherence.

Criterion: Topicality (t) Topicality is generally measured by an IR model. We have investigated the retrieval effectiveness of two well-known IR models, namely the BM25 model [Robertson et al., 1994] and the ln_expB2 Divergence from Randomness (DFR) model [Amati and van Rijsbergen, 2002]. Ultimately, we adopt the ln_expB2 model as it leads to better retrieval performance (refer to Section 7.3). To obtain the required indices, I_{main}, I_{incl}, and I_{excl}, we index the respective document sections and extracted parts using PyTerrier with its default indexing parameters, i.e porter-stemming and stopword removal [Macdonald et al., 2021]. Finally, to obtain the topical relevance scores (i.e the performance scores) associated with the t_{main}, t_{incl} and t_{excl} relevance criteria, we use the original verbose queries and the default parameters of the employed models.

Criterion: Coverage (cov) The estimation of the performance scores associated with the coverage criterion is based on the method proposed by Li et al. [2017a], which requires several steps.

Firstly, the calculation process requires splitting a given document into small chunks of consecutive words to create a set of document windows, using a pre-defined parameter, L, that defines the window size. In the original implementation, the L parameter has been set equal to 16, as the authors have been working with normal sized documents (e.g. web pages) [Li et al., 2017a]. However, in our experiments the considered texts, i.e the inclusion and exclusion parts of a clinical trial, are relatively smaller and mostly consist of a few sentences. Specifically, our analysis has shown that the document part that mentions the inclusion criteria consist on average of 86 tokens (including stop-words), while the part dedicated to a trial's exclusion criteria contains on average 106 tokens (including stop-words). Therefore, we have lowered the value of the L parameter to 8. This procedure is performed offline and the Dcv_{incl} and Dcv_{excl} document representations are stored, so that the estimation of their respective performance scores is performed faster during retrieval.

Then, to obtain the individual performance scores for each document, given a query, we employ the formula presented in Equation 7.3, as introduced by Li et al. [2017a].

$$CoverRatio = \frac{uwL(q)}{\text{windows}} \tag{7.3}$$

In Equation 7.3, $uwL(q)$ is the total number of document windows that contain query terms. In the original implementation, this value is increased if at least one query term is present in a document window; however, in our experiments, the query length is larger as the query size varies from 5 to 10 sentences. Therefore, we have increased the required number of query terms to be present in a document's window to 2. Moreover, to avoid term mismatching, both the words present in every document window and the queries are lowercased. Finally, the total number of document windows that contain at least two query terms, is divided by the total number of windows in a document, denoted as *windows*. A larger CoverRatio value means that the document has a narrower scope and focuses on the query-related content. Here, it means that specific patient-related content, for instance a medical condition, has been found in a trial's inclusion or exclusion criteria.

Criterion: Coherence (*coh*) The estimation of the performance scores associated with the coherence criterion is also based on a method proposed by Li et al. [2017a].

In particular, initially, given a document (here, either the inclusion or exclusion parts), we estimate the static embedding representation of each word (excluding stop-words) using the Google word2Vec toolkit. The obtained embedding representation Dch_{incl} and Dch_{excl} are also stored to aid the retrieval efficiency. Then, during retrieval, the performance scores are calculated using Equation 7.4, where v_w is the stored word embedding vector for a word w present in the document, and v_q is the uniform-weighted sum of word vectors for all query terms (excluding stop-words). Moreover, longer documents are penalized, by dividing the similarity score with the length of the considered document ($docLength$).

$$CoherenceQD = \frac{\sum_{w \in d} \cos(v_w, v_q)}{\text{docLength}} \tag{7.4}$$

Ultimately, a larger coherence value means a better semantic similarity between patient-related content with a trial's inclusion or exclusion criteria.

7.2.1.2 Effectiveness of DtMRF: Single Run Retrieval Approach

This section concerns the evaluation of the single run retrieval approach. We particularly investigate whether the proposed DtMRF can improve retrieval effectiveness, and second, which of the four DtMRF instantiations is more suitable for this search task.

To compare the four instantiations, we conduct a grid parameter search across all of the provided queries, to identify those criteria weights that maximize the MRR measure for each method. Therefore, the retrieval effectiveness presented in the last four rows of Table 7.4 demonstrates the upper retrieval effectiveness bound obtained by each DtMRF instantiation. The corresponding importance weights associated with the three considered criteria are reported in brackets following the [w_{main}, w_{incl}, w_{excl}] format. In addition, Table 7.4 presents the evaluation measures obtained by the two considered baselines DFR_{bsl} and DFR_{cin}. The statistical significance is tested against both baselines according to a paired t-test with Bonferroni multiple testing correction, at significance levels 0.05(°), while (-) means non-significant results. The best result per measure is in boldface.

Table 7.4: Retrieval effectiveness of the single run retrieval approach.

	Bpref	Rprec	MRR	P@1	P@5	P@10	NDCG@10	R@50
DFR_{bsl}	.168	.238	.467	.276	.309	.279	.496	.179
DFR_{cin}	**.218**	**.267**	.538	.370	.326	.298	**.519**	.186
$DtMRF_{WSM}$ [.7,.1,.2]	.196°	.242⁻	.544°	.361⁻	.330⁻	.291⁻	.498⁻	.186°
$DtMRF_{COPRAS}$ [.2,.4,.4]	.147°	.191°	.558°	.432°	.308⁻	.276⁻	.452°	.171⁻
$DtMRF_{VIKOR}$ [.5,.3,.2]	.157°	.213°	.422⁻	.268⁻	.228°	.210°	.290°	.149⁻
$DtMRF_{TOPSIS}$ [.5,.1,.4]	.198°	.252°	**.583°**	**.443°**	**.334⁻**	**.302⁻**	.510⁻	**.210°**

The retrieval results presented in Table 7.4 show that incorporating the negative influence imposed by the similarity of a patient to a trial's exclusion criteria in the retrieval process has the potential to improve retrieval effectiveness. Indeed, three out of the four instantiations outperform the DFR_{bsl} baseline for the precision-oriented evaluation measures, MRR and Precision; these performance increases are in their majority statistically significant. Nonetheless, note that $DtMRF_{TOPSIS}$ also achieves a statistical significant increase in Rec@50, compared to DFR_{bsl}. Regarding the best performing instantiation, $DtMRF_{TOPSIS}$ yields statistically significant improvements over the DFR_{bsl}, while it also improves the performance over the DFR_{cin}, but these improvements are not statistically significant. In fact, concerning the P@1 and the MRR measures, around 65% of the queries reach optimal performance both by $DtMRF_{TOPSIS}$ and DFR_{cin}; regarding the rest, $DtMRF_{TOPSIS}$ improves almost half of them.

Therefore, although these improvements are not statistically significant, employing the $DtMRF_{TOPSIS}$ instantiation still comes with some merits over DFR_{cin}. Specifically, we would like to remind here that the DtMRF retrieval approach allows three distinct query representations to be used in the retrieval process (refer to Figure 7.1). This is an advantage over the DFR_{cin} approach that might further improve the obtained retrieval results, and lead to statistically significance increases. However, we have not exploited this in the current research work, as we use the same query representation to obtain the topical relevance scores. In addition, by employing $DtMRF_{TOPSIS}$, the obtained document ranking is fully interpretable, as the first document would be the one that follows the characteristics of the considered task (high topical relevance to a document's main and inclusion parts and low topical relevance to the exclusion part). In contrast, DFR_{cin} retrieves a document for which the topical relevance to the trial's exclusion criteria is uncertain.

Regarding our second aim, the findings suggest that the $\text{DtMRF}_{\text{TOPSIS}}$ and $\text{DtMRF}_{\text{WSM}}$ instantiations are more suitable for this task as they yield better results compared to $\text{DtMRF}_{\text{COPRAS}}$ and $\text{DtMRF}_{\text{VIKOR}}$. Moreover, by observing the corresponding Bpref measures one can draw further insights. Specifically, both $\text{DtMRF}_{\text{TOPSIS}}$ and $\text{DtMRF}_{\text{WSM}}$ achieve higher Bpref than $\text{DtMRF}_{\text{COPRAS}}$, $\text{DtMRF}_{\text{VIKOR}}$, and DFR_{bsl}, while, at the same time, they yield higher performance for the considered precision-oriented measures. In contrast, for the $\text{DtMRF}_{\text{COPRAS}}$ and $\text{DtMRF}_{\text{VIKOR}}$ instantiations it can be seen that the corresponding Bpref values are lower than those obtained by DFR_{bsl}; this observation suggests that the improvements observed in the precision-oriented measures are due the employed condensed list evaluation, as these methods tend to rank irrelevant documents in higher positions than relevant documents. Concluding, our analysis shows that $\text{DtMRF}_{\text{TOPSIS}}$ is the most suitable instantiation for document ranking in this retrieval approach.

To further support the above claim, we investigated the sensitivity of $\text{DtMRF}_{\text{TOPSIS}}$ to the weights associated with the criteria by measuring the retrieval performance. This analysis is presented in Figure 7.3 and demonstrates how the MRR value is affected by changing the weights associated with the criteria. Firstly, one

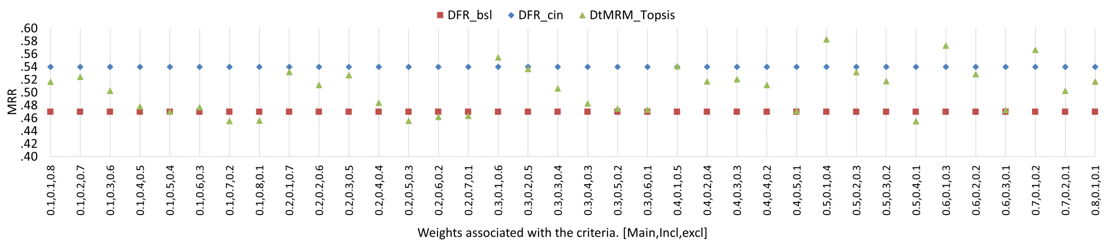

Figure 7.3: Sensitivity analysis of the MRR measure achieved by $\text{DtMRF}_{\text{TOPSIS}}$ for all combinations of the criteria weights. The weights on the x axis are following the format $[w_{main}, w_{inc}, w_{excl}]$.

may observe that DtMRF improves the retrieval performance in this task, when appropriate weights are associated with the three relevance criteria. Specifically, regarding the weights, as it can been seen in the figure, incorporating negative signals into relevance estimation leads to better retrieval effectiveness. In particular, when $w_{excl} \geq w_{incl}$, MRR is always improved compared to the DFR_{bsl}. Secondly, estimating relevance by relying solely to a patient's topical relevance to a trial's inclusion and exclusion criteria is not sufficient in this retrieval approach. Indeed, $\text{DtMRF}_{\text{TOPSIS}}$ reaches its highest effectiveness when $w_{main} \geq .5$ and $w_{excl} > w_{incl}$.

To conclude, the optimal and near-optimal set of weights obtained from these experiments are inline with our initial intuition regarding the contribution of a document's section to its overall relevance, under the studied search task. Specifically, one should primarily consider the similarity to a document's main parts; and then, weight more the negative influence imposed by the similarity to a trial's exclusion criteria, rather than the similarity to a trial's inclusion criteria.

Ranking Analysis. It can be seen in Table 7.4 that in terms of retrieval effectiveness $\text{DtMRF}_{\text{TOPSIS}}$ is a competitive approach. Moreover, the fact that $\text{DtMRF}_{\text{TOPSIS}}$ outperforms the baselines for the P@1 measure, means that it is more likely for the end-user to receive an eligible document in the first position, which is the ultimate retrieval goal in this task. Yet, we have further investigated the underlying reasons for its performance by analyzing the retrieved documents. Specifically, we wanted to investigate if $\text{DtMRF}_{\text{TOPSIS}}$ retrieves the same documents retrieved by the other retrieval approaches, especially in the top-ranked positions. To achieve that we measured the Kendall rank correlation and the intersection of the retrieved documents, across all queries, against the $\text{DtMRF}_{\text{TOPSIS}}$.

Table 7.5: Ranking correlation and retrieved document intersection, across all queries, compared to $\text{DtMRF}_{\text{TOPSIS}}$. The reported values concern the condensed rankings obtained from the experiments presented in Table 7.4.

	DFR_{bsl}	DFR_{cin}	$\text{DtMRF}_{\text{WSM}}$	$\text{DtMRF}_{\text{COPRAS}}$	$\text{DtMRF}_{\text{VIKOR}}$
Inter@1	19	24	33	21	33
Inter@10	288	381	434	301	428
K_cor@1	.20	.35	.34	.31	.43
K_cor@10	.11	.13	.10	.10	.17

The first two rows in Table 7.5 present the number of shared documents in the first (Inter@1) and in the top-10 (Inter@10) positions, while the last two present the Kendall rank correlation obtained at a specific document cut-off. All the reported correlations are statistical significant. An Inter@1 value equal to 75 means that all documents retrieved by the 75 queries were the same across the compared approaches, while a value of zero means an empty intersection set, i.e no common documents were retrieved. As it can be seen, despite the fact that the

DtMRF instantiations used to guide the aggregation are following the same problem formulation and intuition, the obtained document rankings are quite different. This has also been demonstrated in the example presented in Section 5.2.2. Similar behavior is observed for the top-10 retrieved documents, which has been supported by Kendall's correlations. Therefore, even if the absolute effectiveness values are relatively similar across the different decision-theoretic methods, the above analysis shows that the underlying aggregation mechanism behind them leads to different rankings.

Moreover, in Section 4.3.1 we raised some concerns regarding the issue of retrieving unjudged documents. In order to investigate the degree to which unjudged retrieved documents would have affected the performance evaluation in our experiments, we calculated the number of unjudged documents retrieved in the top-10 positions. Across 75 queries, DFR_{bsl} has a median value of unjudged documents retrieved equal to 0 and an average of 1, while $\text{DtMRF}_{\text{COPRAS}}$ has a median of 4 of 3.8. This finding further shows that $\text{DtMRF}_{\text{COPRAS}}$ should not be used for single run retrieval in this task. Regarding the other DtMRF instantiations and the DFR_{cin} baseline, these have a median value of 1 and average values that are spanning from 1.1 to 1.5. That means that all of the proposed retrieval approaches, except $\text{DtMRF}_{\text{COPRAS}}$, retrieve more or less the same number of unjudged documents, and therefore the reported effectiveness is not affected by the employed condensed list evaluation.

7.2.1.3 Effectiveness of DtMRF: Re-ranking Retrieval Approach

This section presents the results obtained from the experiments conducted as part of the re-ranking retrieval approach along with their corresponding result analysis, which is dedicated to the investigation of the obtained weights, and the number of documents to be re-ranked.

As it has been demonstrated in Table 7.3 and Table 7.4, the DFR_{cin} is the best performing standard retrieval approach. Therefore, we employed this retrieval approach to retrieve 1000 documents per query, which we have later re-ranked using the four DtMRF instantiations. To fairly compare the effectiveness achieved across the four DtMRF instantiations, we conducted an exhaustive search to identify, for each instantiation, its optimal set of weights and its optimal number of re-ranked documents. As optimal weights, we have chosen those that maximized the MRR measure. The obtained results are presented in Table 7.6, where its values represent the upper effectiveness bound that can be achieved by each instantiation. For each one, we report its corresponding importance weights associated with the considered criteria and the total number of re-ranked documents (top-n), following the $[w_{tincl}, w_{texcl}, w_{covincl}, w_{covexcl}, w_{cohincl}, w_{cohexcl}]$@top-n format. Similarly to the single run retrieval approach, the $\text{DtMRF}_{\text{WSM}}$, $\text{DtMRF}_{\text{COPRAS}}$ and $\text{DtMRF}_{\text{TOPSIS}}$ instantiations outperform both of the baselines, while these increases are also statistically significant compared to the DFR_{bsl}, as shown in Table 7.6. However, $\text{DtMRF}_{\text{VIKOR}}$ is not performing as well, while $\text{DtMRF}_{\text{TOPSIS}}$ yields the greater improvements.

Table 7.6: Retrieval effectiveness when the top-n retrieved documents by the DFR_{cin} method are re-ranked. Each row presents the obtained effectiveness using all of the considered criteria and instantiations.

	Bpref	Rprec	MRR	P@1	P@5	P@10	NDCG@10	R@50
DFR_{bsl}	.168	.238	.467	.276	.309	.279	.496	.179
DFR_{cin}	.218	.267	.538	.370	.326	.298	**.519**	.186
$\text{DtMRF}_{\text{WSM}}$6C [.1,.3,.2,.1,.1,.2]@10	.212°	.258°	.556°	.376¯	.346¯	.286¯	.496¯	.211°
$\text{DtMRF}_{\text{COPRAS}}$6C [.1,.1,.3,.1,.2,.2]@50	.218°	.266°	.578°	.397¯	.362¯	.328¯	.512¯	.223°
$\text{DtMRF}_{\text{VIKOR}}$6C [.1,.1,.1,.1,.1,.5]@10	.216°	.264°	.502¯	.338¯	.276¯	.312¯	.501¯	.221°
$\text{DtMRF}_{\text{TOPSIS}}$6C [.1,.1,.2,.1,.2,.3]@75	**.224°**	**.269°**	**.602°**	**.443°**	**.387°**	**.331¯**	.514¯	**.224°**

By analyzing the three well-performing instantiations in terms of the obtained criteria weights, one can observe that for $\text{DtMRF}_{\text{COPRAS}}$ and $\text{DtMRF}_{\text{TOPSIS}}$ the weights associated with the coverage and coherence criteria are higher than those associated with topicality. In contrast, for $\text{DtMRF}_{\text{WSM}}$ the weights have been allocated almost evenly among the considered criteria, with a slight preference on topicality. Plausibly, this is the reason why $\text{DtMRF}_{\text{WSM}}$ achieves performance similar to the initial retrieval approach, i.e the DFR_{cin}. Moreover, regarding the re-ranking depth, it can be concluded that $\text{DtMRF}_{\text{COPRAS}}$ and $\text{DtMRF}_{\text{TOPSIS}}$ are more robust compared to the $\text{DtMRF}_{\text{WSM}}$, as their optimal effectiveness was achieved when a larger amount of documents are re-ranked. To conclude, we observe that also in this retrieval approach, $\text{DtMRF}_{\text{TOPSIS}}$ yields statistically significant improvements, while at the same time, it is capable of improving MRR, P@1, and

P@5 more than any instantiation and the considered baselines.

However, the fact that the w_{tincl} and w_{texcl} are equal to .1 suggests that topicality may not be a suitable criterion for document re-ranking, when DFR$_{\text{cin}}$ was used for initial ranking. Therefore, by employing DtMRF$_{\text{TOPSIS}}$, we have conducted an ablation study, aiming to identify the criteria, their associated weights, and the re-ranking depth, that yield the best performance in this task. The obtained results are presented in Table 7.7, where the weights are following the [w_{incl}, w_{excl}, w_{incl}, w_{excl}, ...]@top-n format. From these experiments, as it can be seen in the eighth row of Table 7.7, considering only the coverage and coherence criteria can further improve the retrieval effectiveness, even in the case when all the retrieved documents were re-ranked. These results correspond to the optimal retrieval performance that can be achieved by the DtMRF$_{\text{TOPSIS}}$ instantiation, in the employed retrieval setting. In fact, as we re-rank over the DFR$_{\text{cin}}$ baseline, the top-retrieved documents have high topical relevance to a trial's main and inclusion criteria, and either high or low topical relevance with respect to a trial's exclusion criteria. The later relevance is estimated during re-ranking through the considered criteria, and probably this is the reason why neglecting topical relevance, in this setting, improves the retrieval performance.

Moreover, the findings suggest that the coverage criterion is the most important for the determination of a trial's utility, while topicality and coherence is not as good. In fact, by observing the third row one can notice that when the utility estimation relies solely on topicality, the best effectiveness is achieved when only top-10 documents are re-ranked, and only when the estimation is fully depended on the negative influence imposed by the similarity to a trial's exclusion criteria. Similarly, when the DtMRF instantiations relies solely on the coherence criterion the best performance is achieved when only the top-10 documents are re-ranked, and when $w_{cohincl} > w_{cohexcl}$. From this analysis, we concluded that a combination of two relevance factor, i.e four considered criteria, may yield further improvements. The obtained retrieval results are presented in rows six to eight in Table 7.7. In this case, we observed that using the coverage and coherence criteria and weighting more the *covexcl* criterion, lead to the optimal retrieval performance that can be achieved by DtMRF$_{\text{TOPSIS}}$.

All in all, our experiments showed that the proposed framework can be also used as a re-ranker, while its DtMRF$_{\text{TOPSIS}}$ instantiation is robust to the re-ranking

Table 7.7: Retrieval effectiveness when the top-n retrieved documents by the DFR$_{\text{cin}}$ method are re-ranked using the DtMRF$_{\text{TOPSIS}}$. Each row presents the obtained effectiveness for every possible combination of criteria, in the found optimal setting.

	Bpref	Rprec	MRR	P@1	P@5	P@10	NDCG@10	R@50
DFR$_{\text{bsl}}$	.168	.238	.467	.276	.309	.279	.496	.179
DFR$_{\text{cin}}$	.218	.267	.538	.370	.326	.298	**.519**	.186
DtMRF$_{\text{TOPSIS}}$top [.0,1.]@10	.224°	.271°	.593°	.402^{-}	.353^{-}	.312^{-}	.521^{-}	.220°
DtMRF$_{\text{TOPSIS}}$cov [.5,.5]@75	.220°	.272°	.601°	.412^{-}	.371^{-}	.341°	.521^{-}	.220°
DtMRF$_{\text{TOPSIS}}$coh [.8,.2]@10	.220°	.267°	.571°	.431°	.331^{-}	.312^{-}	.518^{-}	.220°
DtMRF$_{\text{TOPSIS}}$top_cov [.1,.6,.2,.1]@10	.220°	.266°	.602°	.443°	.351^{-}	.312^{-}	.523^{-}	.220°
DtMRF$_{\text{TOPSIS}}$top_coh [.1,.1,.7,.1]@50	.220°	.268°	.612°	.473°	.331^{-}	.332^{-}	.501^{-}	.220°
DtMRF$_{\text{TOPSIS}}$cov_coh [.1,.7,.1,.1]@1000	.220°	**.282**$^{\circ}$	**.631**$^{\circ}$	**.471**$^{\circ}$	.362^{-}	**.353**$^{\circ}$	.501^{-}	.220°
DtMRF$_{\text{TOPSIS}}$6C [.1,.1,.2,.1,.2,.3]@75	**.224**$^{\circ}$	.269°	.602°	.443°	**.387**$^{\circ}$	.331^{-}	.514^{-}	**.224**$^{\circ}$

depth, and further improves retrieval effectiveness. Regarding the studied task, also in this case, the obtained optimal weights support our initial intuition that the performance scores obtained from a trial's exclusion criteria contribute negatively to its utility. Finally, the coverage criterion turned out being the strongest relevance factor for the determination of a trial's eligibility, in this re-ranking setting.

However, the values presented in Table 7.6 and Table 7.7 constitute the upper effectiveness bound achieved by the four DtMRF instantiations. Therefore, we have conducted a detailed analysis of the best performing instantiation, i.e DtMRF$_{\text{TOPSIS}}$, when only the coverage and coherence criteria are considered for re-ranking. This analysis is presented in the following section, and it involves an investigation in terms of achieved performance regarding the criteria weights and the re-ranking depth.

Re-Ranking Analysis. Figure 7.4 presents the MRR value across different combinations of weights associated with the *covincl*, *covexcl*, *cohincl*, *cohexcl* criteria. Due to the high number of possible combinations, the MMR values achieved by the DtMRF$_{\text{TOPSIS}}$ instantiation for re-ranking are presented in descending order, and are compared with the DFR_{cin} baseline and the best performing single run retrieval approach achieved by the $\text{DtMRF}_{\text{TOPSIS}}$.

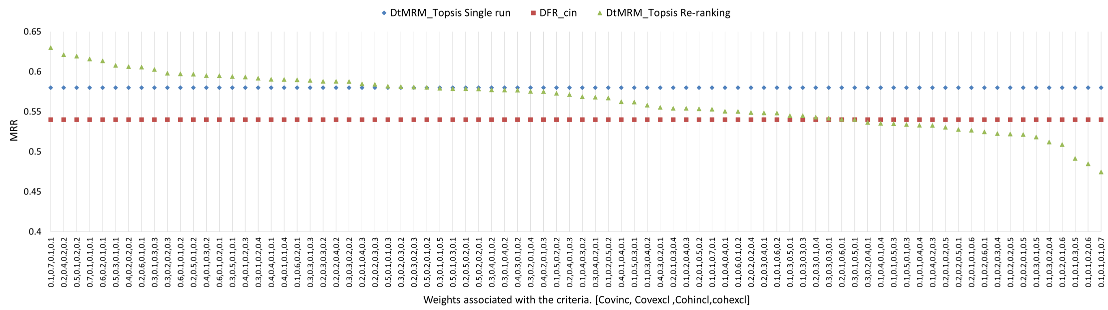

Figure 7.4: $\text{DtMRF}_{\text{TOPSIS}}$'s sensitivity to the weights associated with the criteria, when used to re-rank the 1000 retrieved documents.

As it can be seen in the figure, for the majority of the considered weights, the $\text{DtMRF}_{\text{TOPSIS}}$ instantiation for re-ranking outperforms the DFR_{cin} baseline, suggesting that, indeed, incorporating the negative influence associated with some criteria, improves the retrieval effectiveness in this task. Moreover, $\text{DtMRF}_{\text{TOPSIS}}$ for re-ranking improves the retrieval performance also over the single run $\text{DtMRF}_{\text{TOPSIS}}$ approach, proving that further relevance factors should be considered in this search task. Lastly, we observe that for some combinations of weights, $\text{DtMRF}_{\text{TOPSIS}}$ downgrades the retrieval performance. In these combinations, the weights associated with the coverage criteria are relatively lower compared to those of the coherence criterion. This finding suggests that relying on the coherence criterion is not sufficient for retrieving eligible clinical trials, i.e the evaluation function employed to estimate this criterion is not accurate.

The $\text{DtMRF}_{\text{TOPSIS}}$ instantiation improves the retrieval performance in this task for various criteria weights as long as the utility estimation relies mostly on the coverage criterion and in particular with the *covexcl* criterion. Following that, we have continued our analysis by investigated the $\text{DtMRF}_{\text{TOPSIS}}$'s sensitivity to the re-ranking depth for the optimal weights. This analysis is presented in Figure 7.5. Here, we notice that the this instantiation is also robust with respect to the depth of re-ranking, as it outperforms the considered baseline (DFR_{cin}) for all re-ranking depths (10 to 1000).

Also, in this re-ranking setting, all the employed retrieval approaches have been evaluated using condensed result lists. As the initial retrieval approach (DFR_{cin})

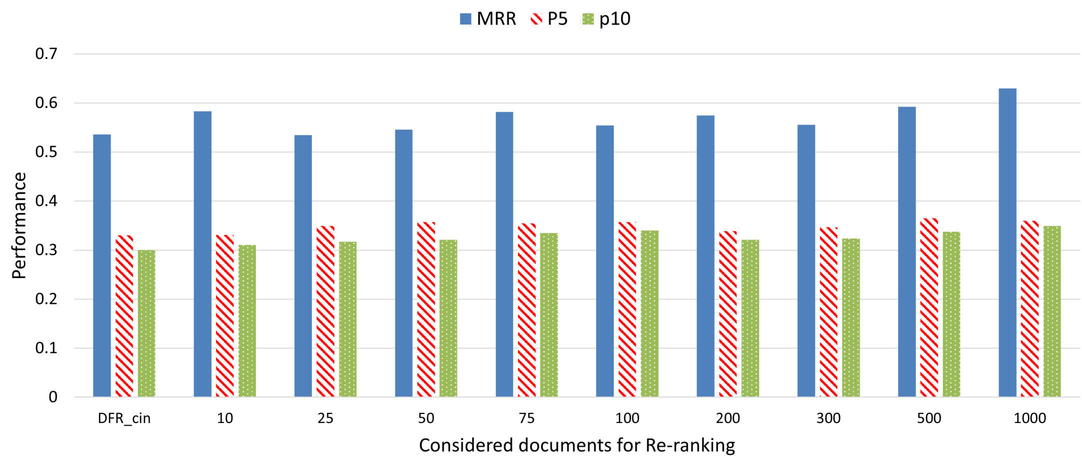

Figure 7.5: $\text{DtMRF}_{\text{TOPSIS}}$'s sensitivity to the number of documents considered for re-ranking, when the optimal weights are used.

retrieves on average 1.1 unjudged documents in the top-10 positions, it is expected that as the depth of re-ranking increases, so does the number of unjudged documents retrieved by $\text{DtMRF}_{\text{TOPSIS}}$ re-ranker. Even when the re-ranking depth is 1000 (extreme case), across all 75 queries, $\text{DtMRF}_{\text{TOPSIS}}$ has a median of 7 and an average of 6.4 unjudged retrieved documents, meaning that at least 3 of the retrieved documents in the top-10 positions have been judged. Therefore, following a condensed list evaluation in this retrieval setting does not significantly affect the measured retrieval effectiveness.

To conclude, also in this retrieval approach our initial intuition is supported by the experimental results as the optimal retrieval effectiveness is obtained when the negative effect imposed by a trial's exclusion criteria is heavily considered by the $\text{DtMRF}_{\text{TOPSIS}}$ instantiation.

7.2.1.4 Comparing DtMRF with Related Studies

In this section, we compare the retrieval performance of our two top-performing experiments conducted in our research study with the top three approaches in TREC clinical trials track 2021 and the median performance achieved in TREC. In addition, we present the retrieval effectiveness that can be achieved if, for each query, we use a set of importance weights that optimize its nDCG@10. This experiment aims to show the upper effectiveness bound of DtMRF, that can be

Table 7.8: Comparison with other approaches reported in TREC 2021 clinical trials track and TREC's median.

	P@10	MRR	nDCG@10
TREC's Median	.160	.290	.300
University of Waterloo (approach: f_t_mt5_2) [Pradeep et al., 2022]	.593	.816	.712
Alibaba Group (approach: damoebrtog)	.410	.610	.600
CSIROmed Team (approach: CSIROmed_inc) [Rybiński et al., 2021]	.320	.500	.530
$\text{DtMRF}_{\text{TOPSIS}}$ [.5,.1,.4] (single run)	.302	.583	.510
$\text{DtMRF}_{\text{TOPSIS\_cov\_coh}}$[.1,.7,.1,.1]@1000 (re-ranking approach)	.353	.631	.501
$\text{DtMRF}_{\text{TOPSIS}}$ [optimal per query] (single run)	.430	.780	.640

achieved if one can predict, for each query an optimal set of weights, instead of using the same weights for all queries.

The retrieval effectiveness achieved by the state-of-the-art approach, introduced by Pradeep et al. [2022], is presented in the second row of Table 7.8. We remind that the authors adopt a multi-stage neural ranking approach in their study. They specifically employ a technique known as neural query synthesis (NQS), which involves utilizing a zero-shot document expansion model. Using NQS, the authors generate forty sentence-long queries given a clinical note. These queries and the original clinical note are independently used as input in a retrieval pipeline that utilizes the BM25 and RM3 models. The document rankings obtained from this process are fused to form the first-stage retrieval. Subsequently, a neural re-ranker, based on the fine-tuned monoT5 model for clinical trial retrieval, is employed to generate the final ranking. Unfortunately, due to the lack of provided experimental details, it is not feasible to comment on the details of the second-performing retrieval approach. However, one can notice a significant performance gap compared to the state-of-the-art method. Finally, the third performing approach introduced by Rybiński et al. [2021], proposes a two-stage framework for clinical trial retrieval. The first stage involves an initial retrieval phase using a Divergence from Randomness model; in the second stage, a neural re-ranking technique based on BioBERT is applied to the top 100 documents obtained from the initial retrieval. During re-ranking, the documents are represented by concatenating their brief titles and inclusion criteria. To determine the final score, the normalized scores from the initial ranker and BioBERT are combined using a ratio of 1:9.

Following the approaches proposed in the literature, Table 7.8 presents the two top performing experiments in the single-run and the re-ranking settings. We remind that the $\text{DtMRF}_{\text{TOPSIS}}$ approach exploits three DFR scores, computed with respect to three distinct representations of a clinical trial document as depicted in Figure 7.1. The $\text{DtMRF}_{\text{TOPSIS}}$_cov_coh approach penalizes those clinical trials which have high coverage between a given clinical note and the trial's exclusion criteria. Also, we remind that in both of these experiments, the weights are constant across all queries. Based on the reported retrieval performance, the $\text{DtMRF}_{\text{TOPSIS}}$_cov_coh approach achieves higher performance compared to the 3rd-performing approach in terms of P@10 and MRR. Also, it achieves a higher MRR value compared to the 2nd approach. Regarding our single-run retrieval approach, it achieves similar retrieval performance to the 3rd approach submitted in the TREC clinical trials track. Finally, when each query is associated with a set of weights that optimize its nDCG@10 measure, one can observe that the retrieval performance can be further improved. This performance has been achieved using the standard query processing from PyTerrier and relying solely on standard IR models (ln_expB2), i.e. it is significantly simpler than the other approaches proposed in the literature. In addition to its simplicity, it also leads to a fully interpretable document ranking.

In conclusion, the $\text{DtMRF}_{\text{TOPSIS}}$ instantiations demonstrate performance levels better or comparable to the top TREC approaches, despite relying solely on standard IR models and simple relevance signals like coverage and coherence. However, as discussed in Section 5.2.1, DtMRF can utilize various evaluation functions to assess the extent to which a document satisfies a given criterion. These evaluation functions can include those used in the top-performing approaches, such as combining BM25 and RM3 models or BioBERT. Furthermore, due to its high retrieval performance and low computational overhead, the single run $\text{DtMRF}_{\text{TOPSIS}}$ approach can be effectively utilized as an initial retrieval method combined with the top-performing approaches. Lastly, the performance achieved when the set of weights that maximize nDCG@10 is used for each query, the retrieval performance can be significantly improve.

7.2.2 Conclusions and Directions for Future Research

The empirical evaluation performed on the medical IR task showed the benefits of incorporating task-related characteristics in the retrieval process, mentioned its shortcomings and outlined several potential improvements and expansions. The consideration of relevance factors that negatively impact the overall relevance in the studied task, lead to greater retrieval performance compared to the standard retrieval paradigm (i.e. considering only positive relevance factors). The $\text{DtMRF}_{\text{TOPSIS}}$ instantiation achieved the highest retrieval performance across all retrieval experiments and considered relevance factors. Also, despite relying on simpler relevance signals, $\text{DtMRF}_{\text{TOPSIS}}$ achieved better or comparable performance to the top-performing approaches in the literature that rely on neural models. Furthermore, DtMRF offers the advantage of lower computational complexity, interpretability, and the possibility to be combined with other approaches for topical relevance estimation, like BERT-based models. In summary, the conducted experimental evaluations support the effectiveness of DtMRF in producing interpretable document rankings, and its ability to leverage both positive and negative relevance aspects while maintaining competitive performance and computational efficiency. Future research could include more relevance factors, or the incorporation of patient-related (e.g. age, gender), or trial-related attributes (e.g. location) in the retrieval process.

7.3 Leveraging Neural-DtMRF for Clinical Trials Retrieval

Table 7.8 presents the retrieval effectiveness that $\text{DtMRF}_{\text{TOPSIS}}$ can achieve if each query is associated with the weights that optimize its nDCG@10 measure. Therefore, this section presents the usage of Neural-DtMRF aiming to predict a set of weights for each query. Specifically, our exploration aims to address the following research questions:

(RQ1) To what degree can a neural model predict the importance weights for $\text{DtMRF}_{\text{TOPSIS}}$ in the context of clinical trials retrieval?

(RQ2) How effectively can $\text{Neural-DtMRF}_{\text{TOPSIS}}$ enhance retrieval performance in comparison to $\text{DtMRF}_{\text{TOPSIS}}$ and BM25?

(RQ3) How does the retrieval performance of $\text{Neural-DtMRF}_{\text{TOPSIS}}$ stand when compared to alternative methods in existing scholarly works?

In answering these questions, we focus our investigation on the single-run retrieval approach presented in Figure 7.1, and use the $\text{DtMRF}_{\text{TOPSIS}}$ instantiation since it provided the best retrieval performance. However, we have replaced the ln_expB2 model with the BM25 model for estimating topical relevance, as it achieved greater retrieval performance results in the TREC 2022 collection.

7.3.1 Experimental Design and Results

Neural-DtMRF shares foundational components with DtMRF; consequently, this section concentrates on those elements related explicitly to Neural-DtMRF. These include the process of selecting the optimal importance weights (model's outputs), the model's inputs, and the neural architecture utilized for prediction. In the context of this search task, creating a specialized training dataset is unnecessary, as the characteristics of the existing benchmark collections meet the requirements for clinical trials retrieval. Thus, the experimental setup employs all three benchmark collections specified in Section 3.3.4, namely TREC 2021, TREC 2022, and *Clinical*.

7.3.1.1 Selection of Optimal weights

The process of selecting the optimal weights to be associated with each query in the collections is described as follows. Initially, retrieval is conducted using the $\text{DtMRF}_{\text{TOPSIS}}$ instantiation, exploring all possible combinations of the $[w_{main}, w_{incl}, w_{excl}]$ weights, incremented by steps of 0.05. We choose the weight combinations that maximize the nDCG@10 metric for each query as it serves as the principal evaluation measure in the TREC 2021 and 2022 clinical trials tracks, making it apt for the search task at hand. Subsequent to this selection, it is possible to categorize the retrieval effectiveness of $\text{DtMRF}_{\text{TOPSIS}}$ into four distinct types, as depicted in Figure 7.6.

The figures present the nDCG@10 measure across four representative queries from the TREC 2021 collection, when $\text{DtMRF}_{\text{TOPSIS}}$ is associated with different importance weights. The red rectangle represents the weight combination of [0.5,0.1,0.4], which results in the optimal performance of $\text{DtMRF}_{\text{TOPSIS}}$ when applied to all queries in the TREC 2021 collection. The $\times$ symbol denotes the weights that specifically optimized the nDCG@10 metric for the presented query. Furthermore, within the legend of each sub-plot, we display the baseline score for the query as determined by the BM25 model.

The top-left sub-figure illustrates the scenario where $\text{DtMRF}_{\text{TOPSIS}}$ outperforms the baseline even when non-optimal weights are used; in this context, the sub-optimal weight combination already yields satisfactory results. The top-right sub-figure depicts a situation where employing $\text{DtMRF}_{\text{TOPSIS}}$ offers no added value, as its optimal performance matches that of the BM25 baseline. The bottom-left sub-figure demonstrates the case where $\text{DtMRF}_{\text{TOPSIS}}$ significantly benefits from the use of the query's optimal weights, as the model performs better than the baseline but does not reach its upper effectiveness bound. Finally, the bottom-right sub-figure highlights the necessity of pinpointing the optimal weights to ensure performance that exceeds the baseline. It is crucial to note that there were queries for which $\text{DtMRF}_{\text{TOPSIS}}$ yielded performance that fell below the baseline, even when its optimal weights were employed.

Two observations can be drawn after examination of the figures, focusing on the color-coded nDCG@10 values. Firstly, one can observe group weights (clusters) that consistently perform well regarding nDCG@10 (high yellow bullets). These clusters can be interpreted as regions in the weight space where $\text{DtMRF}_{\text{TOPSIS}}$ performs well. Furthermore, the figures reveal a degree of "space for errors in predictions," implying that the model exhibits a level of tolerance for sub-optimal weight combinations. In practical terms, this suggests that even if the chosen weights are not perfectly optimized, the degradation in retrieval effectiveness may still be within an acceptable range, thereby providing some freedom in real-world applications.

The collections under consideration contain a total of 185 queries. Utilizing only a single set of optimal weights for each query would result in 185 samples, which must be enhanced to train a neural model effectively. To address this limitation, we assign optimal and near-optimal weights to each query to augment the number of training instances. By implementing this strategy, we generated a training dataset comprising approximately three thousand query-weight instances.

Another constraint inherent in our experimental setup pertains to the limitations of the available search context we could leverage for training, i.e. the model's inputs x. Our experiments leveraged solely the query context, generating an embedding representation utilizing pre-trained models. In the subsequent section, we provide

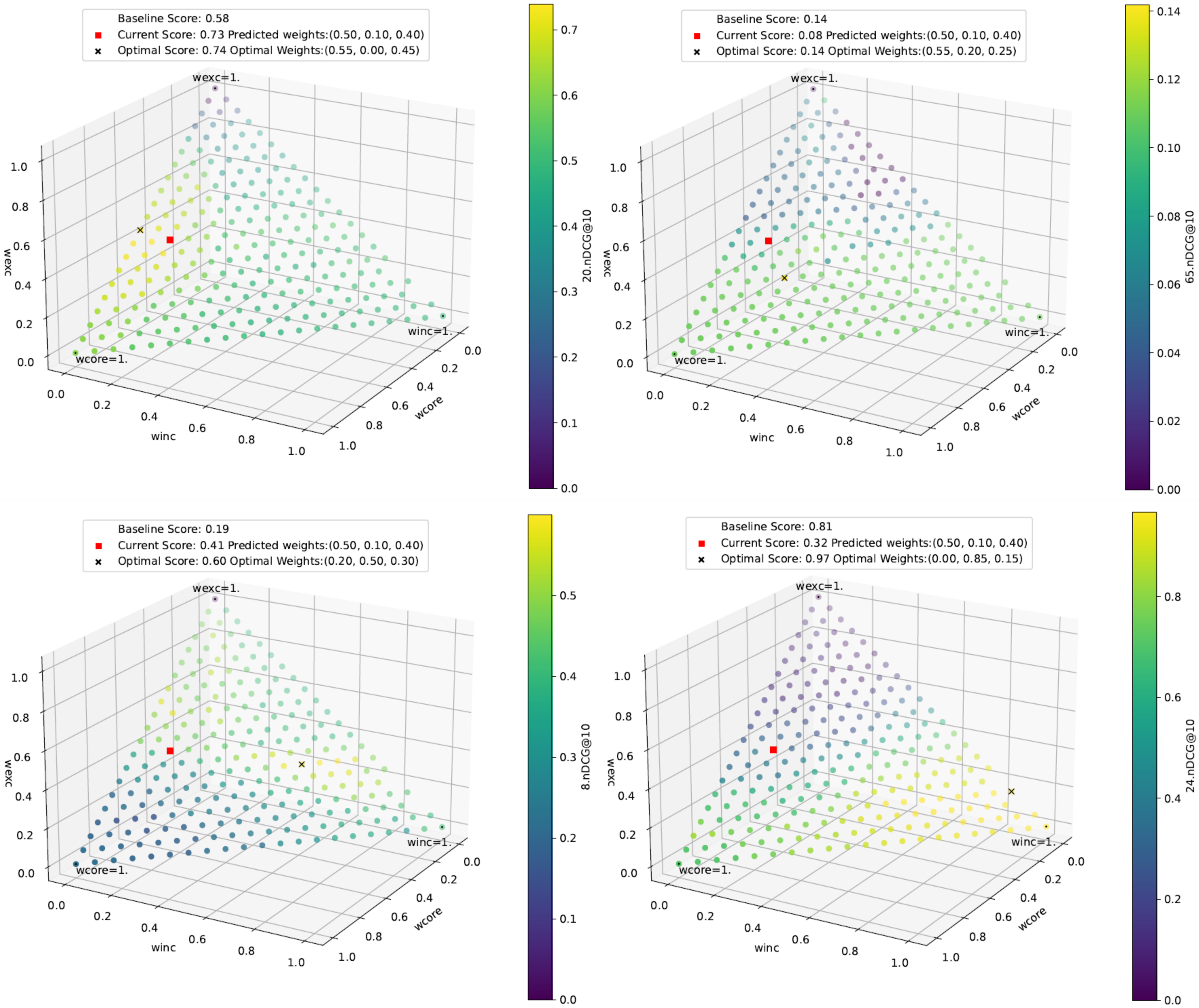

Figure 7.6: Four representative examples of the different retrieval behavior of DtMRF$_{\text{TOPSIS}}$ at different weight combinations.

an in-depth analysis of the neural network employed for weight prediction, as well as elaborate on the training methodology.

7.3.1.2 Neural Architecture and Training Parameters

In our experimental setup, we utilize the Clinical collection exclusively for training purposes. To elaborate, we either train on the combined Clinical and TREC 2021 datasets and subsequently predict on the TREC 2022 collection or train on the combined Clinical and TREC 2022 dataset and make predictions on the TREC 2021 dataset.

The model architecture can be described as follows. The neural model leverages a pre-trained BioBERT [Lee et al., 2020b] for initial text encoding, serving as the embedding layer that converts each token in the input query into a 768-dimensional vector. Following this, a neural network tailored for multi-output regression is employed. It comprises a multihead attention layer with 8 attention heads aiming to capture different aspects of the relationships between the words in the input sequence. Next, two fully connected layers (each with 128 dimensions) are used before the final output layer that transforms their 128-dimensional output into a 3-dimensional space corresponding to the weights w_{main}, w_{incl}, w_{excl}. For the training configuration, the model employs the Mean Squared Error (MSE) as its loss function and utilizes the Adam optimizer with a learning rate of 1×10^{-5}. We train for a total of 100 epochs with a batch size of 16. This setup aims to effectively optimize the model parameters for predicting the optimal and near-optimal weights for each query. The subsequent section delineates the results garnered across both collections.

7.3.1.3 Effectiveness of Neural-DtMRF: Single Run Retrieval Approach

Table 7.9 showcases the retrieval performance of the baseline BM25 model, in comparison with the performances of DtMRF$_{\text{TOPSIS}}$ employing various weight configurations, and Neural-DtMRF$_{\text{TOPSIS}}$ that leverages predicted weights for all queries. The weights that lead to the optimal retrieval performance when used across all of the queries in the collection are equal to [.5,.1,.4] for the TREC 2021 collection and [.65,.25,.1] for TREC 2022. As it can be seen, these weights lead to improvements when compared to the BM25 in TREC 2021, but lead to similar performance in the TREC 2022 collection. Therefore, the need for query-dependent weights becomes apparent.

Table 7.9: Performance comparison between the baseline, the DtMRF$_{\text{TOPSIS}}$, and the Neural-DtMRF$_{\text{TOPSIS}}$.

	TREC 2021			TREC 2022		
	nDCG@10	P@10	MRR	nDCG@10	P@10	MRR
BM25	.469	.264	.471	.394	.272	.507
DtMRF$_{\text{TOPSIS}}$ [Opt. W for collection]	.500$^-$	.300$^-$	.580°	.399$^-$	.267$^-$	.480$^-$
Neural-DtMRF$_{\text{TOPSIS}}$	**.527**$^\circ$	**.321**$^\circ$	**.594**$^\circ$	**.438**$^\circ$	**.312**$^\circ$	**.624**$^\circ$
DtMRF$_{\text{TOPSIS}}$ [opt. W per query]	.640°	.430°	.780°	.530°	.400°	.760°

Neural-DtMRF$_{\text{TOPSIS}}$ exhibits superior performance over the BM25 baseline across all evaluation metrics and on both the TREC 2021 and TREC 2022 collections. This underscores the effectiveness of leveraging a neural model for query-specific weight prediction. All performance gains are statistically significant, as verified by a paired t-test with Bonferroni correction. Furthermore, in the TREC 2021 collection, the model demonstrates low absolute prediction errors for the weights: .183 for w_{main}, .164 for w_{incl}, and .151 for w_{excl}. Comparable performance is noted on the TREC 2022 collection. In response to our first research question, a neural model can be used to predict the optimal weights for a query in the studied task as both the absolute errors are low, and the achieved performance is greater than the baseline. Moreover, we observe effectiveness improvements also when we compare Neural-DtMRF$_{\text{TOPSIS}}$ to the DtMRF$_{\text{TOPSIS}}$ that leverages the same weights for all queries. This further validates the effectiveness of incorporating a neural model for dynamic weight prediction in enhancing the overall retrieval performance. Nonetheless, by observing the last row in the table, one can conclude that the current performance of Neural-DtMRF$_{\text{TOPSIS}}$ is still considerably below the upper effectiveness bound achievable with a more precise weight prediction model.

Regarding the final research question, Table 7.10 presents the retrieval performance of the top three methods in the clinical trials tracks of both TREC 2021 and TREC 2022. Notably, these leading approaches employ multi-stage retrieval processes, followed by a re-ranking step that utilizes pre-trained language models. By contrast, our proposed methodology operates as a single-run retrieval stage, relying on the BM25 model for topical relevance estimation. While most of these leading approaches have been discussed in earlier sections, we direct interested readers to the official TREC proceedings for a more comprehensive understanding[89].

Table 7.10: Performance comparison of Neural-DtMRF$_{\text{TOPSIS}}$ to the top-performing approaches in TREC 2021 and TREC 2022.

	TREC 2021			TREC 2022		
	nDCG@10	P@10	MRR	nDCG@10	P@10	MRR
TREC's 1st Approach	.712	.593	.816	.613	.508	.726
TREC's 2nd Approach	.600	.410	.610	.556	.456	.619
TREC's 3rd Approach	.530	.320	.500	.505	.398	.606
Neural-DtMRF$_{\text{TOPSIS}}$	.527°	.321°	.594°	.438°	.312°	.624°
DtMRF$_{\text{TOPSIS}}$ [opt. W per query]	.640	.430	.780	.530	.400	.760

Even though the current version of Neural-DtMRF$_{\text{TOPSIS}}$ is a single-step retrieval approach based on the BM25 model, its performance in the TREC 2021 clinical trials track is competitive, closely matching that of the third-best performing approach, that leverages BERT for re-ranking. However, in the TREC 2022 track, the model exhibits a performance gap compared to the leading methods. Nonetheless, should we successfully predict the optimal weights, it becomes evident that the performance is on par with that of more intricate retrieval pipelines and BERT-based re-ranking models.

7.3.2 Conclusions and Directions for Future Research

In conclusion, this work sheds light on the potential of utilizing a neural model for predicting optimal weights for each query in the context of clinical trials retrieval. Our experiments suggest that Neural-DtMRF$_{\text{TOPSIS}}$ is a better approach for clinical trials retrieval compared to DtMRF$_{\text{TOPSIS}}$ that use the same weights for all queries. Moreover, our findings indicate that even without leveraging complex BERT-based re-ranking mechanisms, the model can achieve comparable performance when optimal weights are used.

As for future work, there are two distinct directions. One potential direction is related to the created training dataset, and specifically the generation of the optimal weights. One possible modification involves introducing zero or negative weights for those queries where the $\text{DtMRF}_{\text{TOPSIS}}$ retrieval performance is below a considered baseline. By implementing this alteration in generating optimal weights, we aim to train the neural model to recognize scenarios where the utilization of $\text{DtMRF}_{\text{TOPSIS}}$ may not be advantageous. Another avenue for future research involves employing multiple instances of $\text{DtMRF}_{\text{TOPSIS}}$ with varying combinations of weights in a parallel configuration. Lastly, there is the opportunity to adopt more sophisticated evaluation functions that could provide a more accurate estimation of performance scores.

In the subsequent section, we conduct a concluding experiment to examine the potential of LLMs in assessing a patient's eligibility for a clinical trial. The underlying premise of this approach stands in complete contrast to that of DtMRF, specifically in terms of interpretability. While DtMRF offers insights into the decision-making process, the LLM-based approach lacks such interpretability, causing the reasoning behind its decisions to be unclear.

7.4 Leveraging LLMs for Clinical Trials Retrieval

In this section, we present our preliminary findings related to an approach that utilizes LLMs to determine a patient's eligibility for participation in clinical trials. The underlying premise of this idea contrasts with the methodologies employed in our previous experiments, as in this case we have reduced control over the system's operation and determination of relevance (i.e. eligibility). Here, an AI agent takes the role of a decision-maker on behalf of an actual expert user. Consequently, we have constructed a fully automated pipeline characterized by reduced interpretability. Our research provides insights and answers regarding the following research questions.

(RQ1) To what extent does the integration of LLMs into the clinical trials eligibility estimation pipeline impact its effectiveness compared to the previously proposed methods?

(RQ2) What are the computational costs and time requirements associated with the screening process of 6,250 clinical trials (50 trials per patient) using an LLM?

7.4.1 Experimental Design and Results

The proposed retrieval pipeline uses a summary of a patient's information as input in a retrieval model to retrieve potentially eligible clinical trials. For this initial retrieval step, we experimented with two retrieval approaches one that leverages the BM25 model, and one that leverages the Neural-DtMRF$_{\text{TOPSIS}}$. Then, an LLM determines a patient's eligibility by examining the top 50 clinical trials retrieved. The LLM uses a specifically designed prompt, the patient's information, and the eligibility section of each trial. Ultimately, each patient's 1000 retrieved clinical trials are ranked to position eligible trials at the top, sorted based on their initial retrieval scores. Subsequently, non-eligible trials are organized in descending order of their initial retrieval scores. We experimented with the gpt-3.5-turbo model and a quantized version of the Falcon7B Instruct model[1], in two distinct experiments. The employed Falcon7B Instruct model has been quantized into a 4-bit representation leveraging the QLoRA method [Dettmers et al., 2023]. Both LLMs operated on default parameters, with the temperature parameter set to 0 to ensure more deterministic behavior.

Table 7.11: Prompts used with the two LLMs aiming to determine a patient's eligibility.

Prompt Text (Input to GPT3.5)
Based solely on the specific patient and trial information provided, without making any generalizations or assumptions, and disregarding any age, location, or gender requirements, indicate whether the patient is eligible for the clinical trial. Respond with a simple "YES" if eligible, or "NO" if not eligible. Patient information: *{patient information}* Trial information: *{trial's eligibility}*

Prompt Text (Input to Falcon7B Instruct)
Based solely on the specific patient and trial information provided, *{Patient information}* Trial information: *{Trial's eligibility}*. Is the patient an eligible participant "YES" or "NO"? Answer:

[1] Falcon7B Instruct model, 4-bit quantization.

Table 7.11 showcases the two prompts employed in our experiments. The prompt used for the Falcon7B Instruct model is notably simpler than the one provided for the GPT3.5 model. This approach was implemented based on empirical findings from our experiments, which revealed that Falcon7B encountered challenges in delivering a clear binary response of "YES" or "NO" when the prompt text contained additional information.

7.4.1.1 Effectiveness of LLM-based Approach: Re-ranking Retrieval Approach

Table 7.12 provides a comprehensive performance comparison between the proposed LLM-based re-ranking approaches with the top-performing methods at TREC, the BM25, and the Neural-DtMRF$_{\text{TOPSIS}}$ approach. The experiments involving BM25 + Falcon7B (4bit) and BM25 + GPT3.5 are designed to assess the capabilities of these two LLMs in determining a patient's eligibility for clinical trials. The results from these experiments reveal that the Falcon7B model appears to hurt the retrieval performance when combined with BM25. The downgrade in retrieval performance indicates that the Falcon7B model, despite its potential benefits, may not be the optimal choice for this specific task compared to GPT3.5. This finding highlights the importance of selecting the most suitable LLM model for specific information retrieval tasks, as different models may have varying degrees of effectiveness depending on the nature of the task. In the context of patient-trial eligibility assessments, our error analysis showed that Falcon7B mainly categorized patients as non-eligible in most cases. The decrease in the P@10 metric implies that there are instances where the model misclassifies trials as eligible, causing them to appear among the top positions in the rankings.

Table 7.12: Performance comparison of LLM-based re-ranking approaches to the top-performing approaches in TREC 2021 and TREC 2022.

	TREC 2021			TREC 2022		
	nDCG@10	P@10	MRR	nDCG@10	P@10	MRR
TREC's 1st Approach	.712	.593	.816	.613	.508	.726
TREC's 2nd Approach	.600	.410	.610	.556	.456	.619
TREC's 3rd Approach	.530	.320	.500	.505	.398	.606
BM25	.469	.264	.471	.394	.272	.507
Neural-DtMRF$_{\text{TOPSIS}}$	.527°	.321°	.594°	.438°	.312°	.624°
BM25 + Falcon7B Instruct (4bit)	.437°	.250$^-$	.487$^-$	.382$^-$	.258°	.440°
BM25 + GPT3.5	.536°	.342°	.612°	.540°	.416°	.644°
Neural-DtMRF$_{\text{TOPSIS}}$ + GPT3.5	.601°	.483°	.634°	.563°	.462°	.748°

Hence, we opted to use the GPT3.5 model in conjunction with the Neural-DtMRF$_{\text{TOPSIS}}$ model. The Neural-DtMRF$_{\text{TOPSIS}}$ + GPT3.5 experiments show a noteworthy performance that is on par with the top-performing approaches of TREC 2021 and TREC 2022. Notably, it achieves superior MRR performance in TREC 2022. We remind that this approach leverages BM25 for initial retrieval, while the other TREC approaches leverage Pre-trained Language Models for re-ranking. Regarding the second research question, the Falcon7B model is faster than the GPT3.5 model, however it is not capable to improve retrieval performance. GPT3.5 successfully processed 6,250 clinical trials, assessing the eligibility of 125 patients, utilizing OpenAI's API. This task was completed within one hour and thirty minutes, incurring a cost of 25 dollars for the computational resources and services utilized during the processing. The cost associated with using GPT3.5 through OpenAI's API can vary based on several factors, including the length of a clinical trial's eligibility section and the amount of patient information provided. Therefore, the specific details and characteristics of the input data influence the overall cost of using the API for tasks like clinical trial eligibility assessment.

7.4.2 Conclusions and Directions for Future Research

Our research has revealed that the combination of Neural-DtMRF$_{\text{TOPSIS}}$ and GPT3.5 produces results that are on par with state-of-the-art methods, all while maintaining a straightforward and efficient pipeline. This outcome underscores the potential of leveraging a highly interpretable model to obtain a set of clinical trials with specific characteristics combined with advanced language models for clinical trial eligibility assessment.

Moving forward, we intend to explore the development of a smaller model through knowledge distillation, leveraging the insights gained from GPT3.5. This approach seeks to balance model efficiency and performance, crucial for resource-intensive tasks like clinical trial assessment. Additionally, our research will focus on utilizing GPT models for data generation and the subsequent training of smaller models. This synergy between data generation and model training holds promise for enhancing the overall pipeline's efficiency and effectiveness.

7.5 Discussion

In this chapter, we have conducted a comprehensive exploration of different approaches for clinical trials retrieval and eligibility assessment. Our empirical evaluation has shown the effectiveness of the $\text{DtMRF}_{\text{TOPSIS}}$ instantiation in clinical trials retrieval. This approach considers both positive and negative relevance factors, leading to improved retrieval performance compared to the standard retrieval paradigm. $\text{DtMRF}_{\text{TOPSIS}}$ has shown competitive performance, while leveraging simpler relevance signals for relevance estimation. Moreover, it offers advantages in terms of computational efficiency, retrieval interpretability, and ability to put user in control of the search process.

We have introduced the Neural-DtMRF for clinical trials retrieval, which leverages a neural models to predict an optimal set of importance weights for each patient. Our experiments indicate that this approach outperforms the standard $\text{DtMRF}_{\text{TOPSIS}}$ approach that leverages the same weights across queries. This observation indicates that a neural model can improve clinical trials retrieval without the need for its application in a re-ranking setting for relevance estimation.

Finally, we have explored the integration of LLMs, into the clinical trial eligibility assessment process. The combination of Neural-$\text{DtMRF}_{\text{TOPSIS}}$ and GPT3.5 has shown promise, achieving performance comparable to state-of-the-art methods while maintaining a more simple pipeline. The cost associated with this process is within an expected range, making it a viable option for fully automating this search task.

Our future research encompasses several pivotal directions. One direction entails harnessing GPT models for data generation. This approach will enable us to enhance Neural-DtMRF's training. Also, we aim to develop a specialized model for relevance estimation in this particular task, that can be used in place of BM25. Finally, we aim to explore the performance potential of open-source models beyond Falcon7B, which we utilized in this study.

Part IV

From Conceptualization to Development: A Search Prototype

Chapter 8

A prototype Search System for Clinical Trials Retrieval and Patients' Eligibility Screening

This chapter introduces the first version of a search prototype explicitly designed for clinical trials retrieval, focusing on the eligibility screening process. The prototype is formulated to accommodate various requirements inherent to distinct phases of the process. This chapter exploits a user viewpoint to elucidate emerging design constraints and identify procedures that significantly increase human effort. It subsequently presents a comprehensive overview of the proposed prototype, elaborating on its capabilities to address the identified challenges associated with the search task. The chapter examines the technical aspects of each prototype component and provides examples of the user interface to illustrate its applications within a professional setting.

8.1 Introduction

This chapter introduces the first version of our prototype explicitly designed to aid healthcare professionals in retrieving clinical trials and performing eligibility screening for patients. Although Chapter 4 provides an exhaustive analysis of the clinical trials retrieval task, in this chapter, we analyze it from a professional user's viewpoint. In particular, our target user group comprises clinicians and staff of medical organizations with access to confidential patient data, such as electronic health records, as well as to publicly accessible clinical trials, for instance, those published in ClinicalTrials.gov website[1]. In doing so, we examine user-system interactions, identify high-effort processes within the task, and specify design constraints, aiming to optimize the user experience through our prototype. This prototype synthesizes — or will synthesize in future versions— the research methods and findings discussed earlier, particularly in Chapters 6 and 7.

Mainly, the approaches proposed in the literature to address the task of clinical trials retrieval focus on a single phase of the search process. For example, the study by Miotto et al. [2013] presented an interactive system engineered to exclude irrelevant clinical trials from the search results. In this approach, the user inputs a query, which drives the system to retrieve a set of potentially applicable trials. Subsequently, a word cloud of eligibility tags is generated based on the content of these trials. The user is then invited to select tags that most closely align with their specific information needs. Upon selecting these pertinent tags, the system effectively filters out the considered irrelevant trials. Liu et al. [2019a] proposed another system offering interactive features for this specific task. Initially, the system performs offline information extraction of a trial's inclusion and exclusion criteria for each trial in the collection. When a user inputs a query, an information retrieval model is deployed to retrieve an initial set of potentially relevant clinical trials. Subsequently, the system dynamically generates questions derived from the eligibility criteria in the retrieved trials. It presents them to the user in a sequential manner. As the user responds to these questions, the system refines the retrieved trials by progressively excluding those ineligible. The iterative process concludes when the user decides to inspect the remaining set of trials manually. While these approaches offer inspiring solutions for specific aspects of the search process, we aim to develop a comprehensive system that addresses all those steps that require high user effort in this task. To that aim, we analyze the user-system interactions occurring when a professional user aims to identify eligible clinical trials for a patient.

User-System Interactions. An expert user has to read a patient's medical record and identify essential information such as the patient's medical condition, lifestyle habits, and family history. Following this, the user formulates a query that encompasses these various elements and conducts a search for relevant clinical trials through a specialized interface. Upon retrieving a list of potentially relevant clinical trials, the user reviews the primary information on the Search Engine Results Page (SERP), including the trial title, condition under study, and location. At this point, the user can either modify the initial query (in case the results are unsatisfactory) or proceed to read the content of a selected clinical trial. If choosing the latter, the user navigates to the document's eligibility section to examine the inclusion and exclusion criteria for patient eligibility. Often, patients may be excluded due to mismatches in criteria despite having the same medical condition as the trial's focus. Suppose the user encounters a trial for which the patient is not eligible. In that case, two options are available: either return to the SERP to explore additional trials or reformulate the search query to initiate a new search. The search process concludes when a suitable trial is identified. A graphical representation of this user-system interaction can be found in Figure 8.1.

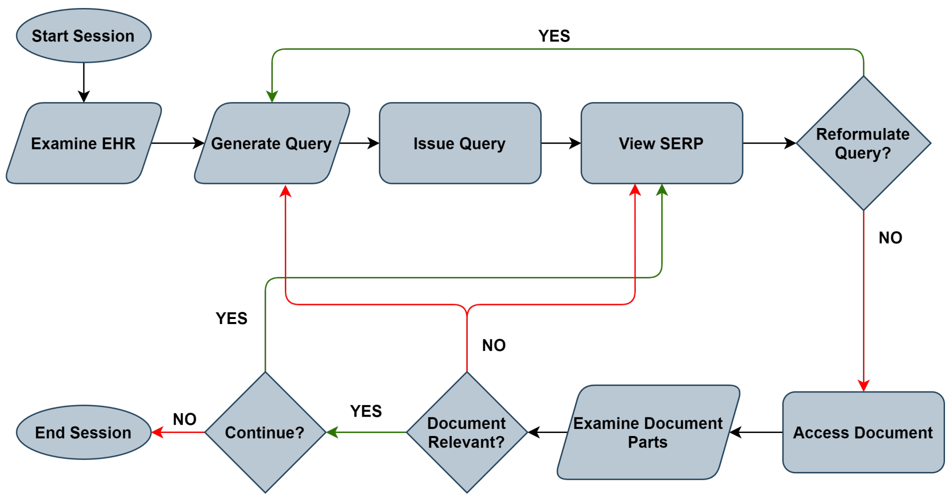

Figure 8.1: The task of eligibility screening for Clinical Trials.

Rewards and Costs in User-System Interaction. In our analysis, we investigate the user and system interactions within a single session, encompassing query formulation, query refinement, result evaluation, and progression to subsequent steps in the process. We identify two distinct phases —query generation and examination of specific document segments— that necessitate considerable cognitive effort on the user's part to advance to the next task stage. Accordingly, we delineate the following user actions:

1. User Action: Load an EHR.
2. User Action: Read the EHR.
3. User Action: (Cost Action) Extract specific aspects from the EHR.
4. User Action: Issue query.
5. User Action: Move to next page (press search button).
6. User Action: Examine SERP.
7. User Action: (Cost Action) Reformulate query.
8. User Action: Open a document.
9. User Action: Examine Document.
10. User Action: (Cost Action) Scroll down to the inclusion/exclusion criteria (Examine documents parts).
11. User Action: Read the specific document parts.
12. User Action: (Reward Action) Examine a subsequent document.

Task's Identified Bottlenecks. The procedure outlined above poses several challenges that make it a less-than-ideal solution for clinical trial retrieval and eligibility screening. Firstly, the process is time-consuming, requiring extensive effort from healthcare professionals. Secondly, its complexity makes it costly in terms of both cognitive load on the user and potential financial costs for medical organizations. Lastly, despite these investments in time and resources, the procedure is not always effective. This ineffectiveness often leads to a high rate of trial cancellations, thereby wasting the already limited resources of medical institutions. These issues interfere with the overall efficiency and risk, undermining the quality and progression of clinical research, as supported by existing studies [Brøgger-Mikkelsen et al., 2020]. In this task, two primary bottlenecks are associated with user actions, and one pertains to the system's functionality. The first bottleneck involves the extraction of information from the patient's health record, corresponding to User

Actions 1-3. The second bottleneck is linked to the time allocated for document scrutiny, represented by User Actions 8-10. Moreover, the retrieval model serves as a latent variable that adversely affects the efficacy of the search. It may retrieve studies that appear relevant to the patient's condition but exclude the patient upon closer examination, thereby diminishing the overall success rate of the search process. The retrieval of non-eligible trials substantially undermines user satisfaction, as indicated by Soboroff [2022].

The prototype is engineered to support the complex process of clinical trials retrieval and eligibility screening. It targets explicitly the three labor-intensive steps within this overarching process:

1. The prototype aims to facilitate the process of information extraction and query generation from patients' health records or summaries containing patient-related information.

2. It leverages the DtMRF model to improve retrieval efficacy, thereby minimizing the inclusion of non-eligible trials in the search results.

3. The system is designed to facilitate the eligibility assessment phase by allowing end-users to directly access specific segments of documents, such as a trial's eligibility criteria, via its user interface.

By strategically addressing these crucial challenges, the prototype seeks to optimize the overall search efficacy and improve user satisfaction.

Additional Design Constraints. Due to the task characteristics and application domain, designing the prototype raises concerns regarding data privacy and security, as it involves handling sensitive patient information. When employed in production, it is essential that the search approach is used in a manner that complies with regulations that ensure the security and confidentiality of patient-related information, such as the Health Insurance Portability and Accountability Act (HIPAA)[1] and General Data Protection Regulation (GDPR)[2].

[1] Health Insurance Portability and Accountability Act of 1996 (HIPAA).
[2] General Data Protection Regulation.

8.2 A Prototype for Clinical Trials Retrieval and Eligibility Screening

The prototype that has been developed can be segmented into three distinct backend components, each tailored to address the specific peculiarities of the search task under examination, as illustrated in Figure 8.2.

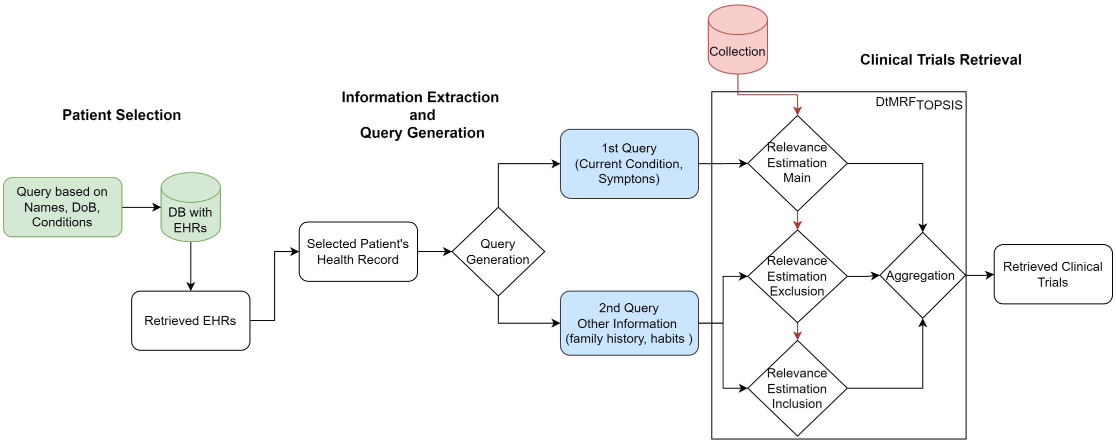

Figure 8.2: Overview of the back-end components of the prototype.

The first component of the system focuses on the stage where the healthcare provider identifies the patient and obtains their electronic health record. The subsequent component deals with query formulation. In this phase, healthcare providers can create two separate queries. The initial query is centered on the patient's medical condition and aims to identify clinical trials relevant to that condition. The secondary query considers multiple attributes of the patient, including lifestyle considerations, familial medical history, and past treatments. This query assesses a signal linked to the patient's potential eligibility for a clinical trial. Finally, the component related to clinical trial retrieval employs the DtMRF model, which incorporates retrieval requirements into the relevance estimation process. DtMRF aims to rank a clinical trial highly if the patient's information is especially relevant to the trial's medical objectives —such as the condition or treatment under investigation— and closely aligns with the trial's inclusion criteria without meeting any of its exclusion criteria.

The subsequent sections provide further details related to the functionality of each component and offer descriptions of their corresponding interfaces. Additionally, we elaborate on the integration of our previous research and findings into the search prototype, either as they currently implemented or as they are planned for future implementation.

8.2.1 Phase 1: Patient Selection

Healthcare institutions such as hospitals usually maintain proprietary databases to store patient information, including electronic health records. As a result, the prototype is engineered to retrieve a patient's EHR from an external database. An elementary MongoDB database has been set up for this prototype, allowing for queries based on medical conditions, patient names, or birth dates. The overarching objective is to provide a framework enabling healthcare institutions to incorporate their existing databases into the broader prototype system effortlessly. Figure 8.3 depicts the landing page of the prototype, highlighting the interface where healthcare providers can commence their search for a patient's EHR within the proprietary database.

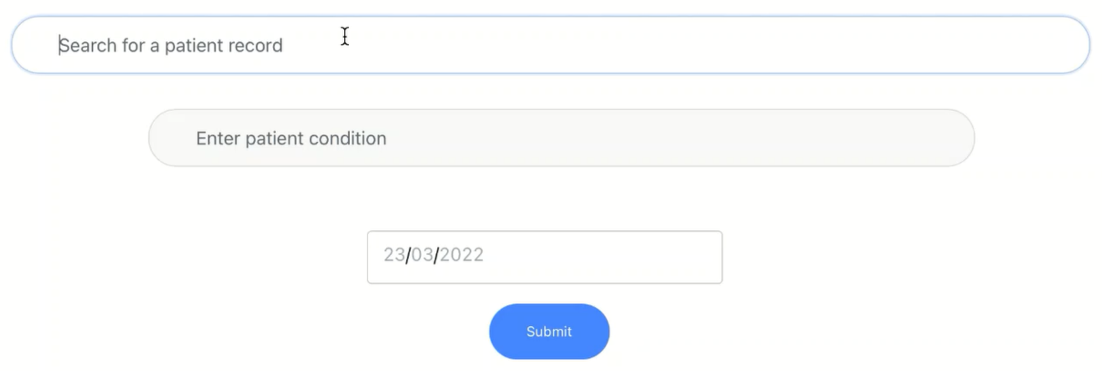

Figure 8.3: Searching the private DB (Landing page of the prototype).

Subsequently, the user is shown a list of available EHRs, as depicted in Figure 8.4. At this time, healthcare providers can open and examine the details of each EHR. This feature aims to facilitate informed selection of patients by offering comprehensive visibility into individual patient data, thereby aiding providers in choosing one patient over others based on specific criteria. Future versions will include more information, like the primary condition and lifestyle factors, extracted based on the approaches presented in Chapter 6.

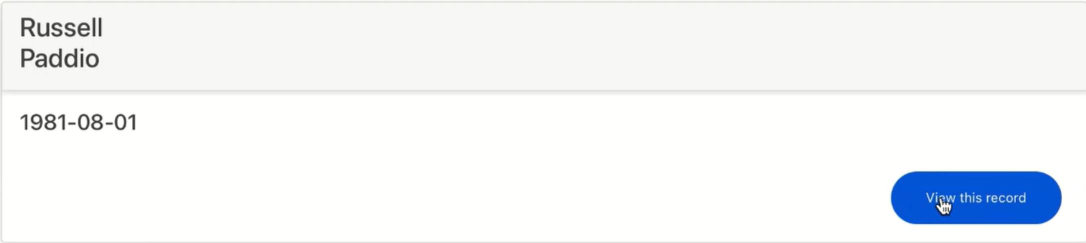

Figure 8.4: An instance of a saved EHR in the database.

Upon clicking the "View this record" button, the user is presented with the view depicted in Figure 8.5. This view allows the user to examine a summary of the patient's information in detail.

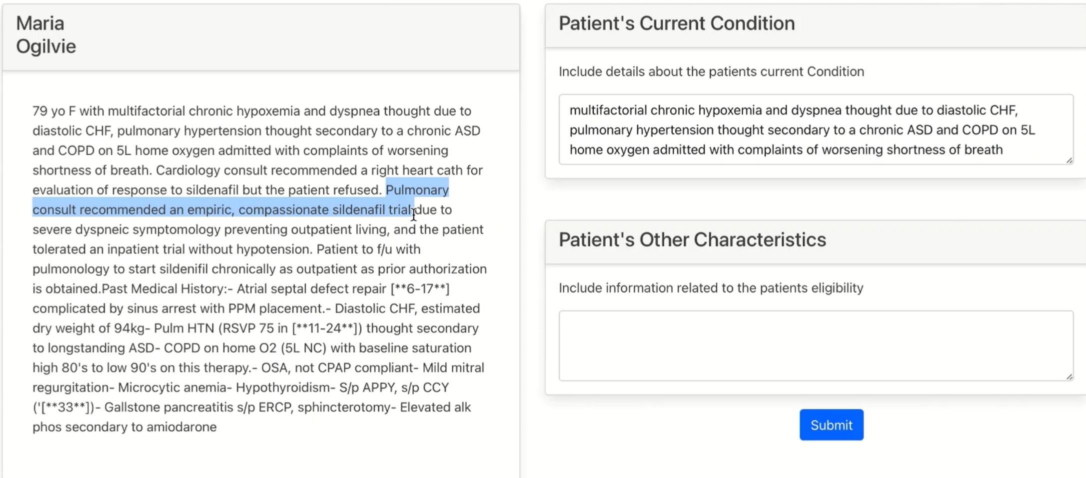

Figure 8.5: Page that allows the healthcare provider to create two distinct queries.

In future version, the system will employ color-coding to emphasize different categories of patient information, thereby enhancing ease of identification, as shown in Figure 6.2. Additionally, this interface allows for the formulation of specific queries, which will be elaborated upon in the subsequent section.

8.2.2 Phase 2: Query Creation

In the interface presented in Figure 8.5, healthcare providers can view and extract selected portions of a patient's EHR that they consider relevant for identifying appropriate clinical trials. At this stage, the user has dual goals: initially, to identify the patient's primary medical condition and generate the corresponding query; subsequently, to identify characteristics that contribute to evaluating the patient's eligibility for a clinical trial. The initial query activates the task-specific retrieval algorithm—specifically, the DtMRF$_{\text{TOPSIS}}$— to identify clinical trials in line with the patient's medical condition. In parallel, the secondary query functions to filter out trials for which the patient is unlikely to be eligible. The system utilizes the BM25 information retrieval model to generate these relevance signals, aiming for a balance of simplicity, robustness, and rapid response time. In forthcoming iterations, the system will incorporate Neural-DtMRF$_{\text{TOPSIS}}$, given its demonstrated superior performance in relevant retrieval tasks. Additionally, the process of relevance estimation will incorporate other factors, including the geographical locations of both the patient and the clinical trials. Furthermore, enhancements in this phase will be informed by our research outcomes discussed in Chapter 6. These improvements will include pre-populating fields with pertinent information for the user, thereby streamlining the process.

8.2.3 Phase 3: Result Presentation and Examination

In the final step, users are presented with the retrieved clinical trials, as depicted in Figure 8.6. Within this interface, users can assess the relevance of each clinical trial, primarily based on its title. Additionally, they can delve deeper by examining the corresponding inclusion and exclusion criteria, which become visible upon clicking on a specific trial.

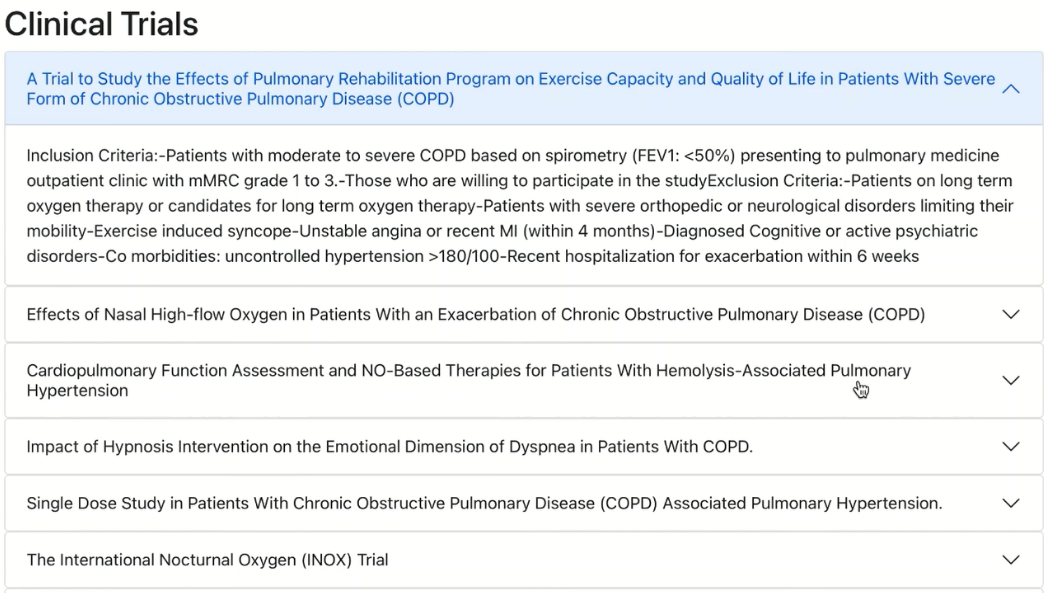

Figure 8.6: Presentation of retrieved clinical trials, along with their corresponding inclusion and exclusion criteria.

It should be noted that while additional details about the trial, including its description and associated MeSH terms, could be displayed to the user, their utility in determining the patient's eligibility for the trial is limited. Given the potential for these details to negatively affect the user experience, a conscious decision was made to exclude these elements from the final display of results in the current prototype. The interface permits the simultaneous opening of multiple trials, thereby enabling the user to conduct comparative evaluations. In future research, we plan to explore the adoption of a card-based presentation format for displaying results, as an alternative to the traditional layout.

8.3 Conclusions and Directions for Future Improvements

The prototype has been carefully designed to accommodate a range of task-specific prerequisites, with the overarching objective of refining the search process for eligibility screening and the retrieval of clinical trials. The emergence of large language models and their advanced text-processing capabilities enable the automation of information extraction and query formulation. As the preliminary findings in

Section 7.4 indicate, LLMs can assess a patient's eligibility for a subset of retrieved clinical trials (subset due to their complexity and cost). This development offers additional avenues for research and enhancements that we intend to explore in future work.

Part V Overall Insights

Chapter 9

Conclusions and Directions for Future Research

This chapter summarizes the key findings, contributions, and implications of the research conducted in this dissertation. It commences by revisiting the research contributions outlined in the introduction and discussing their primary outcomes and results. It then highlights open challenges and directions for future research related to multidimensional relevance estimation in IR and the task of clinical trials retrieval.

9.1 Overview of our Contributions and Results

In this dissertation, our primary research focus was on multidimensional relevance estimation in the field of Information Retrieval, leading to the contribution of a novel framework. Our secondary focus explored the complex task of clinical trials retrieval, where we concentrated on three distinct sub-topics: information extraction from clinical narratives, enhancing its retrieval effectiveness, and developing a specialized search prototype. These focal points guided our research objectives and led to the essential contributions we summarize in this chapter.

Multidimensional relevance estimation in IR. The systematic literature review presented in Chapter 3 sought to understand the multidimensional nature of relevance in the field of IR. Our examination of 70 studies showed an evolving collaborative effort between academia and industry. Our analysis identified 18 unique knowledge domains, each encompassing multiple search tasks, where researchers have advanced multidimensional relevance models. Moreover, we identified and grouped more than 40 relevance factors, some of which were consistently applied across the various domains. However, we noted considerable inconsistencies in both their definitions and how they were operationalized. We developed a structured framework for categorizing relevance factors to address this inconsistency. We classified existing multidimensional relevance estimation approaches into model-driven and data-driven. Most data-driven models in the literature primarily utilize learning to rank techniques. Model-driven strategies are anchored in formal mathematical models. Notably, despite the presence of alternative methods like copulas and Multi-Criteria Decision-Making, most studies continue to rely on simple linear combinations for final relevance estimation. Concluding our analysis, we identified a rising need for benchmark collections annotated with various relevance factors across multiple domains. Our study also emphasized the potential utility of LLMs in facilitating the development of such benchmarks.

Decision theoretic Multidimensional Relevance Framework (DtMRF) and Neural-DtMRF. Chapter 5 lays out our seminal contribution: the DtMRF and its neural extension, Neural-DtMRF. We began by conceptualizing multidimensional relevance estimation in IR as a decision-making process, introducing a formal framework to underpin this perspective. Following that, we performed a comprehensive investigation into the usage of scoring-based and distance-based MADM methods in IR, showcasing how these methods can be employed for document ranking. We used a retrieval simulation to reveal the limitations of commonly used aggregation approaches like linear combination and weighted sum, mainly when a relevance factor is binary. Also, our analysis elucidated how DtMRF leads to interpretable document rankings. Furthermore, we detailed the components necessary to integrate neural models into DtMRF to predict the relevance factors' importance for a given information need. The chapter presented our primary contributions, i.e. a novel retrieval framework applicable to various search tasks, offering advantages in retrieval explainability and enabling user control over the search process.

Information extraction from clinical narratives. Chapter 6 examines the methodologies we have employed for information extraction from clinical narratives in Electronic Health Records, explicitly targeting enhancing clinical trials retrieval. We initiated our investigation by assessing multiple state-of-the-art approaches to information extraction, aiming to improve retrieval effectiveness. Our findings indicated that a well-structured query enhanced with proper medical entities is needed to improve performance. Also, our analysis revealed that transformer-based models fine-tuned on domain-specific datasets outperform traditional rule-based methods for negation detection in this task. We also employed LLMs, particularly GPT-3.5, which exceeded current state-of-the-art techniques, showing an improvement of 8.25% and 9.93% in nDCG@10 across two distinct collections, and demonstrated superior performance to medical experts in specific scenarios. Lastly, we underscored the cost-efficiency and reduced complexity of our GPT-3.5 implementation as distinguishing factors compared to existing state-of-the-art methods. However, we also identified challenges associated with data security concerns while providing potential solutions.

Relevance Estimation in clinical trials retrieval. Chapter 7 provided a comprehensive assessment of the experimental results related to the use of the Decision-theoretic Multidimensional Relevance Framework and its neural extension in the context of clinical trials retrieval. The chapter also presented empirical findings from a retrieval methodology integrating LLMs with Neural-DtMRF to evaluate patient eligibility for clinical trials. We initially showed that compliance with task-specific requirements and evaluating both positive and negative relevance factors led to enhanced retrieval performance. We introduced a re-ranking strategy that utilized additional relevance signals and improved the achieved retrieval performance. Remarkably, these improvements were achieved using the BM25 model for topical relevance in single-step retrieval, matching the performance of more complex neural ranking models like BERT. Although we highlighted the potential for further optimization, the final integration of a Large Language Model for eligibility estimation allowed us to surpass the state-of-the-art approaches in some performance metrics. In conclusion, the deployment of DtMRF and Neural-DtMRF not only improved ranking interpretability but also laid the foundation for future refinements.

Search prototype for clinical trials retrieval and eligibility screening. Chapter 8 introduced the initial version of a search prototype tailored for clinical trials retrieval, explicitly emphasizing the eligibility screening process. The prototype can meet various requirements that arise during different stages of the screening process, considering technical complexity requirements and practical applicability. The system's adaptability indicates its potential to serve as a foundation for more streamlined, user-centric interfaces to aid researchers and healthcare professionals. In conclusion, this prototype serves as a preliminary yet promising solution in the evolving field of clinical trials retrieval, exhibiting the flexibility needed to adapt to future challenges.

9.2 Directions for Future Research

In this section, we highlight potential areas for further research inspired by the results and limitations encountered in this dissertation. The first set of directions involves advancing the Decision-theoretic Multidimensional Relevance Framework (DtMRF) and deepening our understanding of information extraction from clinical narratives. The second set aims to optimize performance in clinical trial retrieval tasks and refine the search system for clinical trial retrieval and patient eligibility assessment.

Decision-theoretic Multidimensional Relevance Framework. In future work focused on the Decision-theoretic Multidimensional Relevance Framework, two primary areas of exploration are identified. Expand the application scope of DtMRF and Neural-DtMRF across a broader range of domains and search tasks. In this regard, leveraging large language models could be particularly beneficial for generating appropriate annotations concerning additional relevance factors. This concept was elaborated upon in Chapter 3. Secondly, augmenting the existing

Multi-Attribute Decision Making methods within the framework is another future direction. Specifically, the focus will be on integrating algorithms that match the current complexity level of the framework while simultaneously allowing the incorporation of both positive and negative relevance factors.

Information Extraction from Clinical Narratives. In the domain of information extraction from clinical narratives, a key avenue for future research is to explore the efficacy of open-source large language models in achieving performance metrics comparable to those garnered in existing experiments. This line of inquiry will help determine whether open-source options can be adequate substitutes for the models we exploited in our experimentation.

Clinical Trials Retrieval. In clinical trials retrieval, numerous potentials for future research can be identified. One is related to creating another dataset for optimal weight prediction to be used with the Neural-DtMRF. In this case, we aim to discern queries where applying the Decision-theoretic Multidimensional Relevance Framework may not be advantageous. To that aim, consideration is given to introducing zero or negative weights for queries where the $\text{DtMRF}_{\text{TOPSIS}}$ retrieval performance falls below an established baseline. Additionally, current experiments utilize the same patient-related information (i.e. same queries) to estimate the relevance of a clinical trial to patients' condition and their eligibility. Given that DtMRF allows for distinct query representations, future work will explore this capability. Further advancements could include incorporating additional relevance factors, such as patient demographics or trial-specific attributes, to align the retrieval process more closely with real user needs in this task. Lastly, employing a smaller model through knowledge distillation, drawing upon insights from GPT-3.5, is considered for assessing patient eligibility for clinical trials.

Search System for Clinical Trials Retrieval and Patients' Eligibility Screening. In the context of the proposed search system for clinical trials retrieval and patients' eligibility screening, the next step involves advancing the system to its second version. Upon completing this upgrade, we aim to conduct user-based evaluations on the updated prototype to validate its functional efficacy.

Bibliography

M. Agosti, G. M. D. Nunzio, and S. Marchesin. An analysis of query reformulation techniques for precision medicine. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019*, pages 973–976. ACM, 2019. doi: 10.1145/3331184.3331289.

M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. A. Sontag. Large language models are few-shot clinical information extractors. In *EMNLP*, pages 1998–2022. Association for Computational Linguistics, 2022.

T. Ahmed, M. M. A. Aziz, and N. Mohammed. De-identification of electronic health record using neural network. *Scientific reports*, 10(1):1–11, 2020.

D. Akestoridis and E. Papapetrou. A framework for the evaluation of routing protocols in opportunistic networks. *Computer Communications*, 145:14–28, 2019. doi: 10.1016/j.comcom.2019.06.003.

J. Albrecht, E. Kitanidis, and A. J. Fetterman. Despite "super-human" performance, current llms are unsuited for decisions about ethics and safety, 2022.

A. Alinezhad and A. Amini. Sensitivity analysis of topsis technique: the results of change in the weight of one attribute on the final ranking of alternatives. *Journal of optimization in industrial engineering*, 4(7):23–28, 2011.

A. Alinezhad and J. Khalili. *New methods and applications in multiple attribute decision making (MADM)*, volume 277. Springer, 2019. doi: 10.1007/978-3-030-15009-9.

E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909.

M. Alsulmi and B. Carterette. Improving medical search tasks using learning to rank. In *2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–8, 2018. doi: 10.1109/CIBCB.2018.8404965.

L. Amancio, C. F. Dorneles, and D. H. Dalip. Recency and quality-based ranking question in cqas: A stack overflow case study. *Information Processing and Management*, 2021. doi: 10.1016/j.ipm.2021.102552.

G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20 (4):357–389, 2002. doi: 10.1145/582415.582416.

A. Arampatzis, G. Peikos, and S. Symeonidis. Pseudo relevance feedback optimization. *Information Retrieval Journal*, 24(4-5):269–297, 2021.

S. Arastoopoor. Domain-specific readability measures to improve information retrieval in the persian language. *The Electronic Library*, 2018. doi: 10.1108/el-01-2017-0007.

M. Aruldoss, T. M. Lakshmi, and V. P. Venkatesan. A survey on multi criteria decision making methods and its applications. *American Journal of Information Systems*, 1(1):31–43, 2013. doi: 10.12691/ajis-1-1-5.

E. Ashoori and M. Lalmas. Using topic shifts for focussed access to xml repositories. In G. Amati, C. Carpineto, and G. Romano, editors, *Advances in Information Retrieval*, pages 444–455, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-71496-5.

A. Askari, G. Peikos, G. Pasi, and S. Verberne. Leibi@coliee 2022: Aggregating tuned lexical models with a cluster-driven bert-based model for case law retrieval. *CoRR*, abs/2205.13351, 2022.

I. Badache and M. Boughanem. Social priors to estimate relevance of a resource. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, page 106–114, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329767. doi: 10.1145/2637002.2637016.

A. Balagopalan, A. Z. Jacobs, and A. J. Biega. The role of relevance in fair ranking. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2650–2660, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591933.

M. Bambia and R. Faiz. Frel: A freshness language model for optimizing real-time web search. In R. Silhavy, R. Senkerik, Z. K. Oplatkova, Z. Prokopova, and P. Silhavy, editors, *Intelligent Systems in Cybernetics and Automation Theory*, pages 207–216, Cham, 2015. Springer International Publishing. ISBN 978-3-319-18503-3.

Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023, 2023.

C. L. Barry and L. Schamber. Users' criteria for relevance evaluation: A cross-situational comparison. *Inf. Process. Manag.*, 34(2-3):219–236, 1998. doi: 10.1016/S0306-4573(97)00078-2.

E. Bassani and G. Pasi. A multi-representation re-ranking model for personalized product search. *Information Fusion*, 2021. doi: 10.1016/j.inffus.2021.11.010.

M. Behzadian, S. K. Otaghsara, M. Yazdani, and J. Ignatius. A state-of the-art survey of TOPSIS applications. *Expert Systems with Applications*, 39(17): 13051–13069, 2012. doi: 10.1016/j.eswa.2012.05.056.

N. J. Belkin. People, interacting with information1. *SIGIR Forum*, 49(2):13–27, jan 2016a. ISSN 0163-5840. doi: 10.1145/2888422.2888424.

N. J. Belkin. People, interacting with information1. In *ACM SIGIR Forum*, volume 49, pages 13–27. ACM New York, NY, USA, 2016b.

N. J. Belkin, M. Cole, and J. Liu. A model for evaluation of interactive information retrieval. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 7–8, 2009.

I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods* *in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371.

M. Bendersky, W. B. Croft, and Y. Diao. Quality-biased ranking of web documents. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, page 95–104, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450304931. doi: 10.1145/1935826.1935849.

J. Berger. *Statistical decision theory: foundations, concepts, and methods*. Springer Science & Business Media, 2013.

O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004. doi: 10.1093/nar/gkh061.

L. H. Bonifacio, H. Abonizio, M. Fadaee, and R. F. Nogueira. Inpars: Unsupervised dataset generation for information retrieval. In E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, and G. Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2387–2392. ACM, 2022. doi: 10.1145/3477495.3531863.

H. Borchani, G. Varando, C. Bielza, and P. Larranaga. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(5):216–233, 2015.

P. Borlund. The iir evaluation model: a framework for evaluation of interactive information retrieval systems. *Information research*, 8(3):8–3, 2003.

O. Bouidghaghen, L. Tamine-Lechani, G. Pasi, G. Cabanac, M. Boughanem, and C. da Costa Pereira. Prioritized aggregation of multiple context dimensions in mobile ir. In M. V. M. Salem, K. Shaalan, F. Oroumchian, A. Shakery, and H. Khelalfa, editors, *Information Retrieval Technology*, pages 169–180, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-25631-8.

J.-P. Brans, P. Vincke, and B. Mareschal. How to select and how to rank projects: The promethee method. *European journal of operational research*, 24(2):228–238, 1986. doi: 10.1016/0377-2217(86)90044-5.

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, apr 1998. doi: 10.1016/s0169-7552(98)00110-x.

M. Brøgger-Mikkelsen, Z. Ali, J. R. Zibert, A. D. Andersen, and S. F. Thomsen. Online patient recruitment in clinical trials: Systematic review and meta-analysis. *J Med Internet Res*, 22(11):e22179, Nov 2020. ISSN 1438-8871. doi: 10.2196/22179.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *SIGIR*, pages 25–32. ACM, 2004.

T. Calders and D. Van Assche. Promethee is not quadratic: An o(qnlog(n)) algorithm. *Omega*, 76:63–69, 2018. ISSN 0305-0483. doi: 10.1016/j.omega.2017.04.003.

D. Carmel, E. Haramaty, A. Lazerson, and L. Lewin-Eytan. Multi-objective ranking optimization for product search using stochastic label aggregation. In *Proceedings of The Web Conference 2020*, WWW '20, page 373–383, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380122.

M. Cascella, J. Montomoli, V. Bellini, and E. Bignami. Evaluating the feasibility of chatgpt in healthcare: An analysis of multiple clinical and research scenarios. *J. Medical Syst.*, 47(1):33, 2023.

T. Chai and R. R. Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3):1247–1250, 2014.

W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. Evaluation of negation phrases in narrative clinical reports. In *AMIA 2001,* *American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 3-7, 2001*. AMIA, 2001a.

W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001b.

J. M. Chenlo, J. Parapar, D. E. Losada, and J. Santos. Finding a needle in the blogosphere: An information fusion approach for blog distillation search. *Information Fusion*, 23:58–68, 2015. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2014.09.001.

R. Chow, J. Midroni, J. Kaur, G. Boldt, G. Liu, L. Eng, F.-F. Liu, B. Haibe-Kains, M. Lock, and S. Raman. Use of artificial intelligence for cancer clinical trial enrollment: a systematic review and meta-analysis. *JNCI: Journal of the National Cancer Institute*, 115(4):365–374, 2023.

A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022.

H. Chu. Factors affecting relevance judgment: a report from trec legal track. *Journal of Documentation*, 67(2):264–278, 2011.

Y.-N. Chuang, R. Tang, X. Jiang, and X. Hu. Spec: A soft prompt-based calibration on mitigating performance variability in clinical notes summarization. *preprint arXiv:2303.13035*, 2023.

V. Claveau. Neural text generation for query expansion in information retrieval. In J. He, R. Unland, E. S. Jr., X. Tao, H. Purohit, W. van den Heuvel, J. Yearwood, and J. Cao, editors, *WI-IAT '21: IEEE/WIC/ACM International Conference* *on Web Intelligence, Melbourne VIC Australia, December 14 - 17, 2021*, pages 202–209. ACM, 2021. doi: 10.1145/3486622.3493957.

C. Cleverdon. Evaluation tests of information retrieval systems. *Journal of Documentation*, 26(1):55–67, 1970.

K. Cobbe, V. Kosaraju, M. Bavarian, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021.

K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag. Personalizing web search results by reading level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 403–412, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450307178. doi: 10.1145/2063576.2063639.

C. Cool, N. Belkin, O. Frieder, and P. Kantor. Characteristics of text affecting relevance judgments. In *National online meeting*, volume 14, pages 77–77. LEARNED INFORMATION (EUROPE) LTD, 1993.

H. Cooper, L. V. Hedges, and J. C. Valentine. *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation, 2019. ISBN 9780871540058. doi: /10.7758/9781610448864.

W. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1):19–37, 1971.

E. Cosijn. Relevance judgments and measurements. In *Encyclopedia of library and information sciences*, pages 4512–4519. CRC Press, 2009.

E. Cosijn and P. Ingwersen. Dimensions of relevance. *Information Processing & Management*, 36(4):533–550, 2000. ISSN 0306-4573. doi: https://doi.org/10.1016/S0306-4573(99)00072-2.

N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, page 416–423, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595930345. doi: 10.1145/1076034.1076106.

F. Crestani, S. Mizzaro, and I. Scagnetto. *Mobile information retrieval.* Springer, 2017.

A. Creswell, M. Shanahan, and I. Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. *CoRR*, abs/2205.09712, 2022.

C. da Costa Pereira, M. Dragoni, and G. Pasi. Multidimensional relevance: A new aggregation criterion. In M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy, editors, *Advances in Information Retrieval*, pages 264–275, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-00958-7.

C. da Costa Pereira, M. Dragoni, and G. Pasi. Multidimensional relevance: Prioritized aggregation in a personalized information retrieval setting. *Information Processing & Management*, 48(2):340–357, 2012a. doi: 10.1016/j.ipm.2011.07.001.

C. da Costa Pereira, M. Dragoni, and G. Pasi. Multidimensional Relevance: Prioritized Aggregation in a Personalized Information Retrieval Setting. *Information Processing and Management*, 48(2):340–357, 2012b. doi: 10.1016/j.ipm.2011.07.001.

N. Dai, M. Shokouhi, and B. D. Davison. Learning to rank for freshness and relevance. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 95–104, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450307574. doi: 10.1145/2009916.2009933.

M. Daoud, M. Daoud, and J. X. Huang. Modeling geographic, temporal, and proximity contexts for improving geotemporal search. *Journal of the Association for Information Science and Technology*, 2013. doi: 10.1002/asi.22648.

H. Deng, C. Yeh, and R. J. Willis. Inter-company comparison using modified TOPSIS with objective weights. *Computers & Operations Research*, 27(10): 963–973, 2000. doi: 10.1016/S0305-0548(99)00069-6.

A. Deshpande, V. Murahari, T. Rajpurohit, A. Kalyan, and K. Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. *CoRR*, abs/2304.05335, 2023.

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *preprint arXiv:2305.14314*, 2023.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019.

D. Diakoulaki, G. Mavrotas, and L. Papayannakis. Determining objective weights in multiple criteria problems: The critic method. *Comput. Oper. Res.*, 22(7): 763–770, 1995. doi: 10.1016/0305-0548(94)00059-H.

A. Dumitrescu and S. Santini. Full coverage of a reader's interests in context-based information filtering. *Journal of the Association for Information Science and Technology*, 72(8):1011–1027, 2021.

V. S. Effoe, J. A. Katula, J. K. Kirk, C. F. Pedley, L. Y. Bollhalter, W. M. Brown, M. R. Savoca, S. T. Jones, J. Baek, A. G. Bertoni, et al. The use of electronic medical records for recruitment in clinical trials: findings from the lifestyle intervention for treatment of diabetes trial. *Trials*, 17:1–10, 2016.

C. Eickhoff and A. P. de Vries. Modelling complex relevance spaces with copulas. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, page 1831–1834, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450325981. doi: 10.1145/2661829.2661925.

C. Eickhoff, A. P. de Vries, and K. Collins-Thompson. Copulas for information retrieval. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, page 663–672, New York, NY, USA, 2013a. Association for Computing Machinery. ISBN 9781450320344. doi: 10.1145/2484028.2484066.

C. Eickhoff, A. P. de Vries, and K. Collins-Thompson. Copulas for information retrieval. In *The 36th International ACM SIGIR conference on research and development in Information Retrieval*, pages 663–672. ACM, 2013b. doi: 10.1145/2484028.2484066.

H. Eyre, A. B. Chapman, K. S. Peterson, J. Shi, P. R. Alba, M. M. Jones, T. L. Box, S. L. DuVall, and O. V. Patterson. Launching into clinical space with medspacy: a new clinical text processing toolkit in python. In *AMIA Annual Symposium Proceedings*, volume 2021, page 438. American Medical Informatics Association, 2021.

L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill. A bibliometric review of large language models research from 2017 to 2023. *preprint arXiv:2304.02020*, 2023.

M. Farah and D. Vanderpooten. An outranking approach for information retrieval. *Information Retrieval*, 2008. doi: 10.1007/s10791-008-9046-z.

J. Feng, H. Li, M. Huang, S. Liu, W. Ou, Z. Wang, and X. Zhu. Learning to collaborate: Multi-scenario ranking via multi-agent reinforcement learning. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1939–1948, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356398. doi: 10.1145/3178876.3186165.

M. Fernández-Pichel, D. E. Losada, and J. C. Pichel. A multistage retrieval system for health-related misinformation detection. *Engineering Applications of Artificial Intelligence*, 115:105211, oct 2022. doi: 10.1016/j.engappai.2022.105211.

T. R. Fleming and D. L. DeMets. Surrogate end points in clinical trials: are we being misled? *Annals of internal medicine*, 125(7):605–613, 1996.

S. Gerani, C. Zhai, and F. Crestani. Score transformation in linear combination for multi-criteria relevance ranking. In R. Baeza-Yates, A. P. de Vries, H. Zaragoza, B. B. Cambazoglu, V. Murdock, R. Lempel, and F. Silvestri, editors, *Advances in Information Retrieval*, pages 256–267, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-28997-2.

W. Goffman and V. A. Newill. A methodology for test and evaluation of information retrieval systems. *Information Storage and Retrieval*, 3(1):19–25, 1966. ISSN 0020-0271. doi: https://doi.org/10.1016/0020-0271(66)90006-4.

M. D. Grandis, G. Pasi, and M. Viviani. Fake news detection in microblogging through quantifier-guided aggregation. In *Modeling Decisions for Artificial Intelligence - 16th International Conference*, volume 11676 of *Lecture Notes in Computer Science*, pages 64–76. Springer, 2019. doi: 10.1007/978-3-030-26773-5_6.

M. Grootendorst. Keybert: Minimal keyword extraction with bert., 2020.

R. B. Gul and P. A. Ali. Clinical trials: the challenge of recruitment and retention of participants. *Journal of clinical nursing*, 19(1-2):227–233, 2010.

B. J. Gutierrez, N. McNeal, C. Washington, Y. Chen, L. Li, H. Sun, and Y. Su. Thinking about GPT-3 in-context learning for biomedical ie? think again. In *EMNLP (Findings)*, pages 4497–4512. Association for Computational Linguistics, 2022.

U. Hahn and M. Oleynik. Medical information extraction in the age of deep learning. *Yearbook of medical informatics*, 29(01):208–220, 2020.

H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42(5):839–851, 2009.

S. Harrer. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *Ebiomedicine*, 90, 2023.

S. P. Harter. A probabilistic approach to automatic keyword indexing. part i. on the distribution of specialty words in a technical literature. *J. Am. Soc. Inf. Sci.*, 26(4):197–206, 1975a. doi: 10.1002/asi.4630260402.

S. P. Harter. A probabilistic approach to automatic keyword indexing. part II. an algorithm for probabilistic indexing. *J. Am. Soc. Inf. Sci.*, 26(5):280–289, 1975b. doi: 10.1002/asi.4630260504.

W. Hersh, W. Hersh, and Weston. *Information retrieval: a biomedical and health perspective*. Springer, 2020.

W. R. Hersh. Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. *Clin Pharmacol Ther*, 81:126–128, 2007.

S. Himani and D. Vaidehi. A survey on medical information retrieval. In S. C. Satapathy and A. Joshi, editors, *Information and Communication Technology for Intelligent Systems (ICTIS 2017) - Volume 1*, pages 543–550, Cham, 2018. Springer International Publishing. ISBN 978-3-319-63673-3.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *preprint arXiv:1503.02531*, 2015.

M. Hobensack, J. Song, D. Scharp, K. H. Bowles, and M. Topaz. Machine learning applied to electronic health record data in home healthcare: A scoping review. *Int. J. Medical Informatics*, 170:104978, 2023.

Y. Hu, I. Ameer, X. Zuo, X. Peng, Y. Zhou, Z. Li, Y. Li, J. Li, X. Jiang, and H. Xu. Zero-shot clinical entity recognition using chatgpt. *preprint arXiv:2303.16416*, 2023.

J. X. Huang, B. He, and J. Zhao. Mining authoritative and topical evidence from the blogosphere for improving opinion retrieval. *Information Systems*, 78: 199–213, nov 2018. doi: 10.1016/j.is.2018.02.002.

C. Hwang and K. Yoon. *Multiple Attribute Decision Making: Methods and Applications - A State-of-the-Art Survey*, volume 186 of *Lecture Notes in Economics and Mathematical Systems*. Springer, 1981. ISBN 978-3-540-10558-9. doi: 10.1007/978-3-642-48318-9.

B. R. S. Idnay, C. N. Dreisbach, C. Weng, and R. Schnall. A systematic review on natural language processing systems for eligibility prescreening in clinical research. *J. Am. Medical Informatics Assoc.*, 29(1):197–206, 2021.

P. Ingwersen and K. Järvelin. *The turn: Integration of information seeking and retrieval in context*, volume 18. Springer Science & Business Media, 2005.

L. B. Jabeur, L. Tamine, and M. Boughanem. Featured tweet search: Modeling time and social influence for microblog retrieval. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 166–173, 2012. doi: 10.1109/WI-IAT.2012.233.

B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713, 2018.

N. M. Jain, A. Culley, T. Knoop, C. Micheel, T. Osterman, and M. Levy. Conceptual framework to support clinical trial optimization and end-to-end enrollment workflow. *JCO Clinical Cancer Informatics*, 3:1–10, 2019.

N. A. Jaleel, J. Allan, W. B. Croft, F. Diaz, L. S. Larkey, X. Li, M. D. Smucker, and C. Wade. Umass at TREC 2004: Novelty and HARD. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*, volume 500-261 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2004.

J. Jang, S. Ye, and M. Seo. Can large language models truly understand prompts? A case study with negated prompts. *CoRR*, abs/2209.12711, 2022.

M. Jang and T. Lukasiewicz. Consistency analysis of chatgpt. *CoRR*, abs/2303.06273, 2023.

Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J. Am. Medical Informatics Assoc.*, 18(5):601–606, 2011.

H. Joho. Diane kelly: Methods for evaluating interactive information retrieval systems with users - foundation and trends in information retrieval, vol 3, nos 1-2, pp 1-224, 2009, ISBN: 978-1-60198-224-7. *Inf. Retr.*, 14(2):204–207, 2011.

P. Jomsri and D. Prangchumpol. A hybrid model ranking search result for research paper searching on social bookmarking. In *2015 1st International Conference on Industrial Networks and Intelligent Systems (INISCom)*, pages 38–43, 2015. doi: 10.4108/icst.iniscom.2015.258417.

C. Kang, X. Wang, Y. Chang, and B. Tseng. Learning to rank with multi-aspect relevance for vertical search. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, page 453–462, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450307475. doi: 10.1145/2124295.2124350.

S. K. Karmaker Santu, P. Sondhi, and C. Zhai. On application of learning to rank for e-commerce search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 475–484, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350228. doi: 10.1145/3077136.3080838.

N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858, 2019.

G. R. Klare. The measurement of readability: useful information for communicators. *ACM Journal of Computer Documentation (JCD)*, 24(3):107–121, 2000.

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022.

T. Komatsuda, A. Keyaki, and J. Miyazaki. A score fusion method using a mixture copula. In S. Hartmann and H. Ma, editors, *Database and Expert Systems Applications*, pages 216–232, Cham, 2016. Springer International Publishing. ISBN 978-3-319-44406-2.

B. Koopman and G. Zuccon. Understanding negation and family history to improve clinical information retrieval. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 971–974. ACM, 2014. doi: 10.1145/2600428.2609487.

B. Koopman and G. Zuccon. A test collection for matching patients to clinical trials. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 669–672. ACM, 2016. doi: 10.1145/2911451.2914672.

B. Koopman and G. Zuccon. Cohort-based clinical trial retrieval. In *ADCS '21: Australasian Document Computing Symposium*, pages 3:1–3:9. ACM, 2021. doi: 10.1145/3503516.3503529.

K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. F. Jones, R. Forshee, M. Walderhaug, and T. Botsis. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J. Biomed. Informatics*, 73:14–29, 2017.

M. Kritz, M. Gschwandtner, V. Stefanov, A. Hanbury, and M. Samwald. Utilization and perceived problems of online medical resources and search tools among different groups of european physicians. *J Med Internet Res*, 15(6):e122, Jun 2013. ISSN 14388871.

W. Kusa, G. Peikos, O. Espitia, A. Hanbury, and G. Pasi. Dossier at medvidqa 2022: Text-based approaches to medical video answer localization problem. In *BioNLP@ACL*, pages 432–440. Association for Computational Linguistics, 2022.

M. Y. Landolsi, L. Hlaoua, and L. B. Romdhane. Information extraction from electronic medical documents: state of the art and future research directions. *Knowl. Inf. Syst.*, 65(2):463–516, 2023.

V. Lavrenko and W. B. Croft. Relevance-based language models. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors, *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages 120–127. ACM, 2001. doi: 10.1145/383952.383972.

R. Leaman, R. Islamaj, and Z. Lu. The overview of the nlm-chem biocreative vii track: full-text chemical identification and indexing in pubmed articles. In *BioCreative VII Challenge Evaluation Workshop*, pages 108–113, 2021.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. Ho So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020a. doi: https://doi.org/10.1093/bioinformatics/btz682.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240, 2020b. doi: 10.1093/bioinformatics/btz682.

W. Lee, M. Chun, H. Jeong, and H. Jung. Toward keyword generation through large language models. In *IUI Companion*, pages 37–40. ACM, 2023.

E. Lehman, E. Hernandez, D. Mahajan, J. Wulff, M. J. Smith, Z. Ziegler, D. Nadler, P. Szolovits, A. E. W. Johnson, and E. Alsentzer. Do we still need clinical language models? *CoRR*, abs/2302.08091, 2023.

B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. In M. Moens, X. Huang, L. Specia, and S. W. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3045–3059. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.243.

P. S. H. Lewis, M. Ott, J. Du, and V. Stoyanov. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *ClinicalNLP@EMNLP*, pages 146–157. Association for Computational Linguistics, 2020.

J. Li, P. Zhang, D. Song, and Y. Wu. Understanding an enriched multidimensional user relevance model by analyzing query logs. *J. Assoc. Inf. Sci. Technol.*, 68 (12):2743–2754, 2017a. doi: 10.1002/asi.23868.

J. Li, P. Zhang, D. Song, and Y. Wu. Understanding an enriched multidimensional user relevance model by analyzing query logs. *Journal of the Association for Information Science and Technology*, 2017b. doi: 10.1002/asi.23868.

X. Li, J. Mao, W. Ma, Y. Liu, M. Zhang, S. Ma, Z. Wang, and X. He. Topic-enhanced knowledge-aware retrieval model for diverse relevance estimation. In *Proceedings of the Web Conference 2021*, WWW '21, page 756–767, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449943.

N. Limsopatham, C. Macdonald, and I. Ounis. Modelling relevance towards multiple inclusion criteria when ranking patients. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1639–1648. ACM, 2014. doi: 10.1145/2661829.2661958.

N. B. Link, S. Huang, T. Cai, J. Sun, K. Dahal, L. Costa, K. Cho, K. P. Liao, T. Cai, and C. Hong. Binary acronym disambiguation in clinical notes from electronic health records with an application in computational phenotyping. *Int. J. Medical Informatics*, 162:104753, 2022.

N. Linna and C. E. K. Jr. Applications of natural language processing in radiology: A systematic review. *Int. J. Medical Informatics*, 163:104779, 2022.

C. Lioma, B. Larsen, W. Lu, and Y. Huang. A study of factuality, objectivity and relevance: Three desiderata in large-scale information retrieval? In *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, BDCAT '16, page 107–117, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450346177. doi: 10.1145/3006299.3006315.

C. Liu, C. Yuan, A. M. Butler, R. D. Carvajal, Z. R. Li, C. N. Ta, and C. Weng. Dquest: dynamic questionnaire for search of clinical trials. *J. Am. Medical Informatics Assoc.*, 26(11):1333–1343, 2019a. doi: 10.1093/jamia/ocz121.

J. Liu. Deconstructing search tasks in interactive information retrieval: A systematic review of task dimensions and predictors. *Information Processing & Management*, 58(3):102522, 2021. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102522.

J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen. What makes good in-context examples for gpt-3? In E. Agirre, M. Apidianaki, and I. Vulic, editors, *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, May 27, 2022*, pages 100–114. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.deelio-1.10.

P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35, 2023a.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019b.

Z. Liu, X. Yu, L. Zhang, Z. Wu, C. Cao, H. Dai, L. Zhao, W. Liu, D. Shen, Q. Li, T. Liu, D. Zhu, and X. Li. Deid-gpt: Zero-shot medical text de-identification by GPT-4. *CoRR*, abs/2303.11032, 2023b.

Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556.

R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T. Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings Bioinform.*, 23(6), 2022.

M. Lupu, M. Salampasis, and A. Hanbury. Domain specific search. In G. Paltoglou, F. Loizides, and P. Hansen, editors, *Professional Search in the Modern World - COST Action IC1002 on Multilingual and Multifaceted Interactive Information Access*, volume 8830 of *Lecture Notes in Computer Science*, pages 96–117. Springer, 2014.

Y. Ma, Y. Wu, Q. Ai, Y. Liu, Y. Shao, M. Zhang, and S. Ma. Incorporating structural information into legal case retrieval. *ACM Transactions on Information Systems*, jul 2023. doi: 10.1145/3609796.

S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, and N. Goharian. Simplified data wrangling with ir_datasets. In *SIGIR*, 2021.

K. R. MacCrimmon. Decisionmaking among multiple-attribute alternatives: a survey and consolidated approach. Technical report, Rand Corp Santa Monica Ca, Santa Monica, CA, 1968.

C. Macdonald and N. Tonellotto. Declarative experimentation ininformation retrieval using pyterrier. In *Proceedings of ICTIR 2020*, 2020.

C. Macdonald, N. Tonellotto, S. MacAvaney, and I. Ounis. Pyterrier: Declarative experimentation in python from bm25 to dense retrieval. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4526–4533, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482013.

S. Madisetty and M. S. Desarkar. A reranking-based tweet retrieval approach for planned events. *World Wide Web*, 25(1):23–47, 2022. doi: 10.1007/s11280-021-00962-8.

T. Magoc, K. S. Allen, C. McDonnell, J.-P. Russo, J. Cummins, J. R. Vest, and C. A. Harle. Generalizability and portability of natural language processing system to extract individual social risk factors. *International Journal of Medical Informatics*, 177:105115, 2023. ISSN 1386-5056. doi: https://doi.org/10.1016/j.ijmedinf.2023.105115.

M. Maistro, L. Chaves Lima, J. Grue Simonsen, and C. Lioma. Principled multi-aspect evaluation measures of rankings. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 1232–1242, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482287.

P. Mandayam Comar and S. H. Sengamedu. Intent based relevance estimation from click logs. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 59–66, New York, NY, USA, 2017.

Association for Computing Machinery. ISBN 9781450349185. doi: 10.1145/3132847.3132870.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval.* Cambridge University Press, 2008.

A. Mardani, R. E. Hooker, S. Ozkul, Y. Sun, M. Nilashi, H. Z. Sabzi, and G. C. Fei. Application of decision making and fuzzy sets theory to evaluate the healthcare and medical problems: A review of three decades of research with recent developments. *Expert Syst. Appl.*, 137:202–231, 2019.

M. McGregor, L. Azzopardi, and M. Halvey. A systematic review of cost, effort, and load research in information search and retrieval, 1972–2020. *ACM Trans. Inf. Syst.*, 42(1), aug 2023. ISSN 1046-8188. doi: 10.1145/3583069.

R. Miotto, S. Y. Jiang, and C. Weng. etacts: A method for dynamically filtering clinical trial search results. *J. Biomed. Informatics*, 46(6):1060–1067, 2013. doi: 10.1016/j.jbi.2013.07.014.

B. Mitra and N. Craswell. An introduction to neural information retrieval. *Found. Trends Inf. Retr.*, 13(1):1–126, 2018.

B. Mitra, N. Craswell, et al. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126, 2018.

S. Mizzaro. How many relevances in information retrieval? *Interacting with Computers*, 10(3):303–320, 06 1998. ISSN 0953-5438. doi: 10.1016/S0953-5438(98)00012-5.

C. N. Mooers. Zatocoding applied to mechanical organization of knowledge. *American documentation*, 2(1):20–32, 1951.

M. Moradi, K. Blagec, F. Haberl, and M. Samwald. GPT-3 models are poor few-shot learners in the biomedical domain. *CoRR*, abs/2109.02555, 2021.

B. Moulahi, B. Moulahi, B. Moulahi, L. Tamine, and S. B. Yahia. Iaggregator: Multidimensional relevance aggregation based on a fuzzy operator. *Journal of the Association for Information Science and Technology*, 2014a. doi: 10.1002/asi.23094.

B. Moulahi, L. Tamine, and S. B. Yahia. Toward a personalized approach for combining document relevance estimates. In V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, and G.-J. Houben, editors, *User Modeling, Adaptation, and Personalization*, pages 158–170, Cham, 2014b. Springer International Publishing. ISBN 978-3-319-08786-3.

B. Moulahi, L. Tamine, and S. B. Yahia. iaggregator: Multidimensional relevance aggregation based on a fuzzy operator. *J. Assoc. Inf. Sci. Technol.*, 65(10): 2062–2083, 2014c. doi: 10.1002/asi.23094.

D. F. Navarro, K. Ijaz, D. Rezazadegan, H. Rahimi-Ardabili, M. Dras, E. Coiera, and S. Berkovsky. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, page 105122, 2023.

M. Neumann, D. King, I. Beltagy, and W. Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5034.

Y. Ni, M. Bermudez, S. Kennebeck, S. Liddy-Hicks, and J. Dexheimer. A real-time automated patient screening system for clinical trials eligibility in an emergency department: design and evaluation. *JMIR medical informatics*, 7(3):e14185, 2019. doi: 10.2196/14185.

D. W. North. A tutorial introduction to decision theory. *IEEE Trans. Syst. Sci. Cybern.*, 4(3):200–210, 1968.

M. I. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan, C. Sutton, and A. Odena. Show your work: Scratchpads for intermediate computation with language models. *CoRR*, abs/2112.00114, 2021.

B. Omidvar-Tehrani, A. Personnaz, and S. Amer-Yahia. Guided text-based item exploration. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 3410–3420, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392365. doi: 10.1145/3511808.3557141.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

S. Opricovic and G.-H. Tzeng. Compromise solution by mcdm methods: A comparative analysis of vikor and topsis. *European Journal of Operational Research*, 156(2):445–455, 2004. ISSN 0377-2217. doi: 10.1016/S0377-2217(03)00020-1.

M. Oroszlányová, C. Lopes, S. Nunes, and C. Ribeiro. Using the characteristics of documents, users and tasks to predict the situational relevance of health web documents. *Journal of Information Systems Engineering & Management*, 2(4): 25, 2017.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

E. C. O'Brien, S. R. Raman, A. Ellis, B. G. Hammill, L. G. Berdan, T. Rorick, S. Janmohamed, Z. Lampron, A. F. Hernandez, and L. H. Curtis. The use of electronic health records for recruitment in clinical trials: a mixed methods analysis of the harmony outcomes electronic health record ancillary study. *Trials*, 22:1–8, 2021.

D. Palacio, G. Cabanac, C. Sallaberry, and G. Hubert. On the evaluation of geographic information retrieval systems: Evaluation framework and case study. *International Journal on Digital Libraries*, 2010. doi: 10.1007/s00799-011-0070-z.

J. Palotti, G. Zuccon, and A. Hanbury. Mm: A new framework for multidimensional evaluation of search engines. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 1699–1702, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3269261.

J. Palotti, G. Zuccon, and A. Hanbury. Consumer health search on the web: Study of web page understandability and its integration in ranking algorithms. *Journal of Medical Internet Research*, 2019. doi: 10.2196/10986.

J. Papathanasiou and N. Ploskas. Multiple criteria decision aid. *Methods, Examples and Python Implementations*, 136:XVII, 173, 2018. doi: 10.1007/978-3-319-91648-4.

T. K. Park. Toward a theory of user-based relevance: A call for a new paradigm of inquiry. *J. Am. Soc. Inf. Sci.*, 45:135–141, 1994.

G. Pasi and M. Viviani. Application of aggregation operators to assess the credibility of user-generated content in social media. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations: 17th International Conference, IPMU 2018, Cádiz, Spain, June 11-15, 2018, Proceedings, Part I 17*, pages 342–353. Springer, 2018.

G. Pasi, M. Viviani, and A. Carton. A multi-criteria decision making approach based on the choquet integral for assessing the credibility of user-generated content. *Inf. Sci.*, 503:574–588, 2019. doi: 10.1016/j.ins.2019.07.037.

S. B. Patel and K. Lam. Chatgpt: the future of discharge summaries? *The Lancet Digital Health*, 5(3):e107–e108, 2023.

G. Peikos and G. Pasi. Multidimensional relevance in legal and health domains. In *IIR*, volume 2947 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.

G. Peikos, O. Espitia, and G. Pasi. UNIMIB at TREC 2021 clinical trials track. *CoRR*, abs/2207.13514, 2022.

G. Peikos, D. Alexander, G. Pasi, and A. P. de Vries. Investigating the impact of query representation on medical information retrieval. In J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, and A. Caputo, editors, *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part II*, volume 13981 of *Lecture Notes in Computer Science*, pages 512–521. Springer, 2023a. doi: 10.1007/978-3-031-28238-6_42.

G. Peikos, S. Symeonidis, P. Kasela, and G. Pasi. Utilizing chatgpt to enhance clinical trial enrollment [submitted in artificial intelligence in medicine]. *CoRR*, abs/2306.02077, 2023b.

V. Penadés-Plà, T. García-Segura, J. V. Martí, and V. Yepes. A review of multi-criteria decision-making methods applied to the sustainable bridge design. *Sustainability*, 8(12), 2016. ISSN 2071-1050. doi: 10.3390/su8121295.

L. T. Penberthy, B. A. Dahman, V. I. Petkov, and J. P. DeShazo. Effort required in eligibility screening for clinical trials. *Journal of Oncology Practice*, 8(6):365–370, 2012.

E. Perez, D. Kiela, and K. Cho. True few-shot learning with language models. In *NeurIPS*, pages 11054–11070, 2021.

M. Pourreza and F. Ensan. Towards semantic-driven boolean query formalization for biomedical systematic literature reviews. *Int. J. Medical Informatics*, 170: 104928, 2023.

R. Pradeep, Y. Li, Y. Wang, and J. Lin. Neural query synthesis and domain-specific ranking templates for multi-stage clinical trial matching. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2325–2330. Association for Computing Machinery, 2022. ISBN 9781450387323. doi: 10.1145/3477495.3531853.

N. Prieto-Chavana, J. Weeds, and D. J. Weir. Automated query generation for evidence collection from web search engines. *CoRR*, abs/2303.08652, 2023. doi: 10.48550/arXiv.2303.08652.

D. G. P. Putri, M. Viviani, and G. Pasi. Social search and task-related relevance dimensions in microblogging sites. In S. Aref, K. Bontcheva, M. Braghieri, F. Dignum, F. Giannotti, F. Grisolia, and D. Pedreschi, editors, *Social Informatics*, pages 297–311, Cham, 2020. Springer International Publishing. ISBN 978-3-030-60975-7.

D. G. P. Putri, M. Viviani, and G. Pasi. A multi-task learning model for multidimensional relevance assessment. In K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 103–115, Cham, 2021. Springer International Publishing. ISBN 978-3-030-85251-1.

J. Qiu, L. Li, J. Sun, J. Peng, P. Shi, R. Zhang, Y. Dong, K. Lam, F. P. Lo, B. Xiao, W. Yuan, D. Xu, and B. Lo. Large AI models in health informatics: Applications, challenges, and the future. *CoRR*, abs/2303.11568, 2023.

J. Qu, J. Arguello, and Y. Wang. Towards explainable retrieval models for precision medicine literature search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1593–1596, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401277.

J. Qu, J. Arguello, and Y. Wang. A deep analysis of an explainable retrieval model for precision medicine literature search. In D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, and F. Sebastiani, editors, *Advances in Information Retrieval*, pages 544–557, Cham, 2021. Springer International Publishing. ISBN 978-3-030-72113-8.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, H. F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. van den Driessche, L. A. Hendricks, M. Rauh, P. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen, S. M. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J. Lespiau, M. Tsimpoukelli, N. Grigorev, D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. de Masson d'Autume, Y. Li, T. Terzi, V. Mikulik, I. Babuschkin, A. Clark, D. de Las Casas, A. Guy, C. Jones, J. Bradbury, M. J. Johnson, B. A. Hechtman, L. Weidinger, I. Gabriel, W. S. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. Ayoub, J. Stanway, L. Bennett, D. Hassabis, K. Kavukcuoglu, and G. Irving. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446, 2021.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

N. C. Rakotonirina, R. Dessì, F. Petroni, S. Riedel, and M. Baroni. Can discrete information extraction prompts generalize across language models? *CoRR*, abs/2302.09865, 2023. doi: 10.48550/arXiv.2302.09865.

S. Ravikumar, K. Talamadupula, R. Balakrishnan, and S. Kambhampati. Raprop: Ranking tweets by exploiting the tweet/user/web ecosystem and inter-tweet agreement. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 2345–2350, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450322638. doi: 10.1145/2505515.2505667.

N. Rekabsaz and M. Lupu. A real-world framework for translator as expert retrieval. In E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, A. Hanbury, and E. Toms, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, pages 141–152, Cham, 2014. Springer International Publishing. ISBN 978-3-319-11382-1.

L. Reynolds and K. McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *CHI Extended Abstracts*, pages 314:1–314:7. ACM, 2021.

A. M. Rinaldi. An ontology-driven approach for semantic information retrieval on the web. *ACM Trans. Internet Technol.*, 9(3), jul 2009. ISSN 1533-5399. doi: 10.1145/1552291.1552293.

K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. R. Hersh, S. Bedrick, A. J. Lazar, and S. Pant. Overview of the TREC 2017 precision medicine track. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017*, volume 500-324 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2017.

K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. R. Hersh, S. Bedrick, and A. J. Lazar. Overview of the TREC 2018 precision medicine track. In *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018*, volume 500-331 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2018.

K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. R. Hersh, S. Bedrick, A. J. Lazar, S. Pant, and F. Meric-Bernstam. Overview of the TREC 2019 precision medicine track. In *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019*, volume 1250 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2019.

S. E. Robertson and K. S. Jones. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.*, 27(3):129–146, 1976. doi: 10.1002/asi.4630270302.

S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In W. B. Croft and C. J. van

Rijsbergen, editors, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 232–241. ACM/Springer, 1994. doi: 10.1007/978-1-4471-2099-5_24.

S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In D. K. Harman, editor, *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST), 1994.

S. E. Robertson, H. Zaragoza, and M. J. Taylor. Simple BM25 extension to multiple weighted fields. In D. A. Grossman, L. Gravano, C. Zhai, O. Herzog, and D. A. Evans, editors, *Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management, Washington, DC, USA, November 8-13, 2004*, pages 42–49. ACM, 2004. doi: 10.1145/1031171.1031181.

O. Rubin, J. Herzig, and J. Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.191.

S. Ruder. An overview of gradient descent optimization algorithms. *preprint arXiv:1609.04747*, 2016.

T. Russell-Rose, J. Chamberlain, and L. Azzopardi. Information retrieval in the workplace: A comparison of professional search practices. *Inf. Process. Manag.*, 54(6):1042–1057, 2018.

M. Rybinski, J. Xu, and S. Karimi. Clinical trial search: Using biomedical language understanding models for re-ranking. *J. Biomed. Informatics*, 109:103530, 2020. doi: 10.1016/j.jbi.2020.103530.

M. Rybinski, S. Karimi, and A. Khoo. Science2cure: A clinical trial search prototype. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2620–2624. ACM, 2021. doi: 10.1145/3404835.3462797.

M. Rybiński, V. Nguyen, and S. Karimi. Csiromed team report of trec 2021 clinical trials track: Experiments with bert reranking methods. In *TREC Proceedings*, 2021.

J. Saad-Falcon, O. Khattab, K. Santhanam, R. Florian, M. Franz, S. Roukos, A. Sil, M. A. Sultan, and C. Potts. UDAPDR: unsupervised domain adaptation via LLM prompting and distillation of rerankers. *CoRR*, abs/2303.00807, 2023. doi: 10.48550/arXiv.2303.00807.

M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011.

A. K. Sahraoui and R. Faiz. Time sensitivity for personalized search. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 585–592, 2017. doi: 10.1109/AICCSA.2017.77.

T. Sakai. Alternatives to bpref. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 71–78. ACM, 2007. doi: 10.1145/1277741.1277756.

G. Salton. *Automatic information organization and retrieval.* McGraw Hill Text, 1968.

T. Saracevic. The stratified model of information retrieval interaction: Extension and applications. In *Proceedings of the annual meeting-american society for information science*, volume 34, pages 313–327. LEARNED INFORMATION (EUROPE) LTD, 1997.

T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part ii: Nature and manifestations of relevance. *Journal of the American society for information science and technology*, 58(13): 1915–1933, 2007.

T. Saracevic. *The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really?* Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2016a. doi: 10.2200/S00723ED1V01Y201607ICR050.

T. Saracevic. *The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really?* Morgan & Claypool Publishers, 2016b.

Y. Sasaki, T. Komatsuda, A. Keyaki, and J. Miyazaki. A new readability measure for web documents and its evaluation on an effective web search engine. In *Proceedings of the 18th International Conference on Information Integration and Web-Based Applications and Services*, iiWAS '16, page 355–362, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450348072. doi: 10.1145/3011141.3011172.

L. Schamber, M. B. Eisenberg, and M. S. Nilan. A re-examination of relevance: toward a dynamic, situational definition. *Information processing & management*, 26(6):755–776, 1990.

H. Schütze, C. D. Manning, and P. Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

E. Sezgin, J. Sirrianni, and S. L. Linwood. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the us health care system: outlook of generative pretrained transformer 3 (gpt-3) as a service model. *JMIR medical informatics*, 10(2):e32875, 2022.

M. Shajalal, Shajalal, M. Aono, and M. Aono. Coverage-based query subtopic diversification leveraging semantic relevance. *Knowledge and Information Systems*, 2020. doi: 10.1007/s10115-020-01470-3.

A. Shekhovtsov and W. Salabun. A comparative case study of the VIKOR and TOPSIS rankings similarity. In *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES-2020*, volume 176 of *Procedia Computer Science*, pages 3730–3740. Elsevier, 2020. doi: 10.1016/j.procs.2020.09.014.

R. Shin, C. Lin, S. Thomson, C. Chen, S. Roy, E. A. Platanios, A. Pauls, D. Klein, J. Eisner, and B. Van Durme. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.608.

T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.346.

R. Shraga, H. Roitman, G. Feigenblat, and M. Cannim. Web table retrieval using multimodal deep learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1399–1408, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401120.

A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 525–534, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595938039. doi: 10.1145/1321440.1321515.

R. F. Silva, C. K. Roy, M. M. Rahman, K. A. Schneider, K. Paixao, and M. de Almeida Maia. Recommending comprehensive solutions for programming tasks by mining crowd knowledge. In *2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC)*, pages 358–368, 2019. doi: 10.1109/ICPC.2019.00054.

V. Singh and M. Dave. Improving result diversity using query term proximity in exploratory search. In S. Madria, P. Fournier-Viger, S. Chaudhary, and P. K. Reddy, editors, *Big Data Analytics*, pages 67–87, Cham, 2019. Springer International Publishing. ISBN 978-3-030-37188-3.

K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. K. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, N. Schärli, A. Chowdhery, P. A. Mansfield, B. A. y Arcas, D. R. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. K. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan. Large language models encode clinical knowledge. *CoRR*, abs/2212.13138, 2022.

S. Sivarajkumar, H. A. Mohammad, D. Oniani, K. Roberts, W. Hersh, H. Liu, D. He, S. Visweswaran, and Y. Wang. Clinical information retrieval: A literature review. *preprint*, 2023.

I. Soboroff. Building test collections: An interactive guide for students and others without their own evaluation conference series. In *SIGIR*, pages 1407–1410. ACM, 2017.

I. Soboroff. Overview of trec 2021. In *TREC Proceedings*. Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, 2022-05-06 04:05:00 2022.

A. Soufan, I. Ruthven, and L. Azzopardi. Untangling the concept of task in information seeking and retrieval. In *ICTIR*, pages 73–81. ACM, 2021.

K. Spark-Jones. Report on the need for and provision of an'ideal'information retrieval test collection. *Computer Laboratory*, 1975.

Y. Sun, Y. Zhang, J. Gwizdka, and C. B. Trace. Consumer evaluation of the quality of online health information: Systematic literature review of relevant criteria and indicators. *J Med Internet Res*, 21(5):e12522, May 2019. ISSN 1438-8871.

D. R. Swanson. Subjective versus objective relevance in bibliographic retrieval systems. *The library quarterly*, 56(4):389–398, 1986.

S. Symeonidis, G. Peikos, and A. Arampatzis. Unsupervised consumer intention and sentiment mining from microblogging data as a business intelligence tool. *Oper. Res.*, 22(5):6007–6036, 2022.

L. Tamine and C. Chouquet. On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings. *Inf. Process. Manag.*, 53(2):332–350, 2017.

L. Tamine and L. Goeuriot. Semantic information retrieval on medical texts: Research challenges, survey, and open issues. *ACM Comput. Surv.*, 54(7):146:1–146:38, 2022. doi: 10.1145/3462476.

L. Tamine, L. B. Jabeur, and W. Bahsoun. *On Using Social Context to Model Information Retrieval and Collaboration in Scientific Research Community*, pages 133–155. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-19047-6. doi: 10.1007/978-3-642-19047-6_6.

R. Tang, X. Han, X. Jiang, and X. Hu. Does synthetic data generation of llms help clinical text mining? *CoRR*, abs/2303.04360, 2023.

P. Thomas, S. Spielman, N. Craswell, and B. Mitra. Large language models can accurately predict searcher preferences. *preprint arXiv:2309.10621*, 2023.

E. Triantaphyllou. Multi-criteria decision making methods. In *Multi-criteria decision making methods: A comparative study*, pages 5–21. Springer, 2000. doi: 10.1007/978-1-4757-3157-6.

M. Tsagkias, T. H. King, S. Kallumadi, V. Murdock, and M. de Rijke. Challenges and research opportunities in ecommerce search and recommendations. In *ACM Sigir Forum*, volume 54, pages 1–23. ACM New York, NY, USA, 2021.

G.-H. Tzeng and J.-J. Huang. *Multiple attribute decision making: methods and applications*. CRC press, 2011. ISBN 9781439861585.

R. Upadhyay, G. Pasi, and M. Viviani. An unsupervised approach to genuine health information retrieval based on scientific evidence. In R. Chbeir, H. Huang, F. Silvestri, Y. Manolopoulos, and Y. Zhang, editors, *Web Information Systems Engineering – WISE 2022*, pages 119–135, Cham, 2022. Springer International Publishing. ISBN 978-3-031-20891-1.

S. Uprety, Y. Su, D. Song, and J. Li. Modeling multidimensional user relevance in ir using vector spaces. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 993–996, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210130.

A. Usta, I. S. Altingovde, R. Ozcan, and Özgür Ulusoy. Learning to rank for educational search engines. *IEEE Transactions on Learning Technologies*, 2021. doi: 10.1109/tlt.2021.3075196.

Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *J. Am. Medical Informatics Assoc.*, 18(5):552–556, 2011.

P. Vakkari. The usefulness of search results: A systematization of types and predictors. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, CHIIR '20, page 243–252, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368926. doi: 10.1145/3343413.3377955.

B. van Aken, I. Trajanovska, A. Siu, M. Mayrdorfer, K. Budde, and A. Loeser. Assertion detection in clinical notes: Medical language models to the rescue? In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, Online, 2021. Association for Computational Linguistics.

J. van Doorn, D. Odijk, D. M. Roijers, and M. de Rijke. Balancing relevance criteria through multi-objective optimization. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 769–772, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340694. doi: 10.1145/2911451.2914708.

M. van Opijnen and C. Santos. On the concept of relevance in legal information retrieval. *Artif. Intell. Law*, 25(1):65–87, 2017. doi: 10.1007/s10506-017-9195-8.

S. Vargas, P. Castells, and D. Vallet. Explicit relevance models in intent-oriented information retrieval diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, page 75–84, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450314725. doi: 10.1145/2348283.2348297.

S. Verberne, J. He, U. Kruschwitz, B. Larsen, T. Russell-Rose, and A. P. de Vries. First international workshop on professional search (profs2018). In K. Collins-Thompson, Q. Mei, B. D. Davison, Y. Liu, and E. Yilmaz, editors, *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1431–1434. ACM, 2018.

S. Verberne, J. He, G. Wiggers, T. Russell-Rose, U. Kruschwitz, and A. P. de Vries. Information search in a professional context - exploring a collection of professional search tasks. *CoRR*, abs/1905.04577, 2019.

B. C. Vickery. The structure of information retrieval systems. In *Proceedings of the International Conference on Scientific Information*, volume 2, pages 1275–1290, 1959a.

B. C. Vickery. Subject analysis for information retrieval. In *Proceedings of the International Conference on Scientific Information*, volume 2, pages 855–865, 1959b.

M. S. Von Itzstein, M. Hullings, H. Mayo, M. S. Beg, E. L. Williams, and D. E. Gerber. Application of information technology to clinical trial evaluation and enrollment: a review. *JAMA oncology*, 7(10):1559–1566, 2021.

E. M. Voorhees and W. R. Hersh. Overview of the TREC 2012 medical records track. In *Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012*, volume 500-298 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2012.

L. Wang, N. Yang, and F. Wei. Query2doc: Query expansion with large language models. *CoRR*, abs/2303.07678, 2023a. doi: 10.48550/arXiv.2303.07678.

S. Wang, H. Scells, B. Koopman, and G. Zuccon. Can chatgpt write a good boolean query for systematic review literature search? *CoRR*, abs/2302.03495, 2023b. doi: 10.48550/arXiv.2302.03495.

X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. *CoRR*, abs/2203.11171, 2022.

Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu. Clinical information extraction applications: A literature review. *J. Biomed. Informatics*, 77:34–49, 2018.

J. Watt, R. Borhani, and A. K. Katsaggelos. *Machine learning refined: Foundations, algorithms, and applications*. Cambridge University Press, 2020.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

G. Wiggers, S. Verberne, and G. Zwenne. Exploration of intrinsic relevance judgments by legal professionals in information retrieval systems. In *Proceedings of the 17th Dutch-Belgian Information Retrieval workshop*, pages 5–8, 2018.

G. Wiggers, S. Verberne, W. van Loon, and G.-J. Zwenne. Bibliometric-enhanced legal information retrieval: Combining usage and citations as flavors of impact relevance. *Journal of the Association for Information Science and Technology*, 2023.

S. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, B. Zhao, and H. Xu. Deep learning in clinical natural language processing: a methodical review. *J. Am. Medical Informatics Assoc.*, 27(3):457–470, 2020.

T. Wu, M. Terry, and C. J. Cai. AI chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *CHI*, pages 385:1–385:22. ACM, 2022.

B. Xu, H. Lin, Y. Lin, Y. Ma, L. Yang, J. Wang, and Z. Yang. Improve biomedical information retrieval using modified learning to rank methods. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(6):1797–1809, 2016.

Y. Xu. Relevance judgment in epistemic and hedonic information searches. *J. Assoc. Inf. Sci. Technol.*, 58(2):179–189, 2007. doi: 10.1002/asi.20461.

Y. Xu and Z. Chen. Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, 57(7):961–973, 2006a.

Y. C. Xu and Z. Chen. Relevance judgment: What do information users consider beyond topicality? *J. Am. Soc. Inf. Sci. Technol.*, 57(7):961–973, 2006b. ISSN 1532-2882. doi: http://dx.doi.org/10.1002/asi.v57:7.

M. Yan, Y. Wen, Q. Shi, X. Tian, and L. Zhao. A multimodal retrieval and ranking method for scientific documents based on hfs and xlnet. *Sci. Program.*, 2022, jan 2022. ISSN 1058-9244. doi: 10.1155/2022/5373531.

H. Yang, P. Gupta, R. Fernández Galán, D. Bu, and D. Jia. Seasonal relevance in e-commerce search. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4293–4301, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3481951.

X. Yang, A. Chen, N. M. Pournejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, A. B. Costa, M. G. Flores, Y. Zhang, T. Magoc, C. A. Harle, G. P. Lipori, D. A. Mitchell, W. R. Hogan, E. A. Shenkman, J. Bian, and Y. Wu. A large language model for electronic health records. *npj Digit. Medicine*, 5, 2022a.

X. Yang, N. M. Pournejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, M. G. Flores, Y. Zhang, T. Magoc, C. A. Harle, G. P. Lipori, D. A. Mitchell, W. R. Hogan, E. A. Shenkman, J. Bian, and Y. Wu. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *CoRR*, abs/2203.03540, 2022b.

P. L. Yu. A class of solutions for group decision problems. *Management Science*, 19(8):936–946, 1973. ISSN 00251909, 15265501.

E. Yulianti, R.-C. Chen, F. Scholer, W. B. Croft, and M. Sanderson. Ranking documents by answer-passage quality. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 335–344, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210028.

D. Zaikis and I. P. Vlahavas. TP-DDI: transformer-based pipeline for the extraction of drug-drug interactions. *Artif. Intell. Medicine*, 119:102153, 2021.

E. K. Zavadskas and Z. Turskis. Multiple criteria decision making (mcdm) methods in economics: an overview. *Technological and economic development of economy*, 17(2):397–427, 2011.

E. K. Zavadskas, A. Kaklauskas, and V. Sarka. The new method of multicriteria complex proportional assessment of projects. *Technological and economic development of economy*, 1(3):131–139, 1994.

M. Zeleny. *Multiple criteria decision making Kyoto 1975*, volume 123. Springer Science & Business Media, 1982.

C. Zhai and J. D. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management, Atlanta, Georgia, USA, November 5-10, 2001*, pages 403–410. ACM, 2001. doi: 10.1145/502585.502654.

Q. Zhang and A. Youssef. An approach to math-similarity search. In S. M. Watt, J. H. Davenport, A. P. Sexton, P. Sojka, and J. Urban, editors, *Intelligent Computer Mathematics*, pages 404–418, Cham, 2014. Springer International Publishing. ISBN 978-3-319-08434-3.

W. Zhang, D. Song, P. Zhang, X. Zhao, and Y. Hou. A sequential latent topic-based readability model for domain-specific information retrieval. In G. Zuccon, S. Geva, H. Joho, F. Scholer, A. Sun, and P. Zhang, editors, *Information Retrieval Technology*, pages 241–252, Cham, 2015. Springer International Publishing. ISBN 978-3-319-28940-3.

Y. Zhang, Y. Zhang, P. Qi, C. D. Manning, and C. P. Langlotz. Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*, 06 2021. ISSN 1527-974X.

Z. Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, pages 1–2. Ieee, 2018.

W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen. A survey of large language models. *preprint arXiv:2303.18223*, 2023.

Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: Improving few-shot performance of language models. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR, 2021.

D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, O. Bousquet, Q. Le, and E. H. Chi. Least-to-most prompting enables complex reasoning in large language models. *CoRR*, abs/2205.10625, 2022.

H. Zhuang, X. Wang, M. Bendersky, A. Grushetsky, Y. Wu, P. Mitrichev, E. Sterling, N. Bell, W. Ravina, and H. Qian. Interpretable ranking with generalized additive models. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 499–507, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450382977. doi: 10.1145/3437963.3441796.

E. Znaidi, L. Tamine, and C. Latiri. Aggregating semantic information nuggets for answering clinical queries. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, SAC '16, page 1041–1047, New York, NY, USA, 2016.

Association for Computing Machinery. ISBN 9781450337397. doi: 10.1145/2851613.2851694.