FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Using health statistics to improve medical and health search

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Medizinische Informatik

eingereicht von

Tawan Sierek
Matrikelnummer 0326328

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung: Privatdoz. Dr. Allan Hanbury

Wien, 23.04.2015

(Unterschrift Tawan Sierek)

(Unterschrift Betreuung)

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Using health statistics to improve medical and health search

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Medical Informatics

by

Tawan Sierek
Registration Number 0326328

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Privatdoz. Dr. Allan Hanbury

Vienna, 23.04.2015

(Signature of Author) (Signature of Advisor)

Erklärung zur Verfassung der Arbeit

Tawan Sierek
Margaretenstraße 61/3, 1050 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum) (Unterschrift Tawan Sierek)

Acknowledgements

I would like to thank my supervisor Dr. Allan Hanbury for his continuous feedback and help. I would also like to thank Marie for her patience and her endless support. I am grateful to my parents, Kingpaka and Andreas, for their support and helpful advice. And I am grateful to my brother, Leo, for welcoming me in his home, so that I can complete my studies.

Abstract

Healthcare professionals often find additional information by consulting information retrieval systems (IR) when treating a patient. But they face an ever growing amount of scientific literature, which makes it harder to find the relevant citations or articles for a given clinical case. Non-professionals now commonly seek information about health on their own, often starting at a web search engine. Both types of users benefit from the effectiveness of IR techniques, which are essential for web search engines or retrieval systems accessing bibliographic databases. A critical part is the ranking process, as this determines which article or web-page is more relevant than others and should, therefore, be ranked higher. Our goal is to improve the ranking process within health searches by taking available health statistics into account. We assume that it is beneficial for the user if text documents that cover more frequent diseases are ranked higher than others. Based on this assumption, we also believe that health search can be contextualized, by adapting the ranking to a patient profile that contains age and sex data. It is common knowledge that a number of diseases are unequally distributed among men and women, as well as among young and old people.

To the best of our knowledge, IR approaches based on health statistics are not covered by scientific literature. We develop a probabilistic model that incorporates an epidemiological measure and a patient profile. We implement a prototype based on the formal model. The prototype re-ranks the top 150 results of a state-of-the-art system. It maps the documents to ICD-9-CM codes and, depending on the probability of a diagnosis with the same code for a patient with a given profile, the document is ranked higher or lower.

The prototype is evaluated using the test collections of two recent evaluation campaigns in the health domain. We establish baselines with the best-performing IR methods, of a widely used open source search engine. At these times, our experiments show only a minor improvement over the baseline, which we can not report as statistically significant. Our prototype maps documents to ICD-9-CM codes automatically, but relies only on Wikipedia articles serving as the ground truth. Due to this sparseness of training data, we can not evaluate this crucial step and, therefore, our results are biased. We suggest conducting further research based on our formal model, but with test collections of manually annotated documents.

Kurzfassung

Behandelnde Ärzte und Ärztinnen beziehen oft zusätzliche Informationen aus Information Retrieval (IR) Systemen. Aber die immer größer werdende Anzahl von wissenschaftlichen Texten in der Medizin machen das Auffinden von relevanter Information immer schwerer. Auch immer mehr Laien informieren sich selbständig im Internet über medizinische Themen und benutzen oft eine Suchmaschine als Ausgangspunkt. Beide Benutzertypen profitieren von der Effizienz von IR Techniken, welche das Kernstück von Suchmaschinen bilden. Ein wichtiger Schritt ist die Reihung von Suchergebnissen. Unser Ziel ist es, die Reihung bei medizinischen Suchanfragen zu verbessern, in dem Statistiken von diversen Krankheiten berücksichtigt werden. Wir nehmen an, dass ein medizinischer Artikel, welcher eine bestimmte Krankheit zum Thema hat, relevanter ist wenn die Krankheit häufiger vorkommt. Aufbauend auf dieser Annahme, glauben wir auch, dass die Reihung von Suchergebnissen an ein Patientenprofil angepasst werden kann, in dem das Alter und Geschlecht berücksichtigt werden. Es ist allgemein bekannt, dass einige Krankheiten unterschiedlich oft bei Männern und Frauen, beziehungsweise jungen und älteren Personen, vorkommen.

Nach unserem besten Wissen existieren keine wissenschaftlichen Arbeiten, die IR Techniken, basierend of Gesundheitsstatistiken, thematisieren. Wir entwickeln ein stochastisches Modell, welches ein epidemiologisches Maß und ein Patientenprofil einbeziehen. Aufbauend auf dem formalen Modell implementieren wir einen Prototyp. Der Prototyp ordnet das Ergebnis einer state-of-the-art Suchmaschine neu, in dem er die Suchergebnisse zu ICD-9-CM Codes zuordnet. Anhand der Wahrscheinlichkeit einer Diagnose mit dem selbem Code, im Bezug zu einem Patientenprofil wird das Suchergebnis höher oder niedriger gereiht.

Der Prototyp wird mit zwei Testkollektionen, von zwei kürzlich organisierten Evaluierungskampagnen, evaluiert und getestet. Wir etablieren Baselines mit den effizientesten IR Methoden, die in einer weit verbreitete Open Source Suchmaschine implementiert sind. Unsere Experimente zeigen innerhalb einer Testkollektion eine minimale Verbesserung. Diese ist aber nicht statistisch Signifikant. Der Prototyp ordnet Suchergebnisse zu ICD-9-CM Codes basierend auf Wikipedia-Artikel, welche als Ground Truth dienen. Durch das spärliche Vorhandensein von Trainingsdaten, können wir diesen kritischen Schritt nicht evaluieren und deshalb sind unsere Ergebnisse beeinflusst. Wir schlagen vor, weiterhin Forschungen auf Basis des formalen Modells durchzuführen, aber mit Testkollektionen mit manuell annotierten Dokumenten.

Contents

CHAPTER 1

Introduction

Finding relevant information in the vast amount of medical literature has always been a challenge, long before computers had been invented. Already in the late 1870s John Shaw Billings created the *Index Medicus* because of the need for organizing biomedical literature, books and articles [56]. It was one of the most important access points to medical literature and the U.S. National Library of Medicine continued to publish it until 2004 [108]. But with the spread of computers and the advancements in the field of information retrieval (IR), search engines became the main source of information. Not only for healthcare professionals, but also for the layperson, who wants to find out more about specific topics within the health and medical field [26].

1.1 Motivation

We introduce this section with a specific use case. It does not cover all aspects of IR in the health domain, but it is sufficient for understanding the motivation behind our work.

We assume a user, who needs to find information about up-to-date treatments of high blood pressure. The user has access to a web-page that displays a list of links to all scientific biomedical articles from the past twenty years. In order to find articles about treatments of high blood pressure, the user just needs to click on the first link, read the abstract and decide whether the article contains useful information or not. If not, then the user proceeds to the next link, clicks on it, reads the abstract, and so on, until a relevant article is found. It is obvious that this approach is pointless, since there have been millions of articles published in the biomedical domain [54], and only a very small subset covers treatments of high blood pressure. In other words, the proportion of relevant articles is too small to be discovered just by chance within a sensible time frame. With the help of a search engine however, the user can submit a query that is composed of keywords that are likely to appear in a relevant article, for example the query `treatment high blood pressure`. The system responds with a smaller list of links, that point only to articles that contain these keywords. The basic assumption is that the proportion of relevant links is higher within the returned subset. Again, the user opens link after link but is more likely

to find a relevant link, and in a much shorter time frame, which is the main benefit of using a search engine. An optimal response would be a list of links, where each of them points to a relevant article. In this case, the user finds useful information immediately, just by opening the first link. Further, the search engine can take the user directly to the article of the first link in the list, without waiting for the user to click it. This shortcut can be taken by the "I'm feeling lucky " button of *Google's* search engine [70]. But the button carries its label for a reason: a search engine can not guarantee that the very first article is relevant, because whether an article is really relevant to the user's information need is subject to the user. The search engine can only make an informed guess based on the query and other factors. An article that contains the keywords must not necessarily be useful to the user's information need. In addition, the first article might be somehow relevant but it does not satisfy the user's information need completely. For example, the article covers drugs that lower the blood pressure, but the user also wants information about treatments based on diet. For these reasons, the user must have the possibility to look at other articles, too. This process can be improved by displaying the titles and preview snippets of the articles. The preview snippets can also highlight the keywords from the query, so that the user can determine if the keywords are used in a relevant context. The user examines the list and only clicks on the links with the most promising titles and preview snippets. Still, depending on how long the list is, this can take more or less time.

In fact, there are two opposing demands on the IR system: (1) a user wants to find all relevant articles, (2) the user wants to invest as little time as possible in the search. The optimal trade-off of these two opposing requirements is subject to the user, it therefore makes sense to give the user the choice of where to draw the line. The search engine can support this decision by returning a rank-ordered list. The user experience can be significantly improved if the search engine responds with a ranked list [80]. By ranking the articles, the search engine implies that a user examine the list from top to bottom; in addition, the user understands that the further down the list the article is, the less relevant it is. The search engine's challenge is to determine the optimal ranking with the available data. Many algorithms and approaches have been developed to solve the ranking problem. A variety of relevance signals from various sources have been studied and successfully applied, but there is still room for improvement. In this thesis, we investigate if epidemiological data can be used as a relevance signal for a search in a clinical setting. The basic hypothesis is that frequent diseases are more relevant, than rare diseases, therefore documents about frequent diseases are more relevant than documents about rare diseases. Furthermore, if this assumption holds, then the relevance of a document covering a disease depends on the age and the sex of a patient, given that the search is performed in regard to a patient. This assumption is based on the fact that various diseases affect various demographic groups differently.

It is common knowledge that diseases and other health disorders are not equally distributed among men and women [66]. It is also obvious that numerous diseases are common among children, whereas many health conditions develop with age [32]. We think that contextualizing search and incorporating knowledge from available health statistics, in combination with the patient's demographic profile, provides an opportunity to improve the performance of existing information retrieval techniques. The main idea is that a search engine does not only rank articles based on the query, but also takes into account the context that is composed of relevant health factors like the age and sex of a patient.

Based on these assumptions, we further motivate our work from the perspectives of two different user groups: (1) the healthcare professional, who is an expert in the medical domain, and (2) the layperson, who consults a search engine on health related topics. Both groups have a different approach to IR in medicine, and their requirements on the systems are different in a number of aspects [97].

A professional's perspective

A physician faces many tasks during the daily work which can often be supported by additional information. These tasks include finding the correct diagnosis or choosing an appropriate treatment based on the diagnosis. Furthermore, with the adoption of the philosophy of *evidence-based medicine* (clinical decision making is based on high qualitative evidence from scientific literature), demand of efficient search systems has gained ever greater importance in medicine. Simultaneously, the amount of medical literature is growing fast, and with it the challenge of finding the right information [34, p.44-45, p. 109-116]. The *Text REtrieval Conference* (TREC) very recently organised (2014) a conference around an IR task, which focuses on the retrieval of biomedical literature within a clinical setting [103]. The task's goal is to retrieve full-text biomedical articles, which satisfy information needs, that may rise from a given clinical case report. Our approach to health search aims to support the physician by ranking articles based on the likeliness of the disease that the article discusses. If a physician issues the query `causes pulmonary hypertension`, it might match articles about *chronic left heart disease* and articles about HIV. Based on our assumptions, we argue that the article about chronic left heart disease is more relevant within the search context since this disease is more common.

A layperson's perspective

A recent study found that one in three US adults has used online resources to figure out a medical condition [26]. However, it is difficult for people to deal with medical vocabulary. Often, a layperson would describe medical concepts with elaborate descriptions, without the use of medical vocabulary [97]. A search engine can significantly improve its effectiveness if it adapts to the lower expertise of users [69]. Another important topic, which was studied by White and Horvitz, is the

> "...the unfounded escalation of concerns about common symptomatology, based on the review of search results and literature on the Web [111].".

Queries about a common symptom like a headache can mislead the user into being concerned about relatively uncommon health condition like a brain tumor. The authors coined the term *Cyberchondria*, which describes this phenomenon. This study is a hint for us, that information on the frequency of diseases should be considered by IR systems. Since consumer health search poses different requirements for IR systems, Goeuriot et al. initiated the creation of a new evaluation benchmark that addresses the needs of laypeople regarding their health condition [28]. In 2013, the Conference and Lab of the Evaluation Forum (CLEF) organized the first user-centered health IR task, which was then repeated in 2014 [29]. The goal of this task is to support laypeople in understanding their discharge summaries.

1.2 Goals and Scope

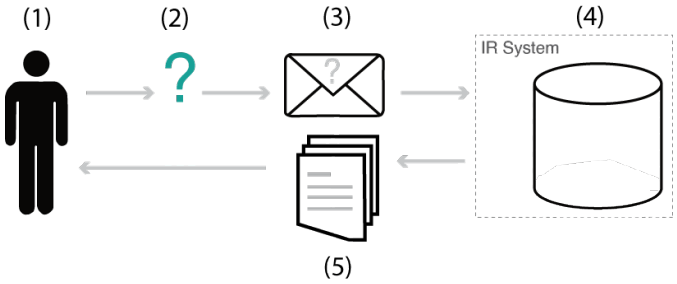

Figure 1.1: The IR use case

Figure 1.1 displays the IR use case, which is composed of a *user* (1) (sometimes referred to as *searcher*), who has an *information need* (2), that is expressed by the user in a *query* (3), and that query is processed by an IR system (4), which responds with a set of references to documents, ranked by relevance to the user's information need (5). This definition is not exhaustive, and others do exist, however it is sufficient for our needs.

In this work we investigate an approach which we refer to as personalized probabilistic health search (PPHS). We define *health search* as the search for information on health and medical topics in collections of unstructured textual data, through the means of search engines that process keyword queries and retrieve documents, references to documents, citations, abstracts or web-pages covering medical and health related topics ranked by relevance to the information need of the user.

We will refer to *documents* throughout this thesis, but our considerations are compliant to any representation of unstructured textual data, as long as the data is composed of discrete units, that are identifiable with unique ids. We also reduce the different concepts of words, phrases, compounds, inflectional variations of words to a single concept, which we refer to as *terms*. In the scope of this work, terms are atomic units which represent the building blocks of any document and search query.

As the title of this thesis suggests, this work does not aim to solve health search, but rather proposes an approach to improve established techniques that are already in use. Our considerations have their foundation in the *Probability Ranking Principle* (PRP), which states that if a retrieval system retrieves documents by decreasing probability of relevance, estimated from the available data, then this system is optimal with regard to the available data [78]. In classical approaches, the "available data" refers to content of the documents, or statistics which are inferred from the document collection, such as term frequencies, average document length and others.

The goal of this thesis is to investigate whether the ranking can be improved by including additional sources of data, such as epidemiological statistics and patient profiles. The thesis' main research questions are:

- Can epidemiological data be used to improve ranking of documents within the domain of health and medical search?

- Which epidemiological data sources are available and suitable for improving health search?
- Can an IR system which uses epidemiological data, improve the ranking, when it is adapted to context which is composed of a patient's sex and age?

It is important to emphasize that the information need is not equal to its query. A query is just the user's expression of the user's information need. Depending on the power of the query language and its syntax, these queries can be more or less accurate. Sometimes the user's current knowledge is not enough to actually convey precisely what kind of information he or she is looking for. Often an information need is not resolved by one single query. A user might refine a query after processing the result, and the information need is only completely satisfied after a search session composed of several search queries. In the course of this work we will refer to the anomalous state of knowledge (ASK) that a user experiences which eventually triggers the search. The ASK is a conceptual framework that was proposed by Belkin [9]. The term ASK tries to capture the vagueness and broad range of manifestations of information needs. To summarize the concept: A person has a state of knowledge about the world. This state of knowledge can come into conflict with new observations of the world, for example by the onset of a symptom, that cannot be explained by the person's current state of knowledge. The person recognizes this anomalous state of knowledge, and therefore tries to resolve it with external information. The reason why we introduced Belkin's ASK framework is that its perspective has consequences on the design of an IR system. It shifts the focus from just matching the query with documents to resolving the ASK.

With the research questions in mind, we conducted a review of scientific literature. Based on the approaches of other authors we developed a formal IR model which incorporates probability estimates based on epidemiological statistics. In order to assess the effectiveness of the formal model we developed a prototype as a reference implementation. The implementation processed the results of state-of-the-art retrieval models and calculated a new ranking score for the top 150 results of each query. Two ad hoc retrieval tasks were completed by the prototype and its performance was quantitatively measured. The results are presented in Chapter 6.

1.3 Overview of this work

The thesis is organized in these topics: Chapter 2 presents the state of the art in IR and also discusses relevant work. Chapter 3 provides a brief overview of the scientific field of epidemiology. It presents important statistical measures and relevant terminology. Furthermore, it demonstrates how diseases affect persons of different age and sex. Chapter 4 presents the formal retrieval model that was developed in the course of this work. Chapter 5 shows how we implemented a prototype based on the formal model. Chapter 6 demonstrates evaluation results of the prototype implementation. Finally, Chapter 7 discusses the results, and we present our conclusions and proposals for future work.

CHAPTER 2

State of the Art

Findings from an review of state-of-the-art methods and further studies of related work are presented in this chapter. The results of the review are organized in a three-stages model, which is depicted in Figure 2.1.

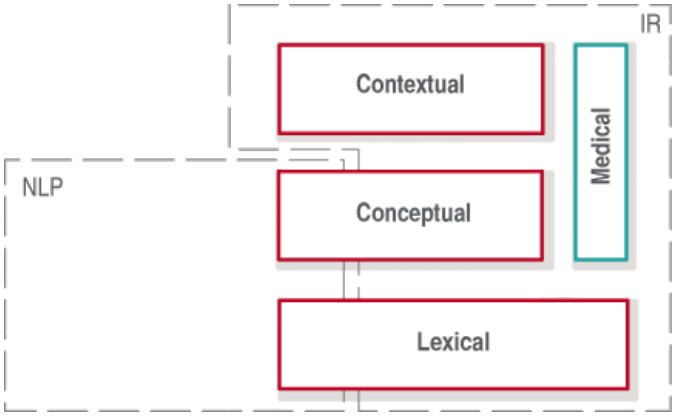

Figure 2.1: The three-stages model.

Each stage (red rectangles) represents a semantical level with increasing explicitness from bottom to top. To illustrate the boundaries of the three stages, we analyse the term `cold`.

On the bottom stage, the *lexcial* stage, terms are not interpreted semantically. This means that the term `cold` is different from the term `low temperature`, even though both terms carry the same meaning. Two terms are only considered to be identical if they have the same spelling. This also implies that, `cold` as in "I'm feeling cold", and `cold` as in "The patient has a cold", are considered to be identical even though the meaning is different in each of the sentences.

At the second stage, the *conecptual* stage, the term `cold` is either interpreted as the concept of (1) low temperature, or of (2) the human sensation (a person feels cold), or of (3) the common

cold, the viral infectious disease, and therefore terms that might have the same spelling can be different with regard to their intended meaning. At the conceptual stage, terms are considered to be identical when they designate the same concept. In fact the spelling is irrelevant. For example, if the term `cold` stands for the disease in a sentence, then `nasopharyngitis` would be an identical term, since it is just the Latin spelling of the same concept.

At the top stage, the *contextual* stage, again terms are interpreted semantically as concepts, but concepts are also related to a certain context. A term is represented as a concept with certain characteristics with regard to specific attributes. For example, the concept *person* can be interpreted in the context of gender, which means that on the contextual stage, the term `person` is interpreted as the concept *person*, and additionally annotated as either male or female.

A mathematical description of the three stages is best explained with a set $\mathcal{T}$ of terms, and an equivalence relation $\sim_n$, where n denotes the stage. Each stage has a different interpretation of *equvivalent*, where $\sim_1$ denotes equivalence at the lexical stage, $\sim_2$ denotes equivalence at the conceptual stage and $\sim_3$ denotes equivalence at the contextual stage. Given two terms $a, b \in \mathcal{T}$ then

- $a \sim_1 b$, if a is composed of the same sequence of characters as b,
- $a \sim_2 b$, if a designates the same concept or idea as b does,
- $a \sim_3 b$, if $a \sim_2 b$, and a and b represent instances with attributes, which fulfill certain conditions.

Under the premise that we retrieve documents by matching terms of a query with terms of documents: according to our model, an IR method can operate at (1) a lexical, (2) a conceptual, and at (3) a contextual stage. The three-stages model distinguishes methods on the basis of data interpretation, whereas specific computational steps of the algorithms are less important. We briefly describe the motivations behind the methods that operate at one of the three stages.

Lexical stage: This stage represents approaches, which Hersh et al. [34, p. 303] summarized as *Lexical-statisticial*. IR methods at this stage form the basis of many IR systems. Many of them are state of the art and technically mature. However, IR methods at this stage ignore the problem of lexical ambiguity [43], which leads to (1) lower precision, because of false positive labeled documents that use the query terms in a different sense than the user, and (2) lower recall, because of false negative labeled documents that contain synonyms of the query term and therefore are incorrectly not retrieved.

Conceptual stage: This stage encompasses techniques that try to circumvent the problem of lexical ambiguity. They introduce an additional layer on top of the lexical representations of terms. Terms are then mapped to unique concepts within this layer. This mapping is then used to improve precision and recall of a pure lexical approach, by either disambiguating terms or by expanding query terms with synonyms. The additional layer adds complexity to a system. This additional complexity can be dealt with by focusing only on a specific domain, for example medicine. Figure 2.1 depicts this focus with the green rectangle. Within the domain of medicine, there exists a variety of *controlled vocabularies* which can serve as the conceptual layer. We will present them in more detail in Section 2.2.

Contextual stage: Methods at this stage use additional data in combination with the pure textual data of the search query and the document collection. One example of this is assigning a geographic scope to a document, which makes them only relevant within the assigned scope. The geographic location of the user can be used then as an additional criteria when retrieving documents [90]. As we mentioned earlier, a term a and a term b are only equivalent at the contextual stage $a \sim_3 b$ if both terms refer to instances of the same concept which fulfill a certain condition. For example, if a user of a location sensitive IR system issues the query `restaurant`, then it is not sufficient that a document contains the term `restaurant` and both designate the same concept. The concepts must represent instances that are equal in the context of geographic location. This means that the term in the documents must refer to a restaurant whose address has the same zip-code as the users current location.

The three-stages model was inspired by the *Stufenmodell* (German for *stages model*), which was presented by Spree et al. [106]. Figure 2.1 also shows dashed rectangles that sketch out the broad disciplines of information retrieval (IR) and natural language processing (NLP). We want to point out that many approaches that we present make use of techniques that are also related to NLP, which is symbolized by the overlapping of the red rectangles in Fig 2.1.

The sections of this chapter are arranged along the three-stages model. In general, the methods at the lexical stage can be considered state of the art, their effectiveness has been proven already, and various production systems make use of them. However, the approaches presented as conceptual or contextual are considered only to be related work, and many of them are still a topic of current research.

2.1 Lexical Stage

A text document is a collection of words in a specific order. The order of words has to follow certain rules (grammar) to build proper sentences and therefore provide meaning. If all the words of a document are taken out of their sentences and put into a bag so that the original order can not be restored, one might think that all useful information is lost. This is not the case from an IR perspective. The fact that a term occurred in a document tells us something about this document. For example, if the term `hypertension` appeared in a document, chances are that this document provides information about hypertension. The exact meaning can not be determined, however many IR models assume that a matched term is a hint for relevance. These heuristic approaches can produce good results, but fail in some cases. Given the query `causes of hypertension` a document with the sentence „This document excludes causes of hypertension“ would be a false positive match. However we assume that the correct matches outweigh the false ones. Upon this hypothesis, many IR methods create a so called *inverted index*, sometimes referred to as *posting file* or *inverted file*.

An inverted index maps terms to document identifiers. The inverted index holds one entry for each term that occurs in one or more documents of the document collection, and each entry points to a list of identifiers of the documents, in which that term occurs. In order to find all documents in which the term `hypertension` occurs, an IR algorithm looks up the entry for this term in the inverted index and reads out the corresponding list of document identifiers. This approach is much faster than scanning through all documents looking for that term.

term	Document IDs
⋮	⋮
hypertension	3, 39, 46
hyperventilation	3, 20, 21
⋮	⋮
the	1, 2, 3, 4, 5 ...
⋮	⋮

Table 2.1: An excerpt of a possible inverted index.

Given that a user issued a query `hypertension`, the logical response of the IR system would be a list of references to the documents where identifiers are enumerated in the list of the index entry `hypertension`. A disadvantage of this simple approach is that the results are not ranked. A user profits immensely if the first suggested hit is also the most relevant [80]. One common approach to impose order is to calculate a similarity measure between a query and a document, then order the results by decreasing similarity. A widely adopted and efficient method to quantify similarity is to represent documents and queries as vectors in a vector space, this was first introduced by the *SMART* IR system at Cornell [85].

Vector Space Model

Given that the inverted index of a document collection holds $|\mathcal{V}|$ different terms, then each document of that collection can be represented by an $|\mathcal{V}|$-dimensional vector $\vec{d}$. Each dimension refers to a term, and depending on how often that term occurs in that document, that dimension's coordinate is higher or lower. A query can be represented in the same manner, as an $|\mathcal{V}|$-dimensional vector $\vec{q}$ from the same vector space.

The cosine of the angle between document vector and the query vector can then be used as a similarity measure, referred to as the *cosine similarity* [50, p. 121].

$$\cos(\theta) = \frac{\vec{q} \cdot \vec{d}}{||\vec{q}|| \cdot ||\vec{d}||} \tag{2.1}$$

The cosine similarity has an advantage over simple vector differences as it more stable when the similarity of two vectors of different magnitude is calculated, which is actually the common case, since a query tends to be very short and a document is often longer.

The vector space model (VSM) is a convenient framework to produce similarity scores between documents and queries. To put the VSM into practice, one needs to provide the values for the individual coordinates. As noted before, one can use the term frequency directly, but another approach that proved to be very efficient is to weigh the term frequency with its *inverse document frequency* (IDF).

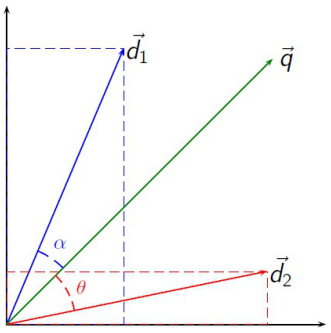

Figure 2.2: Vector Space Model: Two documents, $\vec{d_1}$ and $\vec{d_2}$, and the query vector $\vec{q}$, represented in a term vector space. The angle α between $\vec{d_1}$ and $\vec{q}$, is smaller than the angle θ between $\vec{d_2}$ and $\vec{q}$, therefore document d_1 is more similar to the query then document d_2 [112].

TF · IDF

The term frequency is a good estimator of how much a document covers information that is related to the term. However, some terms are more specific than others. If a term occurs very seldom in a document collection, then it carries more information then terms that appear in almost every document. This property of a term can be expressed by its *inverse document frequency* (IDF):

$$\text{idf}(t) = \log \frac{|\mathcal{D}|}{\text{df}(t)} + 1 \tag{2.2}$$

where $|\mathcal{D}|$ denotes the number of documents in a collection, and $\text{df}(t)$ denotes the number of documents in which term t occurs. The IDF of a term should therefore be taken into account when weighing terms. Referring to the IDF, Salton and McGill introduced the notion of *discrimination value* which expresses how well a term separates relevant from non relevant documents [87]. The combination with the term frequency TF provides us with a powerful term weight which is used to build the coordinates in a VSM within the default scoring method of the Apache Lucene Project [3].

Probabilistic Models

Only the user can decide with complete certainty if a document is relevant to the user's information need or not. An IR system can only make an informed guess. In order to handle these uncertainties it makes sense to analyze the IR problem within principled methods of probability theory and its well established foundation. Inspired by Shannon's work on information theory [88] and its probabilistic approach, Maron and Kuhns introduced probabilistic reasoning to IR in 1960 [52].

Central to probabilistic approaches is to model relevance as an event with a probability mass that is estimated based on available data and statistics. In terms of probability theory, the IR problem can be described with a random variable R which takes on the value 1 if the document

is relevant and 0 if the document d is irrelevant with respect to a query q. In the same vein, we can model the user formulating a query and the drawing of a document from the document collection as events of the probability space. Hence the probability that a document d that was retrieved for query q is relevant, is denoted with:

$$p(R = 1 \mid d, q). \tag{2.3}$$

An IR model that estimates probabilities of relevance and returns a set of documents ordered by decreasing probabilities follows the *Probability Ranking Principle* (PRP).

> „If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of those data [78]."

The PRP has practical implications as it allows us to simplify a model as long as the order of a result is preserved. This means that exact probabilities do not need to be calculated. A practical application is the *Binary Independence Model* [79], which calculates a *Retrieval Status Value* (RSV) based on $p(R = 1 \mid d, q)$:

$$RSV(d, q) = \sum_{t:\ t \in d, t \in q} \log \frac{p_t(1 - u_t)}{u_t(1 - p_t)}. \tag{2.4}$$

The RSV of a document d for a given query q is basically the sum of a fraction over all terms of the query that also occur in the document at least once. The fraction is composed of p_t, which denotes the probability of a term t to appear in a relevant document, and of u_t, which denotes the probability of a term t to appear in a non-relevant document. Manning et al. [50, p. 224-225] show how the RSV is derived so that the following inequality is fulfilled:

$$RSV(d_1, q) \geq RSV(d_2, q), \text{ if } p(R = 1 \mid d_1, q) \geq p(R = 1 \mid d_2, q), \forall d_1, d_2 \in \mathcal{D}. \tag{2.5}$$

The RSV is therefore order preserving and does not violate the PRP. But to calculate the RSV, the probabilities p_t and u_t have to be estimated first.

Based on the assumption that the fraction of relevant documents for any given query is very small, we can estimate the probability of a term appearing in a non relevant document with its overall *relative frequency* in the whole collection [94].

$$\hat{u_t} = \frac{\mathrm{df}(t)}{|\mathcal{D}|}. \tag{2.6}$$

An estimation for the probability of a term appearing in a relevant document is not as clear. However, Croft and Harper [20] proposed a constant estimate to begin with:

$$\hat{p_t} = c, \tag{2.7}$$

where c is for example 0.5. Greiff [30] refined this approach based on empirical observations and proposed to estimate p_t in relation to $\mathrm{df}(t)$:

$$\hat{p_t} = \frac{1}{3} + \frac{\frac{2}{3}\mathrm{df}(t)}{|\mathcal{D}|}. \tag{2.8}$$

The accuracy of the probability estimations can be increased when the IR system allows the user to provide *relevance feedback* for an initial set of retrieved documents. This information can be fed to a *Baysian update process*, which is a standard tool within *Bayasian statistics*. This extension is a direct benefit of expressing the IR problem in terms of probability theory as it opens it to a whole set of established methods and techniques.

However, the Binary Independence Model does not consider term frequency and document length, which are important quantities when the document collection is heterogeneous with regard to these attributes. The *Okapi BM25* [81] approach builds upon a probabilistic basis, but also takes term frequency and document length into account:

$$RSV(d,q) = \sum_{t \in q} \log \left[\frac{|\mathcal{D}|}{\mathrm{df}(t)} \right] \cdot \frac{(k_1+1)\mathrm{tf}(t,d)}{k_1 \left((1-b) + b \times (l(d)/L_{ave})\right) \mathrm{tf}(t,d)}, \tag{2.9}$$

where $\mathrm{tf}(t,d)$ denotes the term frequency, $l(d)$ denotes the document length, L_{ave} denotes the average document length of the whole collection, k_1 and b are tunable parameters. This formula is just a variant, and the BM25 weighting is actually a family of several weighting schemes. The BM25 approaches showed to be a successful advancement of the probabilistic IR approach. Spärck Jones et al. summarized the basics and evaluation results in their paper [94, 95]. The Apache Lucence [3] implements the variant which was introduced by Robertson et al. at the TREC-3 conference [81].

Language Modeling Approach

First introduced by Ponte and Croft [72] in 1998, the language modeling approach in IR has been studied and applied in many ways [10, 38, 55, 57, 119].

The general idea is that a user generates a query by thinking of words that are likely to appear in a document that is relevant to his or her information need. An IR system then tries to model this process with a *statistical generative language model* (LM). It does not create a new model, instead it considers only the LMs it had already created, one for each document. The more likely it is that an LM generates the query, the higher the ranking of the document that the LM is based on.

The language model M_d of a document d is the probabilistic model that assigns a probability to an ordered list of terms $q = t_1 t_2 .. t_n$. This probability defines the likelihood of this sequence to be generated by this language model:

$$p(q|M_d). \tag{2.10}$$

For example, if the language model M_d represents the language of a medical journal article about the causes of hypertension, than the sequence `pressure of blood` has a higher probability to be generated by this language model then the sequence `paris during spring`:

$$p(\texttt{pressure of blood} \mid M_d) > p(\texttt{paris during spring} \mid M_d) \tag{2.11}$$

The core steps of the LM approach can be summarized as:

1. estimate a language model M_d for each document d of a collection $\mathcal{D}$,
2. calculate the probability of the query being generated by a document's language model for each document $p(q \mid M_d)$,
3. rank the documents according the calculated probabilities of step 2.

From this summary it is obvious that LM approaches have a common ground with the probabilistic approaches, such as the Binary Independence Model that we just presented. However, the pure probabilistic approach models relevance directly. Whereas, as mentioned earlier, the LM approach is based on the assumption that the query and a relevant document are generated by the same LM, which happens to be a probabilistic model. Another difference to general probabilistic approaches is that the term order is essential in language models [38]. Other disciplines, like NLP, refer to *n-grams*, which are the basis for statistical language models.

An n-gram is an ordered sequence of n terms. An LM that builds upon 3-grams, for example, predicts the next term of the sequence by taking into account the two previous terms. An LM based on 4-grams, takes the previous three terms into account and so on. Language models based on n-grams of higher order are more complex than LMs based n-grams of lower order. This is one of the reasons why we will present only LMs that are based on a 1-gram (unigram models). Unigram language models predict the next term in a sequence completely independent of the previous terms. Language models based on unigrams are easier to estimate, and models of higher order reach their limits quickly due to data sparseness [50, p. 241]. Nonetheless, some authors have also investigated language models based on bi-grams and tri-grams [57,92]. Since unigram models ignore any dependencies between terms, as does the Binary Independence Model, the main motivation of expressing the documents and queries as generated by LMs, seems actually to be of no concern.

However, by looking at IR as an LM problem, it is linked with the field of speech recognition and natural language processing. Consequently, this view of IR gets access to advancements and techniques that had been made in those disciplines. We will present some *smoothing* methods when estimating LMs, which had been originally proposed in the context of the speech recognition tasks [16].

The first step is to estimate a language model for each document based on the content. The maximum-likelihood estimation (MLE) for a single term t is given by

$$\hat{p}(t \mid M_d) = \frac{\mathrm{tf}(t,d)}{\sum_{t' \in \mathcal{V}} \mathrm{tf}(t',d)}, \tag{2.12}$$

where $\mathcal{V}$, the vocabulary of the collection, is the set of all terms. As we focus only on LMs based on unigrams, the occurrence of terms in a sequence are independent of each other. Therefore, the MLE of the probability of a query $q = t_1 t_2 t_3 .. t_n$ is:

$$\hat{p}(q \mid M_d) = \prod_{i=1}^{n} \hat{p}(t_i \mid M_d). \tag{2.13}$$

However, the text of a document is thought to be only a small sample of its actual language. Due to this data sparseness the MLE of terms that are not present in the document are zero probabilities. Consequently, a query that contains terms that are not present in a document has a zero probability to be generated by the document's language model. On the other hand, if a term only occurs once in a document, it is likely due to chance and therefore the MLE assigns a probability to this term which is too high. These inaccuracies can be reduced by various *smoothing* techniques. Zhai and Lafferty analysed and compared various smoothing approaches [119]. Two of them, which we present briefly, are implemented in the Apache Lucene project [3].

The *Jelinek-Mercer* method is a linear interpolation of the MLE regarding the document as the data sample and the MLE based on the whole collection [40].

$$\hat{p_\lambda}(t \mid d) = (1 - \lambda)\hat{p}(t \mid d) + \lambda \hat{p}(t \mid \mathcal{D}). \tag{2.14}$$

The other smoothing approach uses *Bayesian smoothing with Dirichlet priors*. Due to the independence assumption between terms, we can refer to the language model as a multinomial distribution. Hence, a Bayesian update of Dirichlet priors seems practical. The parameters of the Dirichlet distribution are proportional to a parameter μ which can be tuned and adapted for different collections:

$$\left(\mu \hat{p}(q_1 \mid \mathcal{D}), \mu \hat{p}(q_2 \mid \mathcal{D}), .., \mu \hat{p}(q_n \mid \mathcal{D})\right). \tag{2.15}$$

With this conjugate prior distribution, the smoothed LM estimate is given by:

$$\hat{p}(t \mid d) = \frac{\text{tf}(t, d) + \mu \hat{p}(t \mid \mathcal{D})}{\sum_{t' \in \mathcal{V}} \text{tf}(t', d) + \mu}. \tag{2.16}$$

2.2 Conceptual Stage

Ultimately, terms of human languages stand for concepts of the real world. Often, a concept can be designated by multiple terms within the same language, which we refer to as *synonyms*. In the other direction, one term can designate multiple concepts, we refer to this as a *homonym*. One can think of `rock`, which can refer to a solid stone or a music genre. These ambiguities are no issue in day to day communication between two humans as the intended meaning can be derived from the context in which the terms appear. However, in a computer memory, `rock`, as in *a fossil rock*, and `rock`, as in *rock music*, map to the same character encoding. In IR, the lexical processing of terms and its inherent lexical ambiguity pose a well documented problem [43].

In Figure 2.3, the problem of ambiguity is illustrated with two different users having two different information needs, but both users issue the same query. Because the query carries a

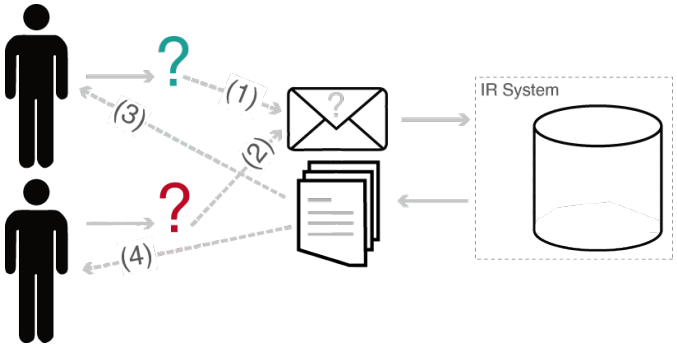

Figure 2.3: Lexical Ambiguity.

homonym it is not sufficiently precise, denoted by the dashed lines (1) and (2). The IR system responds with results that contains documents using the query term in one sense, but also documents that use the term in another sense. Therefore, the result is suboptimal for both users, denoted with the dashed lines (3) and (4).

A logical approach would be to process terms at a higher level. A query term is mapped first to its semantic concept and ideally an IR system responds with documents that contains terms with the same semantic meaning. But to resolve lexical ambiguity is not the only motivation: a concept mapping enables incorporation of background data of any kind related to real world concepts. The main goal of this thesis is to incorporate epidemiological statistics, therefore a conceptual layer composed of medical concepts is essential.

We present a brief overview of *controlled vocabualries*. As we mentioned earlier, a controlled vocabulary can serve as a conceptual layer. We explain their original purpose and application, and we present a few examples.

Controlled Vocabularies

A controlled vocabulary is a set of certified terms that is curated by an institution or organization. Each term is mapped to a unique concept of the real world. A controlled vocabulary might denote synonyms for a term, but there exists only one canonical, preferred term, for each concept. Some controlled vocabularies carry a hierarchical structure in which more specific concepts are subordinated to more general concepts. A hierarchical controlled vocabulary is also called a *Thesaurus*.

The *Medical Subject Headings*, commonly known by the acronym MeSH, is an example for a thesaurus. The main purpose of the MeSH thesaurus is to manually index medical journal articles and books that are collected and archived in the MEDLINE database. Each entry in the database is manually indexed by a human, who uses only terms that are part of the MeSH thesaurus. A user that queries the database can therefore rely on high recall when searching the MEDLINE database by a MeSH term.

A small part of the MeSH thesaurus is illustrated in Figure 2.4. The concept `Hypertension` is a child of the concept `Vascular Diseases`, which is a child of `Cardiovascular Diseases` and so on. It is also encoded with a unique address: `C14.907.489`. A certified

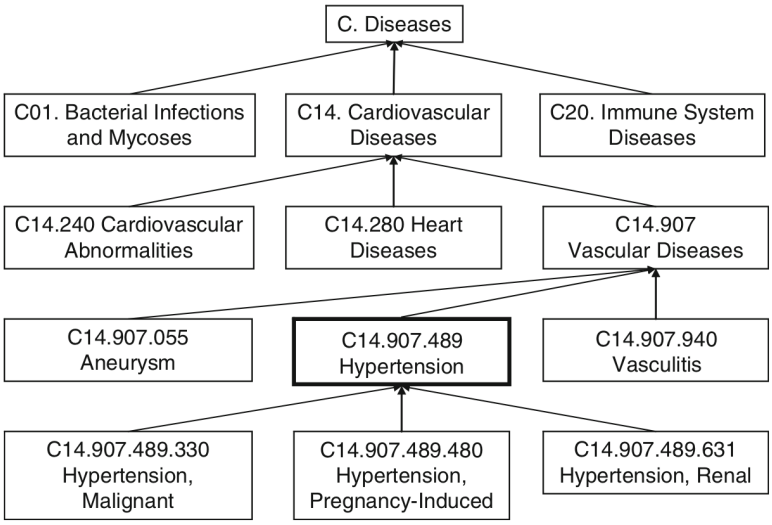

Figure 2.4: The concept `Hypertension` (box in the center), embedded in the MeSH thesaurus [34].

synonym of hypertension is `Blood Pressure, High`, which can be looked up in the MeSH Browser [63].

Pure lexical representation are too ambiguous in the biomedical domain, which makes controlled vocabularies especially useful [62]. Other examples of controlled vocabularies are:

- the International Classification of Diseases (ICD), which is published by the World Health Organization [114],
- Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT), which is now maintained by the International Health Terminology Standards Development Organisation [115],
- and Gene Ontology (GO), which is provides a terminology for molecular biological concepts [27].

These are just a few examples of controlled vocabularies. Since several different vocabularies within the medical domain exist, many of them designed for specific sub-domains, another project was initiated with the goal to link matching concepts between said vocabularies [21]. The *Unified Medical Language System* (UMLS) includes a *Metathesaurus* that enables one to translate a concept from one vocabulary to another one. Figure 2.5 illustrates the linking between controlled vocabularies. We will present a practical application of the UMLS Metathesaurus in Chapter 5, where we present the prototype implementation of the PPHS.

Concept Mapping

This subsection gives an overview of approaches that deal with the problem of recognizing the intended concepts of free text. There seem to exist two general perspectives: (1) The *linguistic*

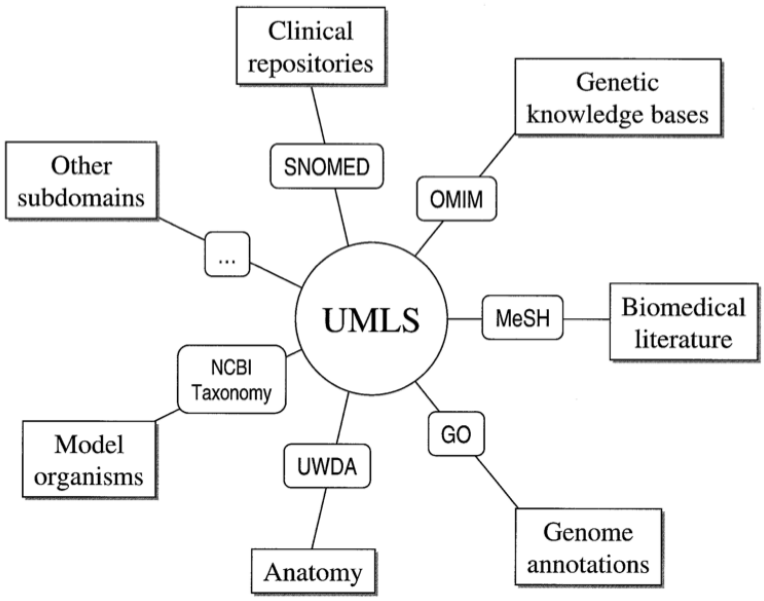

Figure 2.5: The **UMLS** and part of the integrated controlled vocabularies and their attributed sub domains [11].

approach parses an input text and maps words, or phrases, to corresponding concepts. (2) The *text categorization* approach annotates bigger units of text, even whole documents, with one or more concepts from a predefined set of concepts. This makes it essentially a classification problem. The two main differences of both approaches are the scale at which concepts are recognized, and that the linguistic approach uses a lexicon as a knowledge base, whereas text categorization is a typical supervised machine learning approach based on training with training data. The linguistic approach ideally translates term for term to the intended concept, and the text categorization approach assigns text to a more general topic concept.

Linguistic approaches: One of the linguistic approaches is *MetaMap* [5], which is a widely used program that recognizes concepts from the *UMLS Metathesaurus*. Figure 2.6 illustrates the processing steps of *MetaMap*. The input is processed lexically in conjunction with the *SPECIALIST* lexicon[1] [64]. The output of these steps are noun phrases used to generate variants of them. A variant can be a different spelling of the noun phrase, or its acronym notation, or a morphological deviation of the noun phrase. The following step identifies candidates of the UMLS Metathesaurus strings that match the noun phrase and its variants generated by the previous step. The candidates are then combined, and the best matching combination produces the result. Optionally, word sense disambiguation is possible where the mappings are tested within their context and semantically sound mappings are then preferred.

Other linguistic mapping tools are *MicroMeSH* [48], *CHARTLINE* [58], *CLARIT* [24] and *SAPHIRE* [36]. However, we found in our review that in most cases *MetaMap* was chosen to

[1] part of the SPECIALIST NLP System

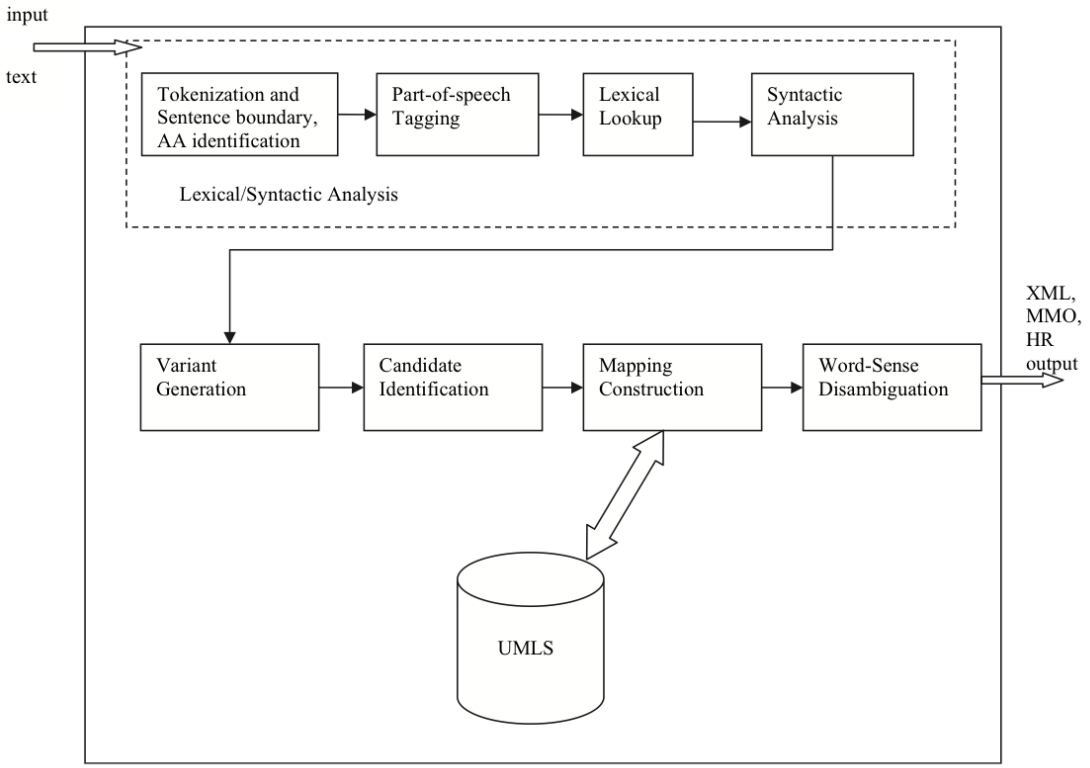

Figure 2.6: etaMap System Diagram [7].

map biomedical text to concepts. Aronson and Lang summarized the development of *MetaMap* in a recent paper [7].

Text categorization approaches: Text categorization or *text classification* is actually a subfield to IR which covers the problem of assigning a piece of text to one or more categories of a predefined set of categories. Indeed, it is just another way of looking at the problem of concept mapping, if said categories are the concepts. As it is presented as a classification problem, the methods used are often supervised machine learning algorithms that train a classifier based on manually annotated training data. Trieschnigg et al. compared 6 different approaches in labeling biomedical text with MeSH concepts [105]. The best performing method was a *K-Nearest Neighbor* (KNN) classifier, which was shown already by Yang to be an efficient approach in a general text categorization evaluation [117]. Its implementation was based on an LM retrieval system, which indexed citations that had already been been annotated manually with MeSH concepts. The text, to classify, is then used to query the retrieval system, and the top-k results form the k-nearest neighborhood. The text is then classified as the most frequent MeSH concept of the neighbors. The prototype implementation of the PPHS performs a similar approach, but instead of MeSH concepts, it uses ICD-9-CM concepts. Yang and Chute investigated a regression based learning method that categorized surgical reports into ICD-9-CM categories. This approach outperformed simple lexical matching significantly, and showed machine learning based

approaches to be effective [118]. Larkey and Croft also investigated the problem of text categorization regarding ICD-9-CM concepts and tested 3 types of classifiers: a KNN approach, relevance feedback and *Baysian Indepence classifiers*. These methods were tested in isolation but also in combinations with each other. The combination of different classifiers showed better results than any single classifier [44].

Concept-based IR

After identifying medical concepts in free text, various authors developed techniques that either replaced pure lexical IR, or aimed to improve lexical approaches by combining them with concept-based techniques. In 1992 Hersh et al. [37] compared IR performance of concept-based indexing and lexical indexing. The retrieval system named *SAPHIRE* indexed medical documents by concepts of the UMLS Metathesaurus. The concepts were extracted by a non-syntactic pattern matching algorithm. The queries were also transformed into their conceptual representation and the documents were then retrieved by a standard lexical IR method operating on the concept-based index. The result of the study showed no significant improvement over a standard lexical index.

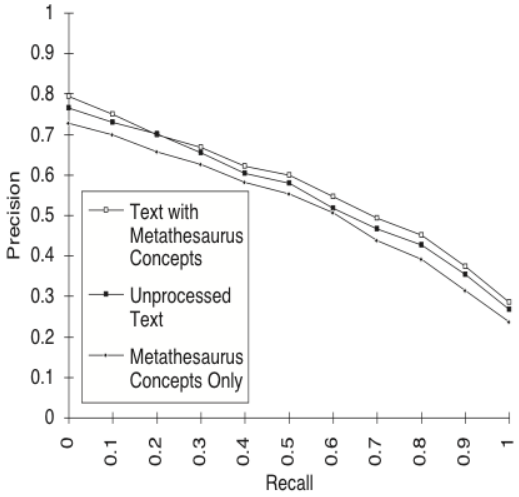

Figure 2.7: Evaluation results of Aronson et al. and their approach of indexing biomedical text, enhanced with recognized UMLS Metathesaurus concepts [8].

In a similar approach, Aronson et al. created enhanced versions of biomedical documents by replacing phrases that were recognized as UMLS Metathesaurus concepts, with their canonical terms. Unmatched text parts were left untouched. They were named *surrogate texts* by the authors, and were indexed with a VSM, used in the *SMART* retrieval system [86]. The concept extraction was done by the same principles as *MetaMap*, which actually was initiated by this work [7]. Figure 2.7 shows results of evaluations on a test collection of 3000 documents and 150 queries. The index of *surrogate texts* performed best and showed modest improvements.

Nadkarni et al. posed the question if concept indexing is production ready and conducted a study. The authors developed a concept indexing algorithm and examined manually the output of the algorithm after processing a set of 12 discharge summaries and 12 surgical notes. They reported 76.3% true positive matches and argued that this rate is too low to be feasibly in a production environment [61].

As pure concept-based approaches did not produce significant improvements, some authors pursued a combination of concept-based and standard lexical-based methods [46, 96, 104]. Limsopatham et al. adapted the approach of Srinivasan [96] and investigated the linear combination of two scores [45].

$$s(d,q) = \lambda_q \cdot s_{BoW}(d,q) + (1-\lambda_q) \cdot s_{BoC}(d,q), \tag{2.17}$$

where $\lambda_q (0 \leq \lambda_q \leq 1)$ is query-dependent parameter that is estimated. $s_{BoW}(d,q)$ denotes the score of a *Bag of Words* model, which we refer to as a lexical model, and $s_{BoC}(d,q)$ denotes the score retrieved from a *Bag of Concepts* model, that we refer to as conceptual, or concept-based approach. Based on this linear combination with unknown λ_q, Limsopatham proposed a classic machine learning approach in order to produce a model that can predict an optimal λ_q setting for unseen queries. The model was trained by *Gradient Boosted Regression Trees* [77] on a set of query features like query length and others. The query-dependent lexical and concept-based score combination method showed to be effective on the TREC 2011 and 2012 Medical Record track [109, 110].

Aronson and Rindflesch adapted *Query Expansion* for UMLS Metathesaurus concepts. In this approach, *MetaMap* identifies concepts from the raw original query and expands the query with these concepts. The results were promising and showed a 14.1 % increase of the 11 point average precision over baseline [6]. We already mentioned the work of Trieschnigg et al. regarding text categorization. Their work also included follow up experiments which investigated the performance of IR systems . The queries were categorized with the text categorization approaches and subsequently expanded with the identified MeSH concepts. The retrieval performance showed significant improvements over baseline [105]. In a similar study, Hersh et al. , expanded queries manually with concepts of the UMLS Metathesaurus, that were suggested by the SAPHIRE program, mentioned earlier [35]. However, Hersh et al. did not report any significant improvements by expanding queries with concepts.

Moskovitch et al. presented *Vaidurya*, a concept-based search engine. In contrast to the concept-based approaches presented so far, concepts are not extracted automatically from a free text query. The user can select concepts optionally in a addition to a free text query. These concepts can be combined by logical AND and OR operators [59, 60]. This approach was evaluated on a repository of clinical guidelines and demonstrated improvements over free text queries, especially when the baseline precision was low.

Zuccon et al. took on the problem of *granularity mismatch*, when employing concept-based retrieval methods. A granularity mismatch is given when a query is composed of a general concept like `Opiate`, and therefore documents that are indexed with more specific concepts, specific types of opiate in this case, are not retrieved. Zuccon et al. implemented a concept-based index of documents. The original texts were mapped to the SNOMED-CT ontology with MetaMap. The relevance scores were calculated by a combination of weights retrieved by

matching the query concepts and by weighting additionally the children concepts of the query concepts. However, this initial attempt did not show any significant improvements over baseline [120].

In 1996, Cimino presented a review of systems that integrate clinical information systems with bibliographic resources and other knowledge resources. The common goal of the reviewed systems was to make use of available clinical data (patient specific data, for example age and sex) when a clinician requests information. One can think that clinical data provides the same context as the PPHS model uses, and therefore these systems are similar in their approach. This is only true to a limited extent. As we explained earlier in this chapter, we differentiate IR systems in respect to the semantic complexity of the data representations they operate on. The systems, reviewed by Cimino, integrate clinical context on a higher level than the PPHS model does. They treat the IR systems as black-boxes, so the data representations were not enhanced. In general, the approaches implemented modules that processed clinical data from clinical information systems and generated augmented queries for IR systems. For example, one system allowed one to select text portions from a patient report, which were then translated into MeSH terms. The user selects the appropriate terms to query an IR system for information on that specific MeSH terms. Therefore, we classified these early context driven systems as conceptual and not contextual according to our three-stages model. At that time, the systems were in a prototype state and thorough evaluation was still missing [18].

2.3 Contextual Stage

As earlier explained, our definition of a contextual approach is that terms at this stage are considered equivalent, if (1) they represent semantically the same concept, and (2) if they refer to instances with equal characteristics.

This definition has some implications when we qualify IR approaches for being contextual. Several authors presented approaches as contextual but not necessarily in the same vein as we understand contextual. The TREC Session track promotes methods that leverage information from previous search sessions [41]. For example, Cheng et al. [17] refer to a *context-aware ranking principle* by taking previous queries from the same search session into account. However session context does not refer to any specific concepts of the real world.

A positive example of our definition is the approach of Lu et al. that define the context to be the user's geographic location. The authors explain their approach with the query `laundry service`. It has an implicit local intent which means that the user is primarily interested in laundry services nearby [49]. Under our definition, the terms of this query are only equivalent to terms in documents, if they refer to specific instances of the same concept that fulfil a criteria. In this approach, the criteria is the geographic location. This means that only laundry services nearby the user's location are considered equivalent to the laundry service mentioned in the query. However, it needs to be pointed out that the implementation does not reflect the context (geographic location) on the level of single terms. The terms are not mapped to its intended concept, and that concept is then annotated with geographical information. One of the presented

methods of Lu et al. re-ranks top-k results by the following equation:

$$s^r(q, d, l) = s(q, d) + \sum_i w_i I_i(d, l), \tag{2.18}$$

where $s(q, d)$ denotes the original ranking score, and $I_i : \mathcal{D} \times \mathcal{L} \rightarrow \{0, 1\}$ denotes the indicator function, which indicates if a section i of a document refers to a location $l \in \mathcal{L}$. The weight w_i is estimated by a supervised learning algorithm. We can see that whole document sections are put into a geographical context and not single identified concepts. A term, such as `laundry service`, would inherit its context from the document section in which it occurred. Implicitly, the concrete instance of the concept gets the same location attribute as the document section itself. The attribute can be thought of as being pushed down to the single concept level. Obviously, in this approach occurs no mapping of the lexical representation, `laundry service` to its actual concept. Only geographic terms, such as city names, are identified and mapped to concepts. However, the ideal case where a term is mapped to its concept, and then the concept is annotated with contextual attributes, is not realized in any of the reviewed works.

The PPHS approach, which is present in this thesis, approximates the ideal, too. The context is composed of the health factors age and sex. This means that concepts, such as symptoms, are ideally only considered equivalent, if they are experienced by people of the same age and sex. However, the PPHS model also puts whole documents into the context of a medical condition and their relationship to the health factors sex and age. The terms of a document inherit the context of that medical condition. We further illustrate this approach by a concrete example: A female user issues the query `burning urination`. A document which covers symptoms of prostate cancer is likely to contain these query terms. Ideally, the term `burning urination` would be mapped to the concept of a controlled vocabulary that represents the symptom and then limited to instances where the symptom is experienced by male patients (since only men can suffer from prostate cancer). However, for practical reasons, only the complete document is mapped to the concept of prostate cancer instead of the symptom. Epidemiological statistics provide the context. It limits the disease to male patients. Therefore, the term `burning urination` inherits the context and is also implicitly limited to male persons. Consequently, the document does not match the query, since we assume a female user.

The probabilistic model for personalizing web search of Sontag et al. [93] is also considered to be operating at the contextual stage. Their model actually inspired our PPHS approach. The domain is different, though. Their model focuses on general web search. The context is composed of a user profile. It models the user's interest in various topics, such as Sport, Arts and Computer Science. Again, whole documents are mapped to concepts. In their work they used nodes of the ontology generated by the *Open Directory Project* [68]. We examine one of the models by Sontag et al. in more detail in Section 4.2.

Personalization of search results is the main motivation behind methods at the contextual stage. In fact most of the approaches, that we have examined, were all labeled as personalizing methods. We named our approach Probabilistic *Personalized* Health Search, too. Pitkow et al. presented the *Outridge Personalized Search System* in 2002 and introduced personalization to IR. The authors described it as a combination of *contextualization* and *individualization*. The main idea is to look at relevance in relation to the user instead of the census [71]. The authors

explained that individualization encompasses past information-seeking behaviour. Various authors continued to incorporate short and long-term search histories in order to build user profiles for personalization [22, 53, 89, 100, 101]. However, these approaches focus on building a profile for users from past searches in order to enhance future search sessions. Our approach, however, does not aim to track a user's behaviour over a time period. It is more concerned with enhancing search results for a demographic group of people based on statistics that are available for this group.

The three-stages model allowed us to classify approaches similar to our approach based on how additional background information acts on the ranking algorithms. We think that by focusing on how concepts are further specified by a context, is what distinguishes it from other contextual, or personalized approaches. Therefore, it became apparent that so far their seem to be no contextual IR methods, covered by academic literature, that incorporates background data from epidemiological statistics. However, the study of White and Horvitz on user behaviour in medical IR noted that the search engine *MSN Health and Fitness* ranks results, covering more common medical conditions, higher than a general search engine [111]. Nonetheless, we could not find any evidence that this search engine enhances its ranking algorithms with the help of epidemiological data.

2.4 Summary

In this chapter we introduced a three-stages model that enabled us to distinguish IR methods based on their semantic complexity. We want to point out that a complete IR system almost always builds on top of the lexical stage. The PPHS model, as well, stretches across all three stages.

We also presented controlled vocabularies. They have a long history in the medical domain. They also seem to be a suitable conceptual layer for IR methods operating on the second and third stage. Under our definition of the third stage, the second stage is a necessary pre-stage. Our review could show that various authors developed IR methods in the medical domain successfully at the second stage. Therefore, it seems feasible for us to develop a medical contextual IR model based on background data from epidemiological statistics. Furthermore, there seem to be no medical IR methods developed at the contextual stage that are discussed by academic literature. This is an additional motivation for us to conduct research in this area.

CHAPTER 3

Epidemiological Background

After thorough observations of statistics and a brilliant process of elimination a Hungarian physician named Ignaz Semmelweis concluded that the cause of a puerperal infection is decomposed "animal-organic" material, which is carried by thr examining fingers of physicians, instruments, bed linen and other things that come into contact with a patient. This discovery was remarkable, since a theory of germs was established only after Semmelweis' death. Without the aid of laborious experiments and knowledge of microbiology at the time, Semmelweis' conclusions were supported only by clinical facts and statistics. During his lifetime, puerperal fever killed many women who had just given birth. Semmelweis worked as the assistant of the head of the first maternity clinic in Vienna. In May 1847, he introduced a programme at the clinic, which demanded that the medical staff wash their hands, and disinfect all sorts of instruments, basins, linens and other things that could potentially have come into contact with the patients. As a consequence, the mortality rate dropped from 18.3% to 2.2% within three months. [116].

Semmelweis is one of the pioneers of a scientific discipline that is nowadays referred to as *epidemiology*. The difference between epidemiology and clinical medicine is primarily the unit of study. Epidemiology is concerned with health and disease conditions in populations, whereas clinical medicine addresses single, individual cases. Rothman et al. provided the following definition [82, p. 33]:

> "Epidemiology is the study of the distribution and determinants of disease frequency in human populations."

The aim of epidemiology is to control relevant health problems in populations based on knowledge inferred from surveillance, observation, screening, analytic research and other studies. In order to control health outcomes and implement prevention programmes, relationships between potential risk factors and diseases are analytically tested for causality. These *analytical studies* are concerned with identifying and measuring the effects of risk factors. In contrast, a *descriptive study* collects data on the occurrence of a disease and its relation to demographic factors such as sex, age, ethnicity, occupation, social class, and geographic location, without testing any causal

relationships [73]. The results of the descriptive studies form the basis on which hypotheses are formulated. These hypotheses are then tested in an analytical study.

This thesis focuses primarily on the descriptive methods, especially the resulting data. The analytical methods of epidemiology are negligible within the scope of PPHS, though our approach incorporates the associations between a disease and demographic characteristics. We are interested to see if a disease is more or less likely for a patient of a certain sex and age. However, these two attributes must not necessarily have a causal relationship with the disease. For example, lung cancer is more common in men than in women. Although the association between sex and lung cancer is apparent in the descriptive studies, a causal connection can not be proclaimed. Public health authorities need to identify the causal risk factors in order to prevent the spread of diseases. Hence, they depend on analytical studies. In the case of lung cancer, studies revealed that the one of the main causes is smoking, a behaviour that is more prevalent among men than women. This explains why the incidence numbers of lung cancer is higher in the male population [99, p. 182-185].

The following section continues with the presentation of basic measures of the occurrence of a medical condition. We will describe what they mean, as well as how they are calculated.

3.1 Measures of Occurrence

In order to estimate the effect of a risk factor on the occurrence of a disease, an epidemiologist needs to express the frequency of a disease, in either relative or absolute terms [82, p. 33-34]. We will describe four basic measures that quantify the distribution of diseases and other medical conditions. First, we briefly discuss the term *case* in the context of epidemiological studies. A case is given, if a certain disease or a particular health condition is found in an individual within the study population. It is necessary to precisely define the criteria (diagnosis, display of symptoms, etc.) that have to be fulfilled, in order to declare a case [73].

Incidence Proportion: The incidence proportion (sometimes referred to as *cumulative incidence*) is the number of new cases in a specified period in relation to the size of the population at risk during that specified time period. We denote the period as an interval $[t_1, t_2]$ and the population at risk as $\mathcal{P}$. Further, we specify the partial function $c\colon \mathcal{P} \to [0, \infty]$ that maps to the onset for individuals who have fallen ill. The function is undefined for individuals, who have not developed the disease. Then, the incidence proportion is defined as:

$$\mathrm{IP}_{t_1,t_2} = \frac{\sum_{x \in \mathcal{P}} 1\left[c(x) \in [t_1, t_2]\right]}{|\mathcal{P}|}, \tag{3.1}$$

where $1\left[c(x) \in [t_1, t_2]\right]$ maps to 1 if the onset is within the time period [1]. The incidence proportion is also a measure of risk [73].

Prevalence: Prevalence is the measure of the occurrence of any health-related factor at a specific point in time (*point prevalence*), or during a given period (*period prevalence*). The considered health factors are not necessarily diseases, but also exposures like smoking or any other health-related condition. An example would be the number of individuals who suffer

[1] Iverson bracket notation: http://en.wikipedia.org/wiki/Iverson_bracket.

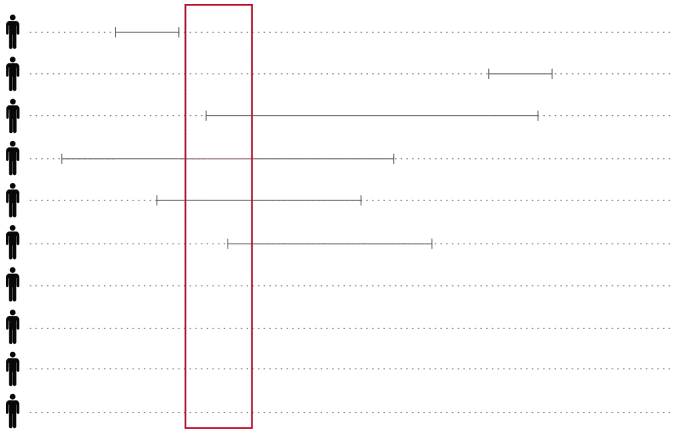

Figure 3.1: Given that the red rectangle specifies the time period of interest, and the solid lines mark the beginning and end of the disease in question: then the period prevalence is $\frac{4}{10}$, whereas the incidence proportion is only $\frac{2}{10}$, since only two new cases occurred.

from hypertension. In order to be meaningful, however, the measure has to refer to a specified population. Furthermore, the prevalence must always refer to a point in time or a time period. [73, 82]. For example, the annual hypertension prevalence is the number of all individuals that suffered from hypertension at one point in time within the specified year. The prevalence of a disease is the proportion:

$$\mathrm{P}_t = \frac{\sum_{x \in \mathcal{P}} 1\left[c(x) \leq t\right] 1\left[r(x) \geq t\right]}{|\mathcal{P}|}, \tag{3.2}$$

where the function, $r\colon \mathcal{P} \to [0, \infty]$, maps to the point of time, when the individual recovered, or ∞, if the individual never developed the disease or has not yet recovered. The point in time for a given point prevalence is denoted as t, and for a period prevalence t refers to the end of that period. Figure 3.1 exemplifies the difference between incidence proportion and prevalence.

Incidence Rate: The incidence rate is the number of onsets C of the disease, divided by the sum of time of persons being at risk. The time of a person, being at risk is denoted by the function $r\colon \mathcal{P} \to [0, t]$, where t denotes the point in time when an individual is followed up the farthermost, then the incidence rate is given as:

$$\mathrm{IR} = \frac{C}{\sum_{x \in \mathcal{P}} r(x)}. \tag{3.3}$$

The incidence rate allows a more accurate interpretation than the incidence proportion, but it is harder to obtain, and usually only measured in controlled experimental studies.

These measures of occurrence serve as the basic description of a population in regard to a disease or health disorder. But a complete review is not within the scope of this thesis. However, the presented measures and terms enable us to analyse and discuss epidemiological data in the upcoming sections.

3.2 Sources of Epidemiological Data

This section reviews potential sources of epidemiological data, that could potentially serve as input at the contextual stage of a PPHS model. We established a list of criteria, which will be discussed in detail.

1. **Machine-readable**: The data needs to be in a format that can be processed by a computer.
2. **Incidence rates**: The data is suitable for estimating incidence rates. As we presented in Definition 3.3, the incident rate is calculated from the number of new cases in relation to the accumulated person-times. The data source needs to provide not only the incidence number, but also how long each person of the population had been at risk. The reasons for preferring incidence rates over the other measures presented earlier are explained in detail in Chapter 4.
3. **Age- and sex-specific**: the data allows age- and sex-specific estimates. The main objective of PPHS is to adapt the search results to the age and sex of a patient. Consequently, it is essential to model the relationship between age and the incidence rate of a disease, stratified by sex.
4. **Completness**: The data allows estimates for all health disorders that are covered by the document collection. Ideally, the incidence rates of each disease covered by a document of the collection can be derived from the data.
5. **Encoding**: The health disorders are encoded with a unique identifier. As mentioned in Chapter 2, Section 2.2, a controlled vocabulary serves as the glue between the IR system and the estimates derived from epidemiological data. In order to establish an unambiguous link between a concept of a disease and its estimated incidence rate, either both, the controlled vocabulary and the data encode diseases with the same code, or there exists a mapping between the encodings.

We analysed five data sources according to our criteria:

- The incidence & prevalence database (IPD) marketed by Thomson Reuters is a commercial product that provides incidence and prevalence numbers for over 4500 diseases and procedures [102].
- The National Hospital Discharge Survey (NHDS) [15], which was conducted annually from 1965 until 2010, by the Centers for Disease Control and Prevention [14], the leading public health organization in the United States. The NHDS is a national probability sample survey of discharges from non-federal hospitals in the United States.
- The data repository of the World Health Organization (WHO) [113] provides data of over 1000 indicators on priority health topics, including mortality and burden of diseases.
- The Statistik Austria (StatAUT) cancer registry publishes incidence and mortality numbers of various types of cancer with respect to the population of Austria [98].

- Finally, the GLOBOCAN project provides estimates of the incidence of major cancer types for 184 countries [25].

Source	machine-readable	incidence rates	age-sex	completeness	encoding
IPD	●	○		○	●
NHDS	●	○	●	○	●
WHO	●	○	○		
StatAUT	●	○	○		●
GLOBOCAN	○	○	○		●

Table 3.1: Data sources evaluated using the PPHS criteria: ● criterion fully met, ○ criterion met partially.

Table 3.1 displays the result of the evaluation of the five data sources, that we examined in detail. Unfortunately, none of them meets all criteria completely. We decided to use the NHDS data set of 2007 to implement the prototype that we used to evaluate PPHS. Chapter 5 explains in detail how we processed the NHDS data set and integrated it into the prototype implementation. We will briefly discuss the characteristics and drawbacks of each data source, individually.

IPD: Lacking a license, we could evaluate the IPD only based on freely accessible samples and the main brochure. The database provides summaries for the most widely searched diseases (according to the vendor) in an Excel spreadsheet format, therefore, the first criterion (machine-readable) is fully satisfied. We could not determine if the set of diseases covers all diseases from the document collections that we used for our evaluations. Since the data set covers more diseases than most other sources, we marked Criterion 4 (completeness) as partially met. The summary sample, which is available for download, includes incidence numbers for various regions, and the size of the populations. This allows a calculation of incidence proportions, but not of the incidence rates (Criterion 2). The brochure also displays a "global incidence & prevalence report" of asthma, which includes age- and sex-specific information. However, the machine-readable format only displays age groups and not sex-specific classifications (Criterion 3). Since the IPD supports queries by ICD-9 codes Criterion 5 (encoding) is fulfilled. The ICD-9 encoding is a version of the International Classification of Diseases (ICD) [114], which is a suitable controlled vocabulary, serving as a conceptual layer.

NHDS: The data produced by the NHDS is available on an FTP server as a text file [15]. Each sampled discharge is represented by a line, with several attributes (diagnosis, age, sex, etc.) encoded numerically (Criterion 1). The data enables calculating a hospitalization rate, which is not exactly the same as the incidence rate, since a person can be admitted several times to a hospital for the same disease; therefore, Criterion 2 (incidence rates) is only partially fulfilled. The detailed information available about the patient of each data record means Criterion 3 (age and sex-specific) is clearly met. Criterion 4 (completeness) is partially fulfilled since the NHDS only covers diseases of cases that are admitted to a short-stay hospital. The 2007 survey included 5536 different primary diagnoses encoded with the ICD-9-CM classification, which is an extension of the ICD-9 (Criterion 5).

WHO: The data repository of the WHO offers the possibility to download datasets in various formats (Excel, CSV); therefore Criterion 1 is met. However, the incidence numbers are not always present. For some diseases, only the mortality numbers are included. Therefore, Criterion 2 (incidence rates) is almost not fulfilled at all. The datasets are usually stratified by sex, but are not age-specific (Criterion 3). The data repository is far from complete. It covers all types of health indicators that are not disease-specific. Hence, Criterion 4 (completeness) is not met. Furthermore, a classification of the datasets with disease codes is missing (Criterion 5).

StatAUT: The cancer registry of Austria annually publishes the incidence numbers of various cancer types in the Austrian population and offers the dataset as an Excel file (Criterion 1). The numbers are age-standardized with the standard population published by the WHO [1]. Further categorization into age groups is not available, which means Criterion 3 is only partially fulfilled. Besides, since it only covers types of cancer, the data does not meet the completeness criterion. The various cancer types are encoded with the ICD-10 [114].

GLOBOCAN: The GLOBOCAN project covers a similar spectrum as the national cancer registry of Austria, but for more than 184 countries [25]. However, we could not find a feasible way to export the data in a machine-readable format. The project offers only an online data browser, which allows visually analysing the data. Again, the cancer types are encoded with ICD-10 codes.

3.3 Age and Sex Differences

Our approach to personalize health search is based on the fact that persons of different sexes and ages are statistically more or less likely to be confronted with different diseases and health disorders. This section illustrates how men and women, as well as young and old people, are affected differently by various diseases. First, we demonstrate the role of sex and gender in health and medicine. Then, we continue to discuss age and its influence. Finally, we provide a quantified analysis of the NHDS data-set that we used to perform our experiments.

The Role of Sex and Gender

Regitz-Zagrosek explained that [66, p. 1]:

> "... being a woman or being a man significantly influences the course of diseases and therefore this fact must be considered in diagnosis and therapy ."

Health- and medicine-related research must pay close attention to clear differentiation between the terms *gender* and *sex*. These terms are not interchangeable and denote two different concepts. Gender is a social construct, whereas sex is a biological construct. Depending on the medical conditions, gender can be the relevant factor, whereas other conditions are only related to the biological sex of the patient. Both factors can also have a synergistic effect [42]. Sex-related health outcomes are determined by biological differences, for example, differences in sexual hormone levels, differences in anatomy, genetic expression, metabolism, or differences in expressions of receptors, enzymes and binding proteins. On the other hand, gender-related health outcomes are influenced by differences in the personal and societal perceptions of women's and

men's roles, different coping strategies of both genders, differences in stereotyping and prevalence attribution, and differences in access to health-care [66, p. 10]. We also want to point out that the female-male division is actually a false dichotomy, as some individuals are intersexual, which means that they are born with male and female characteristics. Hence, a personalized health search interface must regard the sex variable as optional.

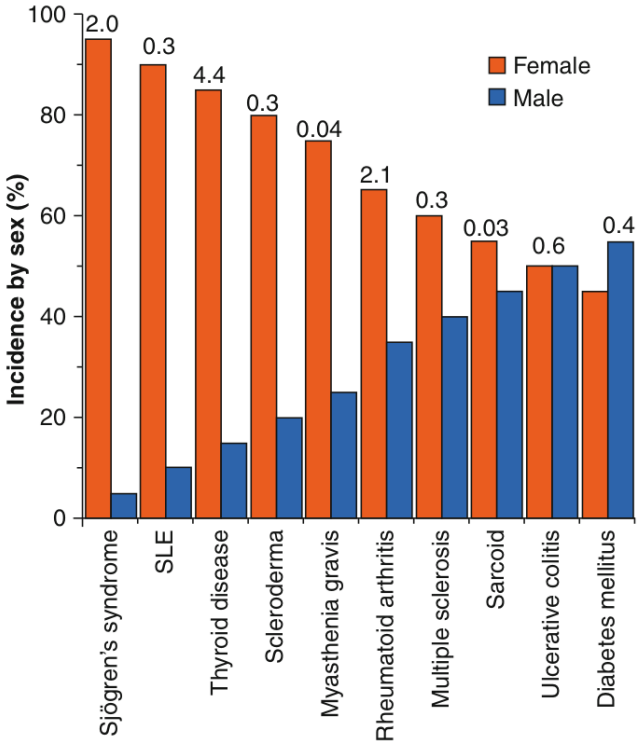

Figure 3.2: Biological factors involved in sex differences in the immune system [66, p. 102].

Various diseases are more common among men, whereas others are found typically in female patients. Autoimmune diseases, like *systemic lupus erythematosus* (SLE), are more likely in the female population than in the male population, as presented in Figure 3.2. Women of childbearing age are mostly affected by this disease. Experimental mouse models of SLE suggest sex hormones as an influencing factor [66, p. 102].

Myocardial infarction has been considered a typical male disease and, probably due to counter measurements, the risk is decreasing for almost all population groups. An exception would be young women, where the numbers are actually increasing. Explanations, which are still being discussed, include a change in life-style and an increasing prevalence of smoking. [66, p. 17-18].

Hypertension (high blood pressure), sometimes also referred to as arterial hypertension, is a medical condition where the patient suffers from elevated blood pressure in the arteries. It affects young men more often than young women. However, with increasing age, the gap between men

and women closes, and the prevalence among women even supersedes the prevalence among men [2].

Hepatocellular carcinoma is a type of liver cancer, which is also unequally distributed among men and women. The male:female ratios usually range between 2:1 and 4:1, sometimes at even around 5:1 in medium-risk European populations. These differences are thought to be due to men being more exposed to alcohol and smoking, which would make it gender related. Furthermore, men are more likely to be infected with *Hepatitis B* and *Hepatitis C* [23]. However, experiments with mice show a two- to eight-fold increase in male subjects, which supports the hypothesis that androgens (hormones that stimulate male characteristics) influence the development of hepatocellular carcinoma [84].

These examples should illustrate how diseases are unequally distributed among men and women, whether influenced by socially defined roles and behaviour (gender), or because of biological differences (sex).

The Role of Age

Reijneveld pointed out that [76]:

> "...even though analyses by age are among the most widely used tools from epidemiological toolbox, the adequate inclusion of age still merits attention ."

Age plays an important role in epidemiological considerations. It is a demographic variable that is easy to measure and, therefore, likely to be available. Most epidemiological studies aim to identify risk factors, so that countermeasures and prevention can be undertaken subsequently . However, aging in a person is inevitable. Therefore, there are no incentives to identify age as single causal risk factor. In fact, analytical studies aim to control age as variable, since it biases the occurrence of many diseases. For example, a study that investigates the influence of physical activities on the risk of suffering from heart failure, will have to consider that the incidence rate of heart failure is higher in the elderly population, but at the same time, that older people tend to engage less in physical activities. An inconsiderate inference could attach too much weight to the influence of physical activities.

Age can be regarded as a continuous variable, or as a categorized variable, by forming age-groups which span across several years. Many authors suggest that age should be recorded as precisely as possible [76]. Although age has continuous characteristics, there are qualitative points in everybody's life, which are marked by significant changes in the hormonal balance. For example children reaching adolescence, and older people hitting menopause and andropause.

Children and elderly persons are more vulnerable to a number of diseases, for example *influenza* (flu) [14]. Some diseases are more prevalent in people of a younger age, but this might be because, after being exposed to a disease, one develops a life long immunity. Many diseases are typically found in the older population, including cancer, Alzheimer's disease, atherosclerosis, cardiovascular disease, arthritis, etc. [32].

Since age is an important confounder, populations under investigation are often adjusted for age to allow meaningful comparisons. The World Health Organization (WHO) published standard populations, which allows for adjusting results and, therefore, enabling comparisons between populations with different age structures [1].

Entropy Estimates

We illustrated, how diseases affect persons of different age and sex in a more or less severe manner in terms of occurrence. We presented some representative exemplary diseases and described their unequal distribution. The remaining part of this chapter analyses the differences in interpretable quantities. We assume the position of a physician, who tries to find the cause of a patient's condition. The diagnosis is the result of putting together pieces of information. The physician consults the patient's history, asks the patient questions, orders laboratory examinations, etc., to narrow down the possible causes of the patient's condition. We can picture the set of diagnoses as a search space, in which the physician navigates towards the correct answer. Shannon's *entropy* measure (H) [88] allows us to express this search space in the basic unit of information, in bits. The entropy of a discrete random variable X, with possible values $\{x_1, x_2, .., x_n\}$ is defined as:

$$H(X) = -\sum_{i}^{n} p(x_i) \log p(x_i). \tag{3.4}$$

We model the medical condition of a patient as a discrete random variable C, with all the diagnoses from the NHDS data set as possible outcomes. The entropy of this variable, measured in bits, can be interpreted as, how many yes/no questions have to be asked on average to find the correct diagnosis, under the assumption that only diagnoses which occur in the NHDS data set are possible. The *conditional entropy* [19] enables us then to calculate the average number of yes/no questions, given that we have already some information- for example, the age of the patient. By comparing the entropy before knowing the age of a patient with the conditional entropy after the age is determined, we can quantify how much information is actually explained by a patient's age.

The NHDS data-set contains 365648 hospital discharges, with 5536 different primary diagnoses. For our studies, we excluded all diagnoses with E and V codes (external causes of injury and supplemental classification). Thereby, the number of different diagnoses decreased to 5269. We calculated the entropy $H(C)$ by estimating the probabilities of the diagnoses by their frequencies within the data-set:

$$H(C) = 9.3164 \tag{3.5}$$

Our calculations show that the entropy of the search space is about nine bits. We introduce another discrete random variable A, which maps to one of four age groups: 0-14 years, 15-44 years, 45-64 years, and 65 and above. We now proceed to calculate the conditional entropy, which is defined as:

$$\begin{aligned} H(C \mid A) &= H(C, A) - H(A), \\ H(C \mid A) &= -\sum_{i}\sum_{j} p(c_i, a_j) \log p(c_i, a_j) + \sum_{i} p(a_i) \log p(a_i). \\ H(C \mid A) &= 8.8065 \end{aligned} \tag{3.6}$$

where $p(c, a)$ denotes the joint probability of a person having a condition c and being a member of the age group a. The joint probabilities are estimated by the frequencies of the diagnosis, stratified by the age groups. The conditional entropy is about half a bit lower:

$$H(C) - H(C \mid A) = 0.5099 \tag{3.7}$$

The entropy decreases by about 5% when the age group of the patient is known.

We can model the sex of a patient as a binary random variable S and calculate the conditional entropy in the same manner:

$$\begin{aligned} H(C \mid S) &= 9.1931 \\ H(C) - H(C \mid S) &= 0.1234 \end{aligned} \tag{3.8}$$

Since the decrease in entropy is much lower, it seems, according to the calculated conditional entropies, that the age of a person provides more information, than the sex. The conditional entropy, incorporating the age group and the sex, is given by:

$$H(C \mid A, S) = H(C, A, S) - H(A, S) \tag{3.9}$$

When both, the age group and the sex of a patient are known, then the entropy drops a little more:

$$\begin{aligned} H(C \mid A, S) &= 8.6407 \\ H(C) - H(C \mid A, S) &= 0.6758 \end{aligned} \tag{3.10}$$

Based on the data from the NHDS, the age and sex of a patient provide at average 7% of the information necessary to determine the correct diagnosis. All results are rounded to four decimal places.

These results have to be interpreted in the context of information theory, which means that the calculated 7% apply only if the correct diagnosis is among the 5269 diagnoses from the data-set, and the physician knows an optimal strategy to obtain information in the form of binary measurements (yes/no questions). Our analysis, therefore, is of a more theoretical nature, and should only provide a formal basis for our hypothesis: that the knowledge of the age and sex of a person can be exploited to narrow down not only the diagnosis space, but, analogously, also the document space for a query with diagnostic intent.

CHAPTER 4

Personalized Probabilistic Health Search

This chapter describes the theory and general consideration behind our PPHS model. The research question is whether health search can be improved by incorporating health statistics. We assume a state-of-the-art model that produces results, which are measurable in terms of ad hoc retrieval effectiveness. The goal of the PPHS model is to produce results for the information needs of health- and medicine-related topics that outperform the state-of-the-art approach with respect to *precision* (P), see Definition 6.1. An IR system's effectiveness is actually an aggregate of its speed, user interface, retrieval performance, and its ability to support the user in completing his or her task (Manning et al. present a broader perspective on system quality and user utility [50, p. 168-169]). Nonetheless we leave out other aspects other than precision within this thesis. Our main concern is that relevant documents are ranked before nonrelevant documents in a response to a query. Another way to formulate our goal is this: The optimal response to a query, is a ranked set of documents, so that the precision of the subset of the top k documents is maximized for all k. Therefore, our approach focuses on identifying relevant documents.

In Section 2.3, we described which properties an IR model needs to have, so that we consider it contextual according to our three-stages model. It interprets terms semantically and recognizes intended concepts (conceptual stage). Furthermore it puts these concepts in context, by interpreting them as instances with specific attributes. The expectation is that a high level of understanding by the IR system yields better results. However, our the state-of-the-art review showed that, from the perspective of the three-stages model, there only exist approaches which approximate the ideal (Chapter 2). The context is not determined for individual terms, but rather for the whole document. We mentioned a model that puts documents in a geographic context [49], as well as a model that assigns broad topics to whole documents [93]. These approaches can also be described as models that estimate a document's prior probability, independent of the query. Document priors are not a novelty, and a number of retrieval algorithms assess a document's relevance before any query has been issued- for example, the *PageRank* algorithm [12]. Our PPHS approach can be described as estimating document priors, too. In fact,

the model is two-fold: (1) it aims to incorporate context by taking a patient profile into account and (2) it estimates document priors, based on the incidence rate of the disease that a document covers.

Our approach to health search is based on several assumptions:

1. The user entered an ASK (see Section 1.2), because of a specific medical case. Therefore, a single patient of known age and sex is given.
2. A document provides information about a disease or health disorder.
3. A document's prior relevance probability is proportional to the incidence rate of the disease that it covers.
4. Diseases have different incidence rates for persons of different age and sex.

Our theoretical justification for Assumption 3 is, that we expect the users' ASK to be induced by the onset of a disease in a person. Therefore, we presume that documents which cover information about the disease are relevant, regardless of the specific query. However, the IR system does not know which disease the patient has, but can make an informed guess based on statistics, and estimates of prior probabilities. Nonetheless, we can not provide any evidence for this assumption. It is also possible that there exists an inverse correlation: If a disease occurs very seldom, than it is more likely to cause an ASK in a physician. The personalization of our approach is based on Assumption 4. However, it is also based on Assumption 3. Basically, if both assumptions hold, then the IR can estimate the document priors more accurately.

4.1 Probability Ranking Principle

In Section 2.1, we already presented the basic idea behind probabilistic models in IR and some specific approaches. We want to recall briefly the probabilistic ranking principle (PRP), which states that an IR system, that returns documents in order of decreasing probability of usefulness to the user, on the basis of whatever data have been made available to the system, is optimal with regard to the available data [78]. The key challenge is to estimate these probabilities, so that the emerging order of documents reflects the true order as accurately as possible. Our approach attempts to improve the estimates by increasing the amount of available data. We consider two additional data sources: (1) a patient profile, that is composed of the sex and age of the patient and (2) epidemiological statistics, which we use to estimate correlations between patient profiles and medical conditions.

Central to a probabilistic IR model is a discrete random variable $R_{d,q} \colon \Omega \to \{0, 1\}$, which we define as:

$$R_{d,q}(\omega) = \begin{cases} 1, & \text{if } d \text{ is relevant to } q \\ 0, & \text{otherwise} \end{cases}. \tag{4.1}$$

Our considerations are based on a random process, for which the set of outcomes consists of all combinations of documents from the collection $\mathcal{D}$, with queries of a query set $\mathcal{Q}$[1], and the

[1]The query set can be thought as a formal language that is defined by the query syntax.

two relevance judgments. Therefore, $p(R_{d,q} = 1)$ denotes the probability that document d and query q are part of the outcome, and d is relevant to q. The process begins with the user issuing a query q and, subsequently, the IR system estimates following conditional probabilities:

$$p(R_{d,q} = 1 \mid d, q), \tag{4.2}$$

for each document $d \in \mathcal{D}$. The documents are then returned to the user in the order of decreasing probabilities. In the remainder of this chapter, we follow the convention of Manning et al. and write just R for $R_{d,q}$ [50, p. 221].

4.2 Personalizing Web Search

Our approach to personalizing health search is based on the work of Sontag et al., who presented probabilistic models to personalize general web search [93]. In this section we discuss the details of one of their models that are relevant to our adaption, which helps understand the motivation behind our approach.

Different users have a deeper interest in some topics than others. At the same time, different documents cover certain topics to varying degrees. The basic assumption is then, that documents covering a topic in which a certain user is more interested in general, have a higher prior probability to be relevant, independent of any query. For example, documents about search algorithms have a higher prior probability for a computer scientist than for a generic user. Sontag et al. model this assumption by introducing a user profile θ_u which is learned from the user's historical data. The conditional relevance probabilities are then not only based on the document and query, but also on the user's profile:

$$p(R_u = 1 \mid d, q, \theta_u). \tag{4.3}$$

Furthermore, their approach also incorporates a relevance signal from a user-independent retrieval model, which is denoted as $\psi(d, q) \in [0, 1]$. The probabilistic model is defined by this formula:

$$p(R = 1 \mid \theta_u, q, d, \psi(d, q)) = \psi(d, q) \sum_{T_d} p(T_d \mid d)\alpha(T_d), \tag{4.4}$$

$$\alpha(T_d) = \sum_{T_u} p(T_u \mid \theta_u, q) p(\text{cov}_u(d, q) = 1 \mid T_u, T_d).$$

As mentioned earlier, the user u is modelled with a user profile and represented as θ_u. The variable T_u represents the topic of the user's information need. The variable T_d represents the topic that is covered by the document d. The variable $\text{cov}(d, q) \in \{0, 1\}$ indicates that topic T_d covers topic T_u.

The distribution $p(T \mid \theta_u, q)$ is derived by applying Bayes' rule and marginalizing the denominator:

$$p(T \mid \theta_u, q) = \frac{p(T \mid \theta_u) p(q \mid T)}{\sum_{T'} p(T' \mid \theta_u) p(q \mid T')}. \tag{4.5}$$

In the evaluated implementation of Sontag et al., the parameters of a user profile θ_u, which determine the distribution $p(T \mid \theta_u)$ are estimated from historical search data of the user in an off-line step. The topic space is formed by the top two levels of the human-generated ontology provided by the Open Directory Project [68]. These are broad topics- for example, computer science, arts or sports. The distribution $p(T_d \mid d)$ is estimated for each document d in the collection by a text-based classifier, that was trained with logistic regression. For the distribution $p(q \mid T)$ the authors propose a unigram language model:

$$p(q \mid T) = \prod_{w \in q} p(w \mid T). \tag{4.6}$$

Sontag et al. conducted a large-scale evaluation of their models and could demonstrate improvements over the baseline, especially for ambiguous queries and queries with acronyms.

4.3 Personalizing Health Search

This section presents how we adapted the model of Sontag et al. to the domain of health search. In contrast to their generic approach, we do not personalize results for a user, but for a patient. We want to point out that the searcher must not necessarily be the same person as the patient. For example the searcher can be a physician treating the patient, or a relative of the patient, trying to gather information about the patient's condition.

A patient is modelled with a two-parametric profile $\theta_{a,s}$ that is determined by the age $a \in [0, 99]$[2] and the sex $s \in \{m, f\}$ of the patient (we will denote the profile as θ from now on). The PPHS model incorporates estimates based on epidemiological data in order to re-rank documents based on the likeliness of the onset of the health disorder, which a document covers. Specifically, we are interested in the probability that a patient of a certain age and sex has developed a medical condition C:

$$p(C \mid \theta). \tag{4.7}$$

Further, we need to estimate the conditional distribution of medical conditions for a given document d:

$$p(C \mid d), \tag{4.8}$$

so that we can adapt Equation 4.4 and Equation 4.5 to:

$$p(R = 1 \mid \theta, q, d, \psi(d, q)) = \psi(d, q) \sum_C p(C \mid d) p(C \mid \theta, q), \tag{4.9}$$

$$p(C \mid \theta, q) = \frac{p(C \mid \theta) p(q \mid C)}{\sum_{C'} p(C' \mid \theta) p(q \mid C')}.$$

[2] we used the NHDS data-set, which recodes ages ≥ 100 to 99

We do not take over an equivalent for the variable cov_u. As a consequence the Equation 4.9 is simplified to a single sum. The distribution $p(q \mid C)$ can be estimated similarly to the proposal of Sontag et al. for their model (see 4.6), by a unigram language model:

$$p(q \mid C) = \prod_{w \in q} p(w \mid C). \tag{4.10}$$

In terms of our PPHS model, the probability of a document being relevant to a query depends on: (1) a profile-independent relevance score, $\psi(d, q)$, (2) the probability that the document covers information about a medical condition, and (3) the probability that a patient has developed the same medical condition.

We continue with our approach to estimate the conditional distribution of medical conditions, depending on a patient profile (see 4.7). Intuitively, the age- and sex-specific prevalence of a health condition seems to be a suitable estimator of its probability being present in a given case. But central to the PPHS model is the assumption that the search was initiated by an ASK, which resulted from the recent onset of symptoms in a patient. Hence, its recency is essential. Definition 3.2 of the prevalence measure states that the onset of a case has to occur before a specified point in time, but not how recent it should be. Therefore, the prevalence measures of diseases and medical conditions are not useful for our needs. On the other hand, estimates based on the incidence rate seem to be a better approach. The higher the rate, the more new cases of the disease occur in any given time frame. Since it is a rate and not a proportion, it is not trivial to derive a probability estimate. However, we circumvent this problem, by multiplying with a normalization factor λ_θ. Its derivation is presented briefly in this section. We estimate the probability of the recent onset of a disease C, given the profile θ by the age- and sex-specific incidence rate, which we denote with $\kappa(C, \theta)$:

$$\hat{p}(C \mid \theta) = \lambda_\theta \kappa(C, \theta). \tag{4.11}$$

In order to estimate probabilities, we convert the incidence rates into proportions. We consider a time frame t, which we assume that the onset of the disease of the patient falls into. The incidence number of a disease within a time frame t is then given by $t\kappa(C)$. The profile-specific incidence is calculated from the profile-specific incidence rate: $t\kappa_\theta(C)$. Under the premise that the patient has developed a medical condition C and $p(\cup_i C_i) = 1$, we can estimate:

$$\hat{p}(\theta) = \frac{\sum_C t\kappa(C, \theta)}{\sum_{C'} t\kappa(C')}, \tag{4.12}$$

and we estimate the joint probability of a disease together with a profile as:

$$\hat{p}(C, \theta) = \frac{t\kappa(C, \theta)}{\sum_{C'} t\kappa(C')}. \tag{4.13}$$

We insert our estimates instead:

$$
\begin{aligned}
p(C \mid \theta) &= \frac{p(C, \theta)}{p(\theta)} \\
&\hat{=} \left[\frac{t\kappa(C, \theta)}{\sum_{C'} t\kappa(C')} \right] / \left[\frac{\sum_{C''} t\kappa(C'', \theta)}{\sum_{C'''} t\kappa(C''')} \right] \\
&\hat{=} \frac{t\kappa(C, \theta)}{t \sum_{C'} \kappa(C', \theta)} && \text{(reducing the fracture)} \\
&\hat{=} \lambda_\theta \kappa(C, \theta), && \text{(the time frame } t \text{ becomes obsolete)} \\
\lambda_\theta &= \frac{1}{\sum_{C'} \kappa(C', \theta)}.
\end{aligned}
$$

For an estimation of the distribution $p(C \mid d)$ we propose a text-based classifier that assigns probabilities to medical conditions for a given text t, which we denote as $\phi(c, t) \rightarrow [0, 1]$. For simplicity's sake, we suggest using the same classifier to estimate the probability of condition to generate a specific query:

$$
\begin{aligned}
\hat{p}(C \mid d) &= \phi(C, d), \\
\hat{p}(q \mid C) &= \phi(C, q).
\end{aligned} \tag{4.14}
$$

At this point we have all necessary estimators. We new replace probability terms of Equation 4.9 with the estimators:

$$
\hat{p}(\theta, d, q) = \psi(d, q) \sum_{C} \phi(C, d) \frac{\kappa(C, \theta)\phi(C, q)}{\sum_{C'} \kappa(C', \theta)\phi(C', q)}. \tag{4.15}
$$

The factor λ_θ can be reduced and we have established a probabilistic personalized ranking model for health search. The model is defined by a four-tuple, PPHS $= (\mathcal{C}, \psi, \kappa, \phi)$. In the following chapter, we will demonstrate a reference implementation, which we used to evaluate the model. The results of the evaluation will be presented in Chapter 6.

CHAPTER 5

Reference Implementation

This chapter presents the implementation of a prototype. It was used to evaluate the PPHS model presented in Chapter 4. The model is defined by a 4-tuple, $(\mathcal{C}, \psi, \kappa, \phi)$, that describes also the cornerstones of the architecture of the prototype. The source code can be otained as a git repository [75]. The three-stages model presented in Chapter 2 serves as a conceptual link between the implementation and the theoretical model. We briefly recall: (1) the lexical stage encompasses algorithms that match terms based on their spelling. (2) IR techniques operating at the conceptual stage identify the intended concept that is represented by a lexical term and, (3) the contextual stage annotates instances of concepts with contextual characteristics.

Figure 5.1 displays the main components and the dataflow between them.

The query is first processed at the lexical stage. At this stage we deployed an Apache Solr instance [4]. This open source full-text search engine is built on top of the search library Apache Lucene. It implements several state-of-the-art retrieval models. In our prototype it processes the query and responds with a search result that is ranked by the relevance score $\psi(d, q)$. The top 150 documents are then further processed at the conceptual stage. The documents are mapped to diseases, which is denoted as $\phi(c, d), \forall c \in \mathcal{C}$. The same component maps the query to diseases: $\phi(c, q), \forall c \in \mathcal{C}$. We deployed a Solr instance at this stage, too. Diseases are encoded in ICD-9-CM. Further on, the documents together with their disease mappings are then processed at the contextual stage. The documents are re-ranked by a score which is calculated according to the PPHS model. The incidence rates, denoted by $\kappa(C, \theta)$ have been estimated in an off-line step. The patient profiles are not detected automatically. They have to be hard-coded.

Figure 5.1 presents which technologies where deployed at which stage. It also presents the off-line dataflows or preprocessing steps. The document collection is indexed by the Solr instance at the lexical stage. Ths Solr instance at the conceptual stage indexes Wikipedia articles which are linked with ICD-9-CM codes. This Solr instance is used to automatically annotate documents or queries with ICD-9-CM codes. In addition to the Wikipedia articles, a description string, which was constructed with the UMLS Metathesaurus API, was indexed, too. At the contextual stage, we used R for the estimation of incidence rates via local regression. The R script retrieved the sample data from a PostgreSQL database, which served as a data-warehouse

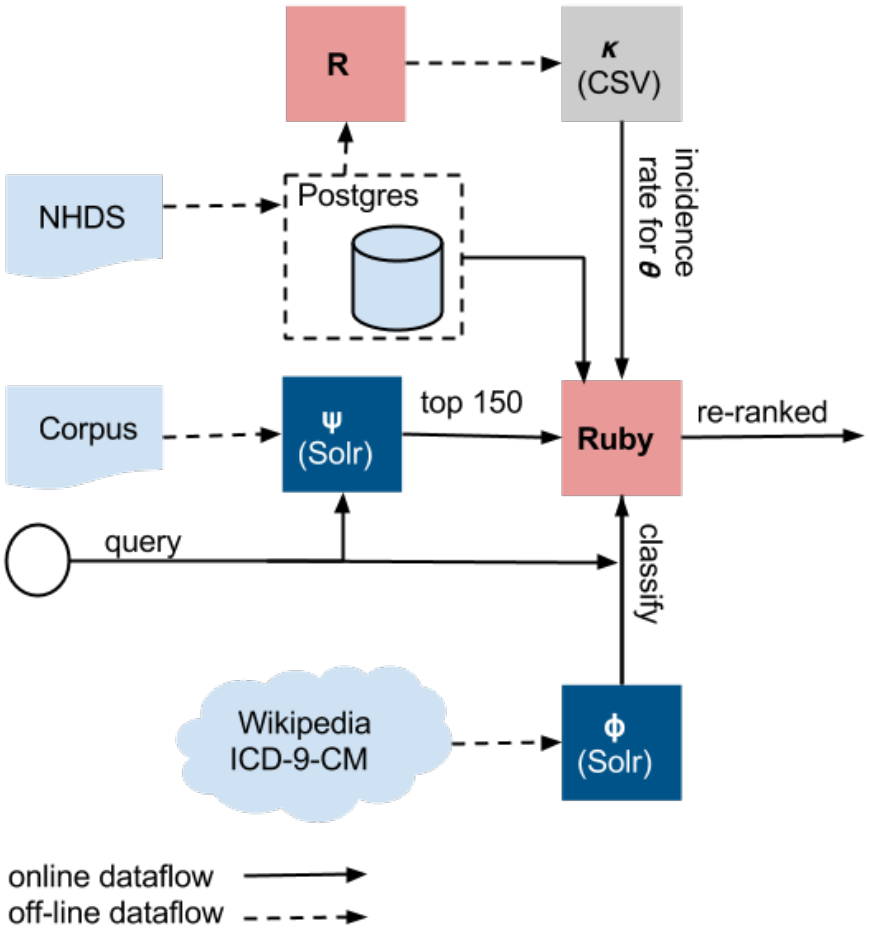

Figure 5.1: Architecture of the prototype.

for the components operating at the conceptual and contextual stage. It stores the discharge records from the NHDS data-set in a relational schema. The glue code, which connected all components, was written in the Ruby language [83]. This component governs the dataflow and performs the final computations of the PPHS model. We used the arbitrary precision type `BigDecimal` during the implementation for all calculations.

We used Apache Solr in version 4.10.2, which we built from the git repository [1]. The schema of the index was composed of a document ID, a title, and a content field. Before documents were added to the index, each document was preprocessed in several steps that were:

1. removing all HTML and XML tags,

2. removing stop words,

[1] https://github.com/apache/lucene-solr

3. normalizing all tokens to lower case,

4. removing possessives from words,

5. and stemming with the Porter stemming algorithm [74].

The search engine is accessed through its JSON API endpoints, both for indexing and querying.

5.1 Preparing Epidemiological Data

We used publicly accessible data obtained from the National Hospital Discharge Survey (NHDS) [15]. The NHDS was an annual survey conducted by the Centers for Disease Control and Prevention, the leading public health organization in the United States [14]. It was first conducted in 1965 and the survey was redesigned in 1988 [51]. We used the data from the 2007 sample, because beginning with 2008, the sample size of the survey was reduced by 50 percent. The 2007 sample had therefore smaller estimated standard errors.

They survey samples hospital discharges nationwide. One sample record includes, among many other attributes, the age, the sex and up to eight diagnoses encoded in ICD-9-CM. This is the reason why we used the ICD-9-CM as our conceptual layer. The 2007 sample records are distributed in one single file that contains more than 30K records. Each line represents a record and the data is encoded with alphanumeric characters. We assumed the file to be ANSI encoded, however, we could not find a definite statement about the encoding in the technical documentation. The discharge attributes are encoded by position in the record line that is 88 characters long. One record carries various attributes. Among them are demographic features, such as age, sex and ethnicity, but also the sample weight.

We imported the records into a relational database. We chose the open source solution PostgreSQL 9.4.0, but any other SQL database would have been suitable. Each record of the NHDS data set became a record in a `Discharges` table. During the creation of the discharge record, a corresponding conditions record in the table `Conditions` with the same ICD-9-CM code as the `Diagnosis code #1` attribute of the discharge record was looked up. If the entry had not been found, a new one was created. The discharge record pointed to the condition with a foreign key attribute. The conditions table was populated with records for each diagnosis that occurred as a primary diagnosis in one of the discharge records.

5.2 Concept Mapping

The objective at the conceptual stage is to implement the function which is denoted as ϕ in the PPHS model. The function scores ICD-9-CM concepts for a given piece of text. Trieschnigg et al. presented concept classifiers based on MeSH documents (see Section 2.2 for a description of MeSH). A sample document, which is refered to as a special "MeSH document", was created for each MeSH concept. These documents were created by merging titles and abstracts of documents that have been indexed manually with the corresponding MeSH term [105].

We scraped the Wikipedia pages that list ICD-9-CM codes for links to articles that cover the corresponding concepts. We analysed the structure of the listing page and scripted a web-crawler. For this purpose we used the tool Nokogiri [65]. The crawler visited each ICD-9-CM chapter from the list page. The individual chapter pages list the ICD-9-CM codes from the hierarchy level that is identified with the first three digits of the code. We created an extra table in the PostgreSQL database for this hierarchy level, which we named `ICD9Chapters`. We created records for all the nodes at this level. For each node we obtained the content of the articles that are linked to them. The articles, that are linked to subordinate nodes, were also scraped. We stored the content in the corresponding database record.

Furthermore, we expanded the ICD-9-CM description of a concept with description of other controlled vocabularies. We discussed briefly the UMLS Metathesaurus in Section 2.2. The thesaurus can be queried via an online API [107]. We obtained corresponding descriptions of the ICD-9-CM concept, which stem originally from other vocabularies. We added them to the concept records in the database table.

The content from Wikipedia and the expanded description were used to create a "ICD-9-CM concept document" for each ICD-9-CM concept at the three digits level. These documents were indexed by the Solr instance at the conceptual stage. In order to map a piece of text to ICD-9-CM codes, the text is sent this Solr instance as a query. The Solr instance is configured to calculate similarity scores with its default scoring method (TFIDFSimilarity). It is a VSM with TF-IDF weighting. The scores are interpreted as confidence estimates in the different ICD-9-CM codes and represent the ϕ function. Unfortunately, due to missing content for some of the ICD-9-CM concepts, we could not create a concept document for all of them. As a result, these concepts had no corresponding concept document that we could have indexed. As a result the number of initially 908 ICD-9-CM concepts fell to 775. The counted concepts are all ICD-9-CM codes at the three-digits hierarchy level, but without the E and V codes and the concepts without any Wikipedia articles.

5.3 Incidence Rate Estimation

We reasoned in Chapter 4 that the incidence rates of diseases are the most suitable measure for the PPHS approach. However, we could not find any data-source that provides incidence rates for an exhaustive set of diseases. The NHDS data set covers a wide range of diseases, though. But its data allows only to estimate a hospitalization rate. We decided that we estimate the annual incidence rate of based on this NHDS data set, nevertheless. We regarded each primary diagnosis of a discharge as an incidence (onset) of the diagnosed disease.

To do so, we exported from the relational database the data in a CSV file format, which was further processed by an R script. The R script performed a nonparametric regression with local polynomials. We used the `locfit` function with its default parameters from the package with the same name [33, 47]. We chose a nonparametric approach, since we performed the regression for several diseases for which we did not assume any predetermined structure in the age incidence relationship. We created three regression models for each disease. Two models were inferred from data sets which were stratified by sex, and one model was created based on the data combining discharges from both sexes. We then estimated incidence numbers for the

hard-coded patient profiles based on the regression models and stored the results in a CSV file. This file was loaded during the evaluation runs and the incidence numbers were looked up from memory.

5.4 List of Software

This section provides a comprehensive overview of all the software packages that we used for our implementation.

- **Apache Solr 4.10.2**: This open source full-text search engine provides a variety state-of-the-art IR models. We used an instance on the lexical stage which filters the initial set of documents. These are then re-ranked on the contextual stage. Furthermore, we used an instance on the conceptual stage to classify documents and queries into ICD-9-CM concept categories.
- **PostgreSQL 9.4.0**: We used this open source relational database as a data-warehouse. The NHDS data set is parsed and stored in this database in a relational schema. The relational schema allows a flexible perspective on the data and generates relational projections as we need them.
- **Nokogiri 1.6.6**: We used this HTML/XML parser library to process the documents collections, the evaluation topics, and further, we used it to scrape Wikipedia articles.
- **locfit 1.5-9.1**: This R library is used to perform nonparametric regression on the NHDS data set. We regarded age as a continuous variable. In order to estimate the incidence rate for any profile, we predicted it with the resulting regression model.
- **R 3.1.2**: We used the statistical software R in order to access the `locfit` library, but we used it also to perform randomized tests for significance on our results.
- **Java 1.8.0_25**: We used the Java platform to access the UML Metathesaurus online API.
- **Ruby 2.1.1**: We use this interpreted scripting language to connect all software components and to manage the data-flow. We further used it do the final computations on the results of the concept classifier and the incidence estimates from the regression model.

CHAPTER 6

Evaluation and Results

The main goal of this thesis is to answer the question of whether retrieval models that incorporate epidemiological data can effectively improve the performance within health search. We formulate null hypotheses for our research questions that we presented in Section 1.2.

1. Using epidemiological data does not improve state-of-the-art IR within the domain of health and medical search.

2. Adapting probabilities to the age and sex of a patient does not improve a search model using epidemiological data.

This chapter demonstrates results of our experiments with the IR models that we presented in Chapter 4. The experiments were conducted in a reproducible environment. We employed two evaluation suites that were used in: (1) the Clinical Decision Support Track (CDS 2014) [103], and (2) The ShARe/CLEF eHealth Evaluation Lab 2014 Task 3 (CLEF 2014) [29]. The source code for the protoype, which was evaluated, can be obtained from a publicly available git repository [75].

6.1 Evaluation Measures

We evaluated our results with regard to these measures:

- *Mean Average Precision* (MAP),
- the precision considering only the top-N documents (P@N),
- the *Normalized Discounted Cumulative Gain* considering only the top-N documents (NDCG@N).

This section explains the measures in detail.

P@N

Precision is defined by the proportion of relevant documents $\mathcal{R} = \{d \in \mathcal{D} \mid \text{relevant}(d) = \textbf{true}\}$ within the set of retrieved documents $\mathcal{D}$.

$$\text{P} = \frac{|\mathcal{R}|}{|\mathcal{D}|}. \tag{6.1}$$

Given that $\mathcal{D}$ is a ranked set of retrieved documents and the set $\mathcal{D}_N = \{d \in \mathcal{D} \mid \text{rank}(d) \leq N\}$, which is the set of retrieved documents from the top document d, with $\text{rank}(d) = 1$ until the document at rank N, and $\mathcal{R}_N = \{d \in \mathcal{D}_N \mid \text{relevant}(d) = \textbf{true}\}$, then

$$\text{P@N} = \frac{|\mathcal{R}_N|}{|\mathcal{D}_N|}. \tag{6.2}$$

We evaluated our results at two levels: $P@5$ and $P@10$.

MAP

With the definition of $P@N$ we can define the *mean avarage precision* (MAP). Given that $\mathcal{R}_q$ is the set of retrieved relevant documents $\mathcal{R}$ for query $q \in \mathcal{Q}$ from a set of queries $\mathcal{Q}$, then

$$\text{MAP}(\mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{|\mathcal{R}_q|} \sum_{r \in \mathcal{R}_q} P@\text{rank}(r). \tag{6.3}$$

This definition is adopted with a slightly different notation. [50, p. 160]. As recommended by Buckley and Voorhees, average precision seems to be a stable and discriminating measure with regard to general purpose retrieval [13].

NDCG@N

Given that we have $R : \mathcal{D} \times \mathcal{Q} \to [0, r]$ a mapping of documents $d \in \mathcal{D}$ with queries $q \in \mathcal{Q}$ to graded relevance scores, where 0 means not relevant and r is the maximum relevance score, then the *normalized discounted cumulative gain* (NDCG) for N is calculated with

$$\text{NDCG@N} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} Z_{Nq} \sum_{d \in \mathcal{D}_N} \frac{2^{R(d,q)} - 1}{\log_2(1 + \text{rank}(d))}, \tag{6.4}$$

where Z_{Nq} denotes a normalization factor that has the value, so that a perfect ranking for query $q \in \mathcal{Q}$ results in $\text{NDCG@N} = 1$. $\mathcal{D}_N$ denotes the same set of retrieved documents until rank N as in the definition for P@N, see Equation 6.2. This definition is also adopted with a slightly different notation [50, p. 163]. As we have graded relevance judgments for both evaluation tracks that we performed, NDGC@N is a measure of interest. The measure is user-oriented, since it weights down the gain from relevant documents that are ranked poorly [39]. We evaluated our results at two levels: NDCG@5 and NDCG@10.

These measures have different properties and allow us to compare our results from different perspectives. P@N is easy to interpret and allows a quick judgment on how the run performed. MAP is stable and discriminative. We used this measure to test for statistical significance. NDGC@N puts more emphasis in the graded judgments and reflects the user experience more accurate, since users tend to expect relevant documents to be ranked higher.

6.2 Evaluation Tracks

As mentioned in the introduction to this chapter, we evaluated the PPHS model on two IR test tracks. This section describes the tracks, their test collections and goals.

TREC Clinical Decision Support Track

This track's goal is to advance the development of tools that retrieve relevant information for medical cases. The users of such systems would be physicians in need of information when presented with a medical case narrative. The information needs can be of three different types: (1) information to determine the diagnosis, (2) information that supports the choice of an appropriate treatment plan, and (3) information that supports the selection of medical tests in order to find a diagnosis.

The test collection is a snapshot of the Open Access Subset [67] of the PubMed Central (PMC) Archive. The snapshot was taken on January 21, 2014 and contains 733,138 articles. PMC is a freely accessible database of biomedical journal articles.

The track is composed of 30 topic. Of these 30 topics, a subset of 10 topics is assigned to one of the three question categories: *diagnosis*, *test* and *treatment*. Each topic is composed of a longer description and a short summary, see Table 6.1 as an example.

topic	1
type	diagnosis
description	A 58-year-old African-American woman presents to the ER with episodic pressing/burning anterior chest pain that began two days earlier for the first time in her life. The pain started while she was walking, radiates to the back, and is accompanied by nausea, diaphoresis and mild dyspnea, but is not increased on inspiration. The latest episode of pain ended half an hour prior to her arrival. She is known to have hypertension and obesity. She denies smoking, diabetes, hypercholesterolemia, or a family history of heart disease. She currently takes no medications. Physical examination is normal. The EKG shows nonspecific changes.
summary	58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back.

Table 6.1: Exemplary topic of the CDS track.

The relevance assessment was conducted by the Department of Medical Informatics of the Oregon Health and Science University (OHSU). The relevance scale consists of three discrete levels: (0) definitely not relevant, (1) potentially relevant and (2) definitely relevant.

The pool of articles for assessment was formed by the highest ranking documents of the runs submitted by the track's participants. For further information, please see [103].

ShARe/CLEF eHealth Evaluation Lab 2014 Task 3

In contrast to the CDS track, the Share/CLEF eHealth Evaluation Lab Task 3 focuses on the information needs of laypeople in their role as patients. The goal of the task is to retrieve information in order to answer questions of patients that may rise induced by their health condition. The concrete use case depicts a patient, who had examined a discharge summary, and now wants to find out more about the stated diagnosis and the details that are presented in the discharge summary.

The test collection was made available by the Khresmoi project [31]. It is the result of a large-scale web crawl of publicly available web pages that cover health topics and are targeting the general public and healthcare professionals. In total the collection includes 1,104,337 documents.

The test suite is composed of 50 topics and was created by experts based on given discharge summaries and diagnoses. A topic is composed of a reference to its discharge summary, a title, a description, a narrative stating the intent of the search and a profile. Table 6.2 presents an example topic.

id	qtest2014.1
discharge summary	00211-027889-DISCHARGE_SUMMARY.txt
title	Coronary artery disease.
description	What does coronary artery disease mean?
narrative	The documents should contain basic information about coronary artery disease and its care.
profile	This positive 83 year old woman has had problems with her heart with increased shortness of breath for a while. She has now received a diagnosis for these problems having visited a doctor. She and her daughter are seeking information from the internet related to the condition she has been diagnosed with. They have no knowledge about the disease.

Table 6.2: Exemplary topic of the CLEF 2014 track.

The assessment pool was formed of 6,800 documents taken from the results of the submitted runs. The professional assessors judged the documents on a four-step scale: (0) irrelevant, (1) on topic but unreliable, (2) relevant and (3) highly relevant. For further information and a review of the results see [29].

6.3 Runs

We evaluated 12 runs in both evaluation suites, CDS 2014 and CLEF 2014. Seven runs were dedicated establishing a strong baseline for each test collection and its test queries. Five runs were conducted with the PPHS improvement and variants of it.

Baseline Selection

In order to select a baseline we conducted runs with state-of-the-art retrieval models, which are implemented within the Lucene project and, therefore, available in Solr search engines. See Table 6.3 for a list of the runs and their parameter settings. We used three, and two different parameter settings for the language model approaches respectively. As concluded by Zhai and Lafferty [119] in their study of smoothing methods for language models, an optimal smoothing factor depends on the type of query to some extent. Short keyword queries respond better to a Jelinek-Mercer model with a small λ whereas, for longer verbose queries, a higher λ is preferred. The optimal μ parameter of the Dirichlet prior tends to be around 2000. In general, the Dirichlet prior is suitable for keyword queries and loses on verbose queries.

BM25	$k_1 = 1.2$ $b = 0.75$	The similarity measure by Robertson et al. [81], as implemented in the Lucene Project.
TFIDF		The cosine similarity between document vector and query vector in the vector space [50], with term frequencies weighted with the inverse document frequency, as implemented in the Lucene project.
LM Dirichlet	$\mu = 400$	The language model approach with Bayesian smoothing using Dirichlet priors [119], as implemented in the Lucene Project.
LM Dirichlet	$\mu = 2000$	
LM Dirichlet	$\mu = 3000$	
LM JelinekMercer	$\lambda = 0.7$	The language model approach based on the Jelinek-Mercer smoothing method [119], as implemented in the Lucene Project.
LM JelinekMercer	$\lambda = 0.05$	

Table 6.3: Runs to determine a baseline.

As we test using short keyword queries from the CLEF 2014 test suite and verbose queries from the CDS 2014 test suite, we conducted runs with small and large smoothing factors, for both, the LM with Dirichlet prior and the LM based on Jelinek-Mercer smoothing.

PPHS runs

We conducted five runs in order to investigate the influence of a probabilistic model, based on estimates from epidemiological data. One run was performed on a variation that was not personalized, which means that the estimates were based on the incidence rates of the general population, independent of sex and age. One run was performed, with the PPHS model, which means that is adapted to the sex and age of a patient. Then we performed additional runs. In one run, only the sex of the patient was incorporated, and in the other run, only the age was considered. A final run was performed, in which the incidence rate estimates were all set to 1. This run should isolate the influence of the conceptual stage, where documents are mapped to ICD-9-CM concepts.

The scores were calculated by the linear combination of the score observed from the baseline IR model and the score that was calculated using the PPHS model, respective the adapted version: score $= \psi(d, q) +$ PPHS.

PHS	Probabilistic Health Search (PHS), this run was not personalized, which means that the estimation of the incidence rates was not age or sex specific.
PAHS	Probabilistic Age-specific Health Search (PAHS), the estimated incidence rates were age specific.
PSHS	Probabilistic Sex-specific Health Search (PSHS), the estimated incidence rates were sex specific.
PPHS	Personalized Probabilistic Health Search (PPHS), this run was personalized, which means that the estimation of the incidence rates was age and sex specific.
Control	This run was performed with the probabilistic improvement, except that all incidence rates of all diseases were estimated with 1. This means that the frequency of a disease does not influence the ranking. Everything else was left in its original configuration.

Table 6.4: PPHS runs

See Table 6.4 for a description of the runs.

6.4 Results

Results of Baseline Selection

From the baseline runs on the CDS 2014 track (see Table 6.5) TFIDF produced the best results for all considered measures. BM25 and LM Dirichlet came in second with $\mu = 3000$. Hence, we defined TFIDF run as our baseline and configured $\psi(d, q)$ of the PPHS runs to be the TFIDF Similarity score of Solr, for all subsequent CDS 2014 runs.

From the baseline runs on the CLEF 2014 track (see Table 6.6) the language model with Dirichlet smoothing produced the best results for all considered measures. A μ around 2000 seems to be a good setting. Hence, we defined the LM Dirichlet with $\mu = 2000$ as our baseline

			LM Dirichlet			LM JelinekMercer	
Measure	BM25	TFIDF	$\mu = 3000$	$\mu = 2000$	$\mu = 400$	$\lambda = 0.7$	$\lambda = 0.05$
MAP	0.1088	**0.1208**	0.1104	0.1088	0.1013	0.1042	0.0914
NDCG@5	0.2829	**0.3188**	0.2935	0.2946	0.2556	0.2555	0.2536
NDCG@10	0.2564	**0.2732**	0.2538	0.2519	0.2318	0.2325	0.2319
P@5	0.3267	**0.3667**	0.3333	0.3467	0.2933	0.28	0.3
P@10	0.29	**0.3033**	0.28	0.2867	0.2667	0.2633	0.2733

Table 6.5: Results of runs to determine baseline on the CDS 2014 track. The results in bold indicate the best result with regard to a single measure. The TFIDF run produced the best results with regard to all considered measures.

and configured $\psi(d, q)$ of the PPHS runs to be the LMDirichlet Similarity with $\mu = 2000$, for all subsequent CLEF 2014 runs.

			LM Dirichlet			LM JelinekMercer	
Measure	BM25	TFIDF	$\mu = 3000$	$\mu = 2000$	$\mu = 400$	$\lambda = 0.7$	$\lambda = 0.05$
MAP	0.3671	0.3136	0.3821	**0.3951**	0.3866	0.3059	0.2785
NDCG@5	0.6624	0.5235	0.7147	**0.7261**	0.6717	0.5122	0.4906
NDCG@10	0.6476	0.5264	0.6944	**0.7123**	0.6539	0.4955	0.5043
P@5	0.688	0.516	0.728	**0.752**	0.696	0.52	0.504
P@10	0.658	0.528	0.694	**0.718**	0.666	0.486	0.522

Table 6.6: Results of runs to determine baseline on the CLEF 2014 track.The results in bold indicate the best result with regard to a single measure. The language model with Dirichlet prior smoothing with $\mu = 2000$ produced the best results.

Results of PPHS runs

Measure	Baseline	PHS	PPHS	PSHS	PAHS	Control
MAP	0.1208	**0.1222**	0.1221	0.1221	**0.1222**	0.1215
NDCG@5	0.3188	**0.3308**	0.3286	**0.3308**	0.3286	0.321
NDCG@10	0.2732	**0.2885**	0.2863	**0.2885**	0.2864	0.2836
P@5	0.3667	**0.3733**	0.3667	**0.3733**	0.3667	0.3667
P@10	0.3033	**0.3167**	0.3133	**0.3167**	0.3133	**0.3167**

Table 6.7: Results of PPHS runs on the CDS 2014 track. The results in bold indicate the best result with regard to a single measure.

The results of the PPHS runs on the CDS 2014 track are presented in Table 6.7. All probabilistic models performed better than baseline. However, the absolute improvements are minimal with regard to every measure. The non-personalized run performed well in comparison to all measures. However, the differences between the probabilistic models are also very small. The control run, which was performed without any epidemiological statistics, performed worse with regard to all measures except P@10. But it is interesting that it still produced better results than baseline. Therefore, concept mapping appears to have a positive influence. In order to reject both null hypotheses, we performed a statistical test for significance. We chose $\alpha = 0.05$ as the significance level and performed a randomized test as suggested by Smucker et al. [91]. Based on the calculated p-values, we can not reject any of the null-hypotheses in regard to any measure.

Measure	Baseline	PHS	PPHS	PSHS	PAHS	Control
MAP	0.3951	0.396	0.3961	**0.3964**	0.3956	0.3944
NDCG@5	**0.7261**	0.7164	0.7139	0.7163	0.7117	0.7157
NDCG@10	**0.7123**	0.706	0.7042	0.7074	0.7018	0.7034
P@5	**0.752**	0.74	0.736	0.74	0.74	0.736
P@10	0.718	0.72	**0.722**	**0.722**	0.72	0.718

Table 6.8: Results of PPHS runs on the CLEF 2014 track. The results in bold indicate the best result with regard to a single measure.

The PPHS runs on the CLEF 2014 track did not improve the ranking. The results are displayed in Table 6.8. The baseline run performed better with regard to most of the measures. However, the different runs performed quite similarly to each other.

Single Topic Analysis

It was interesting to investigate which specific queries improved performance and which queries performed poorly.

Figure 6.1 displays the difference between the PHS run in comparison to the TFIDF run on the CDS 2014 track for each single topic . Topic 1, among others, seems to have been positively affected. This topic belongs to the *diagnosis* category and its summary is:

> 58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back.

We analysed the top 10 result set of Topic 1 which is displayed in Table 6.9. The document ranked third was assessed as relevant, according to the ground truth, and gained four ranks. We looked into the assigned probabilities and found that the ICD-9-CM concept with the code `414` (other forms of chronic ischemic heart disease) was assigned the highest probability. That can be due to the content of the document as well as to the estimated incidence rate. The document ranked sixth, on the other hand, has the highest probability for concept `553` (other hernia of abdominal cavity without mention of obstruction or gangrene). According to the NHDS data

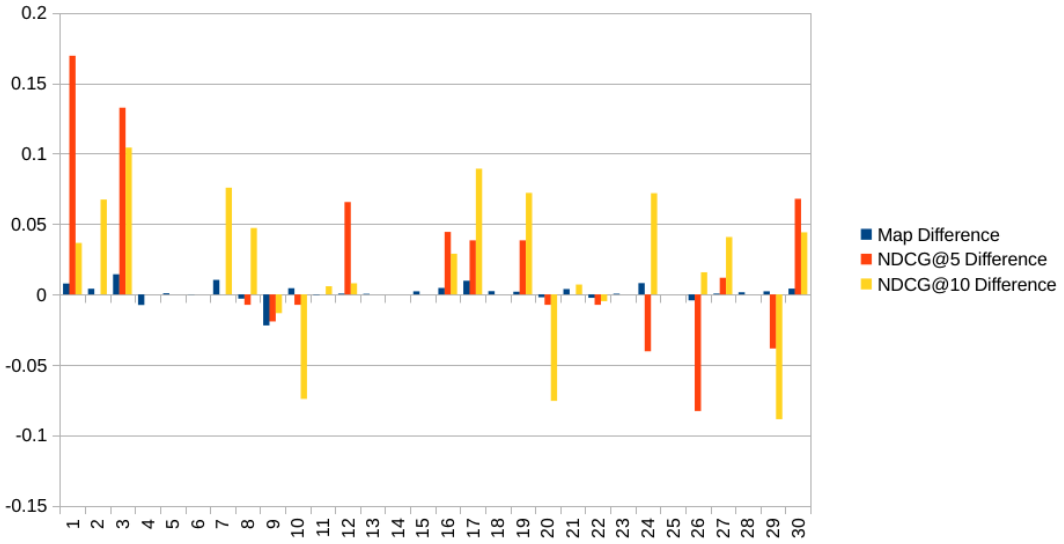

Figure 6.1: Single topic differences of PHS and baseline (PHS - baseline) on the CDS 2014 track.

set, however, the higher-ranking concept was the diagnosis of 840656 discharges, whereas the lower-ranking concept was diagnosed only 94315 times.

We also looked at Topic 26 in detail (see Table 6.10) because it derogated with regard to the NDCG@5 measure. The summary of this topic is:

> Group traveling to the Amazon rainforest, including 3 pregnant women. All members' immunizations are up-to-date but they require malaria prophylaxis.

We found that all top 10 documents had the highest probability for the concept `486` (pneumonia, organism unspecified), which is a common diagnosis. The NHDS data set estimates 912394 discharges. However, this diagnosis is not related to the topic. This topic represents a case where our assumption (documents about frequent diseases are more relevant) does not hold, because in this topic, the concept for Malaria `084` would be better suited, but the NHDS data set estimates only 1366 cases.

Summary and Comparison of all Runs

Considering all runs, we can say that the PPHS runs on the CLEF 2014 track did not perform better than baseline. Only the PPHS runs on the CDS 2014 track produced better-than-baseline results. We explain this difference by the better-performing baseline run of the CLEF 2014 track. It produced a higher bar to begin with. The baseline MAP of CLEF 2014 is 0.3951, whereas the baseline MAP of CDS 2014 is much lower, at 0.1208. Considering these results, we can not reject any of the null-hypotheses based on the runs on the CLEF 2014 track.

Rank	Doc ID	Title	Change	Assessment
1	2790183	Depression with Panic Episodes and Coronary Vasospasm	0	0
2	3258729	Epipericardial fat necrosis; a rare cause of pleuritic chest pain: case report and review of the literature	0	2
3	3853238	Chest pain in primary care: is the localization of pain diagnostically helpful in the critical evaluation of patients? - A cross sectional study	+4	2
4	2801475	Gender differences in presentation and diagnosis of chest pain in primary care	+1	2
5	3809224	Chest Pain as a presenting complaint in patients with acute myocardial infarction (AMI)	-1	2
6	2984347	Cough-induced abdominal intercostal hernia	-3	0
7	2731044	GPs' reasons for referral of patients with chest pain: a qualitative study	-1	0
8	3557637	An Extensive Stanford Type A Aortic Dissection Involving Bilateral Carotid and Iliac Arteries	0	0
9	3487367	Resource Utilization Reduction for Evaluation of Chest Pain in Pediatrics Using a Novel Standardized Clinical Assessment and Management Plan (SCAMP)	+1	0
10	2721934	A Correlation between Low Back Pain and Associated Factors: A Study Involving 772 Patients who Had Undergone General Physical Examination	+5	0

Table 6.9: Top 10 of Topic 1 from the PHS run on the CDS 2014 track.

The runs on the CDS 2014 track performed better than baseline. However, the differences are minimal. We proceeded to perform statistical tests for significance. We followed the recommendation of Smucker et al. and used Fisher's randomization test [91]. In this regard, we briefly recall the first null-hypothesis, which states that incorporating epidemiological data does not improve performance. We first tested the significance of the improvement of the MAP, which is 0.1208 for the baseline run and 0.1222 for the PHS run. Given that we have results that are better than the baseline, we rephrase it as a null-hypothesis for a two-sided randomized test:"the results of both runs were produced by equally well-performing IR systems."

We set the level of significance to $\alpha = 0.05$ and we produced 100000 permutations of the original dataset. In each permutation, we randomly switched the average precision values, which were calculated for single topics. In other words, we randomly picked n topics and switched the AP values. We than computed the MAP for both runs and calculated the absolute difference.

Rank	Doc ID	Title	Change	Assessment
1	2891813	Malaria in Brazil: an overview	0	2
2	3766208	The history of 20th century malaria control in Peru	+3	0
3	3224336	The position of mefloquine as a 21st century malaria chemoprophylaxis	0	2
4	3395692	Prophylaxis of Malaria	-2	2
5	2621393	Clinical development of new prophylactic anti-malarial drugs after the 5th Amendment to the Declaration of Helsinki	-1	2
6	3375659	Epidemiology of Imported Malaria in the Mediterranean Region	0	0
7	3117262	Clinical practice	+1	0
8	2020466	The low and declining risk of malaria in travellers to Latin America: is there still an indication for chemoprophylaxis?	+8	2
9	2823607	Pre-elimination of malaria on the island of..	0	0
10	3381416	Malaria among Military Personnel, French Guiana;2008	-3	0

Table 6.10: Top 10 of Topic 26 from the PHS run on the CDS 2014 track.

These steps were repeated a 100000 times and we counted how often the absolute difference was more extreme than the original difference. If the null-hypothesis holds, then it is equally likely to observe a more extreme difference of MAP than the observed difference. The p-value, therefore, is estimated as:

$$\hat{p} = \frac{\sum_{i=1}^{10000} 1\left[R_i \geq \text{OR}\right]}{100000}, \tag{6.5}$$

where R_i denotes the absolute difference between the MAP values of both runs, but with switched AP values for randomly chosen topics in the i^{th} iteration. The constant OR denotes the original difference.

The p-value for the significance of MAP is 0.2386. The p-value for the NDCG@5 measure is lower at 0.1945. However, both p-values are still far above the significance level.

The results of the control run (all incidence rates were set to 1) were a slightly lower than the results of the PHS run. If we set aside the baseline for a moment, we can investigate if the incidence rates play a significant role. We calculated the p-value for the MAP measure, which is 0.09897. Again, the p-value is above the significance level.

CHAPTER 7

Conclusions and Future Work

This chapter revisits the results of this thesis and draws conclusions from the work that was conducted. We will also discuss future work and point to directions, where we think research on IR models based on epidemiological data can be improved.

7.1 Summary

We began this thesis with the assumption that disease frequencies play a role, when a user of a health search engine assesses the relevance of documents. This assumption was partly motivated by the fact that other scientific fields already produce data, which can be used to estimate frequencies. Another peculiarity of the medical domain is the availability of controlled vocabularies, which are accessible for automatic processing. We introduced the three-stages model in Chapter 2 and, from its perspective, concluded that such controlled vocabularies are a prerequisite. From this point of view, our approach can be considered to be economical, since it exploits artifacts, which already exist in the medical domain.

The review of related work revealed that various authors investigated methods to improve IR effectiveness with the help of controlled medical vocabularies. Some approaches expand queries with the concepts of a thesaurus, while others improve indexing by mapping free text to canonical terms. Trieschnigg et al. compared various methods to categorize documents into MeSH terms. The authors could also demonstrate improved ad hoc IR effectiveness, based on automatically categorized biomedical documents [105]. Our approach can be seen as going a step further. We classify documents into disease categories, as well, but additionally integrate another data source, namely epidemiological studies.

Based on the first hypothesis, we inferred, that if the disease frequencies are contextualized, and the age and sex of a patient considered, then a health search engine can be further improved. We found that research in the personalization of IR systems, is currently active, and various approaches have been published recently. The models of Sontag et al. provided us with important input in formulating the PPHS model [93]. Based on their work, we developed a model on a

formal basis. The requirements of this model, guided the review of epidemiological studies. Our findings showed that the ideal data sources, which can be automatically processed, do not exist, or are not publicly available. Nevertheless, we declared the NHDS of 2007 as good enough.

We presented a reference implementation and demonstrated how the various components can be integrated. With this reference implementation, we were able to evaluate our approach with two evaluation frameworks, that were created very recently.

From the summary perspective, we also want to note that according to our review, there exists no scientific literature that demonstrates an IR approach, which incorporates epidemiological data. Therefore, any outcome of this thesis can be seen as a valuable contribution.

7.2 Conclusions

We conducted several experiments within two evaluation frameworks. We compared our results with the performance of state-of-the-art retrieval models. Though we could demonstrate slight improvements with regard to one of the evaluation tracks, the differences were not statistically significant at a $\alpha = 0.05$ level.

It is hard to assess our preliminary assumption, which is that the relevance of documents and frequencies of diseases correlate. A critical part of our model is identifying the diseases, which a document covers. In our experiments, we used our prototype implementation, which estimates this link similarly to approaches of other authors. Our implementation is challenged by assigning ICD-9-CM concepts to documents. We were constrained to this particular vocabulary, since the disease statistics are encoded with this classification. Other authors presented approaches based on the MeSH thesaurus, for which a large data set of manually annotated documents exist. Therefore, it was possible to evaluate their methods, since there is enough data that serves as the ground truth. We could not evaluate our concept-mapping component, since we had no ICD-9-CM annotations for documents (except the Wikipedia articles). Some authors investigated approaches, based on ICD-9-CM annotated data-sets, but these are out of date, unavailable and, rendered for a specific sub-task [118]. We also attempted to translate the ICD-9-CM concepts to MeSH terms with the UMLS API, but this approach was unreliable and in most cases, there was no unambiguous translation. Given, that the quality of our conceptual stage is not quantifiable, the results are biased and have to be carefully interpreted.

We now return to our research questions, which were presented in Section 1.2: At this time, we can not demonstrate statistically significant improvements of IR systems, that re-rank results, based on relevance signals, which are inferred from epidemiological data. The NHDS was found to be the most suitable source for such data. However, the data resources that are publicly available do not fulfil all requirements to the full extent.

The integration of profiles that contain the sex and the age of a patient, did not improve the results, either. We presented a quantitative estimation of the amount of information, which is gained by knowing these two attributes in Section 3.3. However, with the results of our experiments, it is still not clear if this information gain can be consumed in an automated way.

7.3 Future Work

We recommend continued research on the potential of epidemiological statistics as a relevance signal. In order to isolate the influence of the conceptual stage, we propose to either use manually annotated test collections, or integrate a concept-mapping technique that produces quantifiable estimates. With the MeSH thesaurus as a conceptual layer, the development of concept-mapping modules can access a large pool of manually annotated documents. However, then the key challenge is to estimate epidemiological statistics for MeSH concepts, for which we could not find a suitable data source.

We further suggest, investigating the relationship between disease frequencies and document relevance, independent of any queries. One possible approach would be to manually annotate the documents of IR test collections, for which relevance assessments are available. Subsequently, the incidence rate distributions can be estimated for each document. Based on this data, it might be possible to quantify the relationship between relevance and incidence rates.

Another open question is the optimal granularity of disease concepts. The mentioned concept vocabularies are organized in hierarchical structures. We chose the three-digit level of the ICD-9-CM tree, because it was the most fine-grained level for which Wikipedia articles exist. However, it is possible that a more general partitioning would yield better results.

Bibliography

[1] Omar B. Ahmad, Cynthia Boschi-Pinto, Alan D. Lopez, Christopher JL Murray, Rafael Lozano, Mie Inoue, and others. *Age standardization of rates: a new WHO standard*. World Health Organization Geneva, 2001.

[2] Roberto Antonicelli, Rosaria Gesuita, and Enrico Paciaroni. Sexual dimorphism in arterial hypertension: an age-related phenomenon. *Archives of Gerontology and Geriatrics*, 29(3):283–289, February 2000.

[3] Apache Lucene Project. http://lucene.apache.org. Accessed: 2015-02-04.

[4] Apache Solr. http://lucene.apache.org/solr/. Accessed: 2015-02-04.

[5] A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*, pages 17–21, 2001.

[6] A. R. Aronson and T. C. Rindflesch. Query expansion using the UMLS Metathesaurus. *Proceedings of the AMIA Annual Fall Symposium*, pages 485–489, 1997.

[7] Alan R. Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, May 2010.

[8] Alan R. Aronson, Thomas C. Rindflesch, and Allen C. Browne. Exploiting a Large Thesaurus for Information Retrieval. In *RIAO*, volume 94, pages 197–216, 1994.

[9] NJ Belkin. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, (5):133–143, 1980.

[10] Adam Berger and John Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 222–229. ACM, 1999.

[11] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270, January 2004.

[12] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, April 1998.

[13] Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 33–40. ACM, 2000.

[14] Centers for Disease Control and Prevention. http://www.cdc.gov/. Accessed: 2015-02-04.

[15] Centers for Disease Control and Prevention NHDS data. ftp://ftp.cdc.gov/pub/Health_statistics/NCHS/Datasets/NHDS/. Accessed: 2015-02-04.

[16] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393, 1999.

[17] Shiwen Cheng, Vagelis Hristidis, and Michael Weiner. Leveraging User Query Sessions to Improve Searching of Medical Literature. *AMIA Annual Symposium Proceedings*, 2013:214–223, November 2013.

[18] James J. Cimino. Linking patient information systems to bibliographic resources. *Methods of information in medicine*, 35:122–126, 1996.

[19] Thomas M Cover and Joy A Thomas. Elements of information theory 2nd edition. *Willey-Interscience: NJ*, 2006.

[20] W Bruce Croft and David J Harper. Using probabilistic models of document retrieval without relevance information. *Journal of documentation*, 35(4):285–295, 1979.

[21] Lindberg Da, Humphreys Bl, and McCray At. The Unified Medical Language System. *Methods of information in medicine*, 32(4):281–291, August 1993.

[22] Mariam Daoud, Lynda Tamine-Lechani, and Mohand Boughanem. Learning User Interests for a Session-based Personalized Search. In *Proceedings of the Second International Symposium on Information Interaction in Context*, IIiX '08, pages 57–64, New York, NY, USA, 2008. ACM.

[23] Hashem B. El–Serag and K. Lenhard Rudolph. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastroenterology*, 132(7):2557–2576, 2007.

[24] David A Evans, Kimberley Ginther-Webster, Mary Hart, Robert G Lefferts, and Ira Monarch. Automatic indexing using selective NLP and first-order thesauri. In *RIAO*, volume 91, pages 624–643, 1991.

[25] J Ferlay, I Soerjomataram, M Ervik, R Dikshit, S Eser, C Mathers, M Rebelo, DM Parkin, D Forman, and F Bray. GLOBOCAN 2012 v1. 0. *Cancer incidence and mortality worldwide: IARC CancerBase*, (11), 2013.

[26] Susannah Fox and Maeve Duggan. Health online 2013. *Health*, 2013.

[27] Gene Ontology Consortium. http://geneontology.org/. Accessed: 2015-02-04.

[28] Lorraine Goeuriot, Liadh Kelly, Gareth JF Jones, Guido Zuccon, Hanna Suominen, Allan Hanbury, Henning Müller, and Johannes Leveling. Creation of a new evaluation benchmark for information retrieval targeting patient information needs. 2013.

[29] Lorraine Goeuriot, Liadh Kelly, Wei Li, Joao Palotti, Pavel Pecina, Guido Zuccon, Allan Hanbury, Gareth J. F. Jones, and Henning Mueller. ShARe/CLEF eHealth evaluation lab 2014, task 3: user-centred health information retrieval. In *CLEF 2014 Working Notes*, volume 1180, Sheffield, UK, September 2014. CEUR-WS.

[30] Warren R. Greiff. A Theory of Term Weighting Based on Exploratory Data Analysis. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 11–19, New York, NY, USA, 1998. ACM.

[31] A Hanbury and H Müller. Khresmoi–multimodal multilingual medical information search. *MIE village of the future*, 2012.

[32] Leonard Hayflick. The not-so-close relationship between biological aging and age-associated pathologies in humans. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, 59(6):B547–550; discussion 551–553, June 2004.

[33] Niel Hens, Ziv Shkedy, Marc Aerts, Christel Faes, Pierre Van Damme, and Philippe Beutels. *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data: A Modern Statistical Perspective*. Springer Science & Business Media, October 2012.

[34] William Hersh. *Information Retrieval: A Health and Biomedical Perspective*. Springer, New York, NY, auflage: 3rd ed. 2009 edition, November 2008.

[35] William Hersh, Susan Price, and Larry Donohoe. Assessing thesaurus-based query expansion using the UMLS Metathesaurus. In *Proceedings of the AMIA Symposium*, page 344. American Medical Informatics Association, 2000.

[36] William R Hersh and Robert A Greenes. SAPHIRE—an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. *Computers and Biomedical Research*, 23(5):410–425, 1990.

[37] William R. Hersh, David H. Hickam, and T. J. Leone. Words, concepts, or both: optimal indexing units for automated information retrieval. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 644. American Medical Informatics Association, 1992.

[38] Djoerd Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *Research and advanced technology for digital libraries*, pages 569–584. Springer, 1998.

[39] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[40] Frederick Jelinek and Robert L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *In Proceedings of the Workshop on Pattern Recognition in Practice*, pages 381–397, Amsterdam, The Netherlands: North-Holland, May 1980.

[41] Evangelos Kanoulas, Ben Carterette, Mark Hall, Paul Clough, and Mark Sanderson. Session track 2011 overview. In *20th Text REtrieval Conference Notebook Proceedings (TREC 2011)*, 2011.

[42] Nancy Krieger. Genders, sexes, and health: what are the connections—and why does it matter? *International Journal of Epidemiology*, 32(4):652–657, August 2003.

[43] Robert Krovetz and W. Bruce Croft. Lexical Ambiguity and Information Retrieval. *ACM Trans. Inf. Syst.*, 10(2):115–141, April 1992.

[44] Leah S. Larkey and W. Bruce Croft. Combining classifiers in text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 289–297. ACM, 1996.

[45] Nut Limsopatham, Craig Macdonald, and Iadh Ounis. Learning to Combine Representations for Medical Records Search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 833–836, New York, NY, USA, 2013. ACM.

[46] Nut Limsopatham, Craig Macdonald, and Iadh Ounis. A task-specific query and document representation for medical records search. In *Advances in Information Retrieval*, pages 747–751. Springer, 2013.

[47] Clive Loader. *Local regression and likelihood*, volume 47. Springer New York, 1999.

[48] Henry J. Lowe and G. Octo Barnett. MicroMeSH: a microcomputer system for searching and exploring the National Library of Medicine's Medical Subject Headings (MeSH) Vocabulary. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 717. American Medical Informatics Association, 1987.

[49] Yumao Lu, Fuchun Peng, Xing Wei, and Benoit Dumoulin. Personalize Web Search Results with User's Location. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 763–764, New York, NY, USA, 2010. ACM.

[50] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[51] Margaret Jean Hall, Ph.D., Carol J. DeFrances, Ph.D., Sonja N. Williams, M.P.H., Aleksandr Golosinskiy, M.S., and Alexander Schwartzman. National Hospital Discharge Survey: 2007 Summary. Technical Report 29, Division of Health Care Statistics, October 2010.

[52] M. E. Maron and J. L. Kuhns. On Relevance, Probabilistic Indexing and Information Retrieval. *J. ACM*, 7(3):216–244, July 1960.

[53] Nicolaas Matthijs and Filip Radlinski. Personalizing Web Search Using Long Term Browsing History. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 25–34, New York, NY, USA, 2011. ACM.

[54] MEDLINE Citation Counts. http://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html. Accessed: 2015-02-04.

[55] Donald Metzler, Victor Lavrenko, and W. Bruce Croft. Formal multiple-bernoulli models for language modeling. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 540–541. ACM, 2004.

[56] D Miles Wyndham. A History of the National Library of Medicine: The Nation's Treasury of Medical Knowledge, Bethesda, Md.: US Dept. of Health and Human Services. *Public Health Service, National Institutes of Health, National Library of Medicine*, 1982.

[57] David RH Miller, Tim Leek, and Richard M. Schwartz. A hidden Markov model information retrieval system. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 214–221. ACM, 1999.

[58] Randolph A. Miller, Filip M. Gieszczykiewicz, John K. Vries, and Gregory F. Cooper. CHARTLINE: providing bibliographic references relevant to patient charts using the UMLS Metathesaurus Knowledge Sources. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 86. American Medical Informatics Association, 1992.

[59] Robert Moskovitch, Susana B. Martins, Eytan Behiri, Aviram Weiss, and Yuval Shahar. A Comparative Evaluation of Full-text, Concept-based, and Context-sensitive Search. *Journal of the American Medical Informatics Association : JAMIA*, 14(2):164–174, 2007.

[60] Robert Moskovitch and Yuval Shahar. Vaidurya: a multiple-ontology, concept-based, context-sensitive clinical-guideline search engine. *Journal of Biomedical Informatics*, 42(1):11–21, 2009.

[61] Prakash Nadkarni, Roland Chen, and Cynthia Brandt. UMLS Concept Indexing for Production Databases A Feasibility Study. *Journal of the American Medical Informatics Association*, 8(1):80–91, January 2001.

[62] Goran Nenadic, Irena Spasic, and Sophia Ananiadou. Mining Biomedical Abstracts: What's in a Term? In Keh-Yih Su, Jun'ichi Tsujii, Jong-Hyeok Lee, and Oi Yee Kwong, editors, *Natural Language Processing – IJCNLP 2004*, number 3248 in Lecture Notes in Computer Science, pages 797–806. Springer Berlin Heidelberg, 2005.

[63] NLM MeSH Browser. http://www.nlm.nih.gov/mesh/2014/mesh_browser/MBrowser.html. Accessed: 2015-02-04.

[64] NLM SPECIALIST Lexicon. http://www.nlm.nih.gov/pubs/factsheets/umlslex.html. Accessed: 2015-02-04.

[65] Nokogiri. http://www.nokogiri.org/. Accessed: 2015-02-04.

[66] Sabine Oertelt-Prigione and Vera Regitz-Zagrosek. *Sex and Gender Aspects in Clinical Medicine*. Springer Science & Business Media, November 2011.

[67] Open Access Subset of PubMed Central. http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/. Accessed: 2015-02-04.

[68] Open Directory Project. http://www.dmoz.org/. Accessed: 2015-02-04.

[69] João Palotti, Allan Hanbury, and Henning Müller. Exploiting Health Related Features to Infer User Expertise in the Medical Domain. 2014.

[70] PCWorld. http://www.pcworld.com/article/261363/google_changes_im_feeling_lucky_button.html. Accessed: 2015-02-04.

[71] James Pitkow, Hinrich Schütze, Todd Cass, Rob Cooley, Don Turnbull, Andy Edmonds, Eytan Adar, and Thomas Breuel. Personalized Search. *Commun. ACM*, 45(9):50–55, September 2002.

[72] Jay M. Ponte and W. Bruce Croft. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, New York, NY, USA, 1998. ACM.

[73] Miquel S. Porta, Sander Greenland, Miguel Hernán, Isabel dos Santos Silva, and John M. Last. *A Dictionary of Epidemiology*. Oxford University Press, 2014.

[74] Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[75] Prototype Source Code. https://github.com/tawan/pphs. Accessed: 2015-23-04.

[76] S. A. Reijneveld. Age in epidemiological analysis. *Journal of Epidemiology and Community Health*, 57(6):397–397, June 2003.

[77] Steffen Rendle, Zeno Gantner, Christoph Freudenthaler, and Lars Schmidt-Thieme. Fast Context-aware Recommendations with Factorization Machines. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 635–644, New York, NY, USA, 2011. ACM.

[78] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[79] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[80] Stephen E Robertson and Nicholas J Belkin. Ranking in principle. *Journal of Documentation*, 34(2):93–100, 1978.

[81] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, and others. Okapi at TREC-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109, 1995.

[82] Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash. *Modern Epidemiology*. Lippincott Williams & Wilkins, 2008.

[83] Ruby Language. https://www.ruby-lang.org. Accessed: 2015-02-04.

[84] Karl Lenhard Rudolph, Sandy Chang, Melissa Millard, Nicole Schreiber-Agus, and Ronald A. DePinho. Inhibition of experimental liver cirrhosis in mice by telomerase gene delivery. *Science*, 287(5456):1253–1258, 2000.

[85] Gerard Salton. The SMART retrieval system—experiments in automatic document processing. 1971.

[86] Gerard Salton. Developments in Automatic Text Retrieval. *Science*, 253(5023):974–980, August 1991.

[87] Gerard Salton and Michael J McGill. Introduction to modern information retrieval. 1983.

[88] Claude E Shannon and Warren Weaver. 1949the mathematical theory of communication. *Urbana: University of Illinois Press*.

[89] Xuehua Shen, Bin Tan, and ChengXiang Zhai. Context-sensitive Information Retrieval Using Implicit Feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 43–50, New York, NY, USA, 2005. ACM.

[90] Mário J. Silva, Bruno Martins, Marcirio Chaves, Ana Paula Afonso, and Nuno Cardoso. Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, 30(4):378–399, July 2006.

[91] Mark D. Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 623–632. ACM, 2007.

[92] Fei Song and W. Bruce Croft. A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 316–321. ACM, 1999.

[93] David Sontag, Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Susan Dumais, and Bodo Billerbeck. Probabilistic Models for Personalizing Web Search. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 433–442, New York, NY, USA, 2012. ACM.

[94] K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information Processing & Management*, 36(6):779–808, November 2000.

[95] K Sparck Jones, S Walker, and S. E Robertson. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information Processing & Management*, 36(6):809–840, November 2000.

[96] Padmini Srinivasan. Optimal document-indexing vocabulary for MEDLINE. *Information Processing & Management*, 32(5):503–514, 1996.

[97] Isabelle Stanton, Samuel Ieong, and Nina Mishra. Circumlocution in Diagnostic Medical Queries. In *SIGIR (Information Retrieval)*. ACM, July 2014.

[98] Statistik Austria. http://www.statistik.at/web_de/statistiken/gesundheit/krebserkrankungen/. Accessed: 2014-09-22.

[99] Bernard W. Stewart, Paul Kleihues, International Agency for Research on Cancer, and others. *World cancer report*, volume 57. IARC press Lyon, 2003.

[100] Bin Tan, Xuehua Shen, and ChengXiang Zhai. Mining Long-term Search History to Improve Search Accuracy. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 718–723, New York, NY, USA, 2006. ACM.

[101] Jaime Teevan, Susan T. Dumais, and Eric Horvitz. Personalizing Search via Automated Analysis of Interests and Activities. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 449–456, New York, NY, USA, 2005. ACM.

[102] Thomson Reuters Incidence & Prevalence Database. http://www.tdrdata.com/ipd/ipd_init.aspx. Accessed: 2015-02-04.

[103] TREC Clinical Decision Support Track. http://www.trec-cds.org/. Accessed: 2015-02-04.

[104] Dolf Trieschnigg, Djoerd Hiemstra, Franciska de Jong, and Wessel Kraaij. A cross-lingual framework for monolingual biomedical information retrieval. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 169–178. ACM, 2010.

[105] Dolf Trieschnigg, Piotr Pezik, Vivian Lee, Franciska de Jong, Wessel Kraaij, and Dietrich Rebholz-Schuhmann. MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics*, 25(11):1412–1418, June 2009.

[106] Ulrike Spree, Nadine Feißt, Anneke Lühr, Beate Piesztal, Nina Schroeder, and Patricia Wollschläger. *Semantic Search - State-of-the-Art-Überblick zu semantischen Suchlösungen im WWW*. Number 2 in Handbuch Internet-Suchmaschinen. D. Lewandowski, Heidelberg, 2011.

[107] UMLS Terminology Services. https://uts.nlm.nih.gov/metathesaurus.html. Accessed: 2015-02-04.

[108] U.S. National Library of Medicine. http://www.nlm.nih.gov/hmd/about/collectionhistory.html. Accessed: 2015-02-04.

[109] E. Voorhees and William R Hersh. Overview of the TREC 2012 medical records track. 2012.

[110] E. Voorhees and R. Tong. Overview of the TREC 2011 medical records track. In *The Twentieth Text REtrieval Conference Proceedings TREC*, 2011.

[111] Ryen W. White and Eric Horvitz. Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. *ACM Trans. Inf. Syst.*, 27(4):23:1–23:37, November 2009.

[112] Wikipedia Vector Space Model. http://en.wikipedia.org/wiki/Vector_space_model. Accessed: 2015-02-04.

[113] World Health Organization. http://apps.who.int/gho/data/?theme=main. Accessed: 2015-02-04.

[114] World Health Organization. http://www.who.int/classifications/icd/en/. Accessed: 2015-02-04.

[115] Andrew S. Wu, Bao H. Do, Jinsuh Kim, and Daniel L. Rubin. Evaluation of Negation and Uncertainty Detection and its Impact on Precision and Recall in Search. *Journal of Digital Imaging*, 24(2):234–242, November 2009.

[116] Helmut Wykticky and Manfred Skopec. Ignaz Philipp Semmelweis, The Prophet of Bacteriology. *Infection Control & Hospital Epidemiology*, 4(05):367–370, January 1983.

[117] Yiming Yang. An evaluation of statistical approaches to text categorization. *Information retrieval*, 1(1-2):69–90, 1999.

[118] Yiming Yang and Christopher G. Chute. An example-based mapping method for text categorization and retrieval. *ACM Transactions on Information Systems (TOIS)*, 12(3):252–277, 1994.

[119] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214, 2004.

[120] Guido Zuccon, Bevan Koopman, Anthony Nguyen, Deanne Vickers, and Luke Butt. Exploiting Medical Hierarchies for Concept-based Information Retrieval. In *Proceedings of the Seventeenth Australasian Document Computing Symposium*, ADCS '12, pages 111–114, New York, NY, USA, 2012. ACM.