FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Query Expansion Strategies for Laypeople-Centred Health Information Retrieval

Ricardo Daniel Soares da Silva

DISSERTATION PLANNING

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Carla Alexandra Teixeira Lopes

July 24, 2016

Query Expansion Strategies for Laypeople-Centred Health Information Retrieval

Ricardo Daniel Soares da Silva

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Sérgio Sobral Nunes

External Examiner: José Luís Oliveira
Supervisor: Carla Alexandra Teixeira Lopes

July 24, 2016

Abstract

One of the most common activities on the web is the search for health information. This activity has been gaining popularity among users, but the majority of them has no training in health care, which leads to difficulties in understanding medical terminology present in the documents and in formulating their queries.

In the field of Health Information Retrieval several works focus on query expansion to solve one of the biggest difficulties for users in the search of health information: formulating queries with limited knowledge of medical terminology. This lack of knowledge influence the formulation of queries and the expectations regarding the retrieved documents. The query expansion process complements the original query with additional terms.

The most popular way to define the relevance of a document is trough its topicality. However, if a document is relevant for a topic but the user does not comprehend its contents it ceases to be useful. The field of medicine is associated with complex and specific terms that lay people have difficulty in understanding. Considering only topical relevance is therefore insufficient.

The main objective of this thesis is to propose, evaluate and compare methods to improve health information retrieval by consumers.

We propose several query expansion methods using different sources and methodologies to identify which terms will be added to the original query. To reduce the problems caused by medico-scientific terminology, we re-rank the results obtained through the query expansion approaches based on the documents' readability. Readability is assessed through a score obtained with the most used readability metrics: SMOG, FOG and Flesch-Kincaid.

To evaluate these approaches we use a test collections provided by the CLEF initiative from their lab CLEF eHealth 2015. These approaches will also be evaluated on the CLEF eHealth 2016 collection when the relevance judgements are provided.

To evaluate the relevance of the retrieve documents we use precision at 10 (P@10) and nDCG at 10 (nDCG@10). For evaluating the readability we use the understandability-based Rank Biased Precision (uRBP) and its graded version (uRBPgr).

Overall all the approaches improve the relevance score. The MTI approach is the one that brought the best results, proving that medical concepts related to the query are good terms for the query expansion. Regarding the readability evaluations, most of the runs have low scores. The cause of this may be because the readability metrics give a score based on the number of polysyllabic words and sentence length, and this may not be well suited to evaluate documents of a specific area.

Resumo

Uma das atividades mais comuns na web é a pesquisa de informação relativa à saúde. Esta atividade tem vindo a ganhar popularidade entre os utentes, contudo a maioria destes não possuem uma formação na área da saúde, o que leva a dificuldades ao nível da terminologia utilizada nos documentos e na formulação de interrogações.

Na área da recuperação de informação de saúde já foram realizadas diversas investigações para suprimir uma das maiores dificuldades dos utilizadores na pesquisa de informação de saúde: a formulação de interrogações com conhecimento reduzido de terminologia médica. Esta falta de conhecimento influencia a formulação de interrogações e as expectativas dos documentos devolvidos pela pesquisa. A expansão de interrogações complementa a interrogação original com termos adicionais.

A forma mais popular para definir a relevância de um documento é descobrir se este contém informações sobre o tópico da pesquisa. Contudo, se um documento for relevante para um tópico mas o utilizador não compreender o seu conteúdo este deixa de lhe ser útil. A área da medicina está associada com termos complexos e específicos que os leigos têm dificuldade em compreender. Considerar apenas o tópico para medir a relevância de um documento é insuficiente.

O objetivo principal desta tese é propor, avaliar e comparar métodos para melhorar a recuperação de informação de saúde por parte dos consumidores.

Iremos propor diversos métodos de expansão de interrogação usando diferentes fontes e metodologias para identificar quais termos serão adicionados à interrogação original. Para propor uma solução para a diferença linguística entre terminologia médica e leigos, vamos reordenar os resultados obtidos através das abordagens de expansão de interrogações com base na legibilidade dos documentos. O calculo da legibilidade será obtido através das métricas de legibilidade mais utilizadas: SMOG, FOG e Flesch-Kincaid.

Para avaliar estas abordagens vamos utilizar a coleção de teste do CLEF eHealth 2015. Estas abordagens também serão avaliadas na coleção CLEF eHealth 2016 quando os julgamentos de relevância desta coleção forem disponibilizados.

Para avaliar a relevância dos documentos recuperados usaremos a *precision* a 10 ($P@10$) e o nDCG a 10 ($nDCG@10$). Para avaliar a legibilidade, vamos utilizar a *understandability-based Rank Biased Precision* (*uRBP*) e a sua versão graduada (*uRBPgr*).

No geral todas as abordagens melhoraram os resultados de relevância. A abordagem que utilizava o MTI foi a que obteve os melhores resultados, provando que conceitos médicos relacionados com a interrogação são bons termos para a expansão de interrogações. Em relação aos resultados de legibilidade, a maioria das *runs* obteve valores abaixo da *baseline*. A causa disto pode ser porque as métricas de legibilidade fazem a avaliação de um documento tendo em conta o tamanho das frases e das palavras, podendo este método não ser o melhor para avaliar documentos de uma área cientifica.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor Prof. Carla Teixeira Lopes, for the continuous support, patience and motivation.

A word of thanks goes to all my colleagues and friends in InfoLab, particularly José Devezas for his support and insights on Information Retrieval and Terrier, without him my work would take twice as long.

I would also like to thank all my friends from FEUP and Lans for their companionship and support over the years. Their friendship was what drove me to pursue several life goals.

My deepest thank goes to all my family for all the love and support they have given me.

Last, but not least, I am grateful to my parents, for giving me the opportunity of the education that took me to where I am today.

Ricardo Daniel Soares da Silva

Contents

List of Figures

LIST OF FIGURES

List of Tables

Abbreviations

IR	Information Retrieval
HIR	Health Information Retrieval
Web	World Wide Web
TF-IDF	Term Frequency - Inverse Document Frequency
P	Precision
nDCG	Normalized Discounted Cumulative Gain
TREC	Text REtrieval Conference
SIGIR	Special Interest Group on Information Retrieval
CLEF	Conference and Labs of the Evaluation Forum
NTCIR	NII Testbeds and Community for Information access Research
NIST	National Institute of Standards and Technology
HMD	History of Medicine Division
MeSH	Medical Subject Headings
UMLS	Unified Medical Language System
RBP	Rank Biased Precision
uRPB	understandability-based Rank Biased Precision
uRPBgr	understandability-based Rank Biased Precision graded
MTI	Medical Text Indexer
PRF	Pseudo Relevance Feedback
NLM	US National Library of Medicine
JWPL	Java Wikipedia Library
ICD	International Classification of Diseases
LDA	Latent Dirichlet Allocation

Chapter 1

Introduction

This first chapter situates the reader with the context of the problem and provides the motivation behind this project. The following sub-section 1.1 introduces the background and scope of the project. Afterwards, the sub-section 1.2 describes the main goals to accomplish. In sub-section 1.3 are mentioned the contributions made by this thesis. Finally, the sub-section 1.4 gives an overview of the organization of the document.

1.1 Context

According to the Pew Research Center report from 2013, of the 85% of U.S. adults that uses the Internet 72% have looked for health related information within the past year [Zic13, FD13]. In this way, the Internet has become the dominant source for health information.

Health Information Retrieval (HIR) focus on the application of IR concepts and techniques to the domain of healthcare. This field has largely evolved in the last few years. Habits of health professionals and consumers (patients, their family and friends) have been changing as a result of several factors like the increasing production of information in a digital format [LV01], the greater availability and the easier access to health information.

Most health related articles employ medical terminology, yet laypeople do not have the necessary knowledge to express their information need using such vocabulary, thus struggling to satisfy their information needs [ZKP+04]. This represents a language gap which is difficult to overcome either by laypeople or by experts, requiring different vocabularies for their information needs. One of the main reasons for failures of retrieval engines is this language gap [CYTDP06].

1.2 Objectives

The main objective of this thesis is to propose, evaluate and compare methods to improve health information retrieval by consumers.

Query expansion is the main approach used in this thesis. Query expansion (or term expansion) is the process of supplementing the original query with additional terms, and it can be considered as a method for improving retrieval performance [Eft96]. Different sources and methodologies will be used to identify which terms will be added by the query expansion.

To propose a solution to the language gap between medical terminology and laypeople we will re-rank the results obtained through the query expansion approaches based on the document readability. The readability score will be obtained through the most widely used readability metrics: SMOG, FOG and Flesch-Kincaid [SC01, MP82]. These metrics estimate the educational grade level necessary to understand a document.

1.3 Contributions

Two of our proposed approaches were submitted to the lab “CLEF eHealth 2016 Task 3 Patient-Centred Information Retrieval”, both approaches were re-ranked using the SMOG readability scores. This lab aims to evaluate systems that support people in understanding and searching for their health information [CLE16b]. The task “Patient-Centred Information Retrieval” is split into three subTasks: ad-hoc search, query variation and multilingual search [CLE16b]. We participated in the ad-hoc search and query variation subTasks. At the time of writing this document the paper submitted to the CLEF eHealth 2016 was already approved.

1.4 Thesis Outline

Besides the Introduction, this thesis contains five more chapters. Chapter 2 describes the state-of-art and work related to the subject. Chapter 3 explains the methods and describe the test collections used to evaluate the query expansion approaches. On Chapter 4 are illustrated the experiments and the results done with each approach. On Chapter 5 a comparison between approaches is made to evaluate their efficiency. Finally, Chapter 6 presents the concluding remarks, taking into account the thesis objectives. Also, future expansions for this project are presented in this last chapter.

Chapter 2

Literature Review

In this Chapter we provide some background about the concepts, methods and techniques that we use in the rest of this thesis. The main goal of this chapter is to present a comprehensive analysis of the work that has been done by other researchers and developers in the fields of Information Retrieval and Query Expansion.

2.1 Information Retrieval

The term Information Retrieval (IR) was adopted by Calvin Mooers who defined it as:

> "The name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him." [Moo51]

Understand the meaning of this term is not an easy task to accomplish, because even the fact of searching a dictionary for the meaning of a word is a form of Information Retrieval. A more recent definition defines IR as:

> "...finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). " [MRS08]

The concept of unstructured data refers to the types of documents that don't have a clear structure, making the processing of these documents, in a computer, an arduous task.

At the other hand structured data, usually associated with relational databases following a rigid structure which is easily interpreted by a computer [MRS08].

Information Retrieval was an activity in which a few groups, such as librarians and the legal community were involved, however, with the appearance of the World Wide Web (Web) thousands of people are involved, daily, in the act of IR when they use a search engine, thus expanding the interest in this area to a wider audience, ceasing to be a practice carried out by specific communities [MRS08, Lop08].

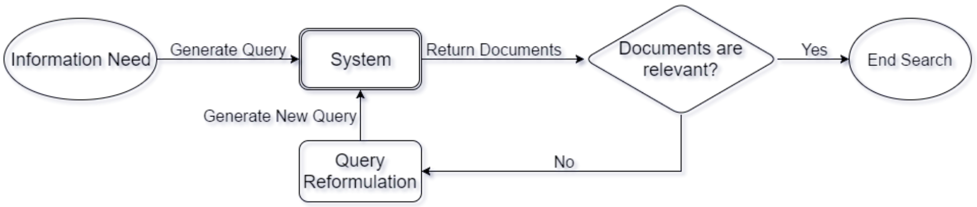


Figure 2.1: Example of the stages traversed in IR from the viewpoint of a user

Web pages are considered to be semi-structured, i.e., contains fluid text blocks (unstructured) in conjunction with certain tags such as `<h1>` and `<cite>` which defines titles and quotes, and links (`<a>`) which allowed the development of algorithms like PageRank [Fra11]. It is this type of structure that assisted the growth of the Web Information Retrieval practices, by simplifying the process of such documents compared to the "unstructured" ones.

The IR process, from the user point of view, follows the steps described in Figure 2.1, as soon as a user has an information need he provides to an information retrieval system a formulated question from his necessity, then the system returns to the user the most relevant documents, if the returned documents do not fully satisfy the need of the user this can rephrase his question and provide it back to the system. To retrieve the most relevant documents for a user, the IR system requires various modules to interact with each other. These modules can be divided, in general, in three operations: indexing, retrieval and ranking. Indexing is responsible for organizing and storing data to enable quick and easy access for research that, in turn, retrieves indexed information to satisfy the user's information needs. Finally we have the ranking which is an optional but very important task on Web IR, its task is to sort the returned documents, based on heuristics, by their potential relevance to the user.

2.1.1 Indexing

For a system to be able to perform efficient searches over a collection of documents it needs a specific structure called index. This index consists of a dictionary of terms (sometimes also referred to as a vocabulary or lexicon) along with a list that records which documents the term occurs in. Each item in the list is conventionally called a posting [MRS08].

Statistics are also stored in the dictionary, document frequency (number of documents which contain each term) which represents the length of each postings list. This information is not vital but it allows an improvement to the efficiency of the search engine at query time [MRS08]. The postings are secondarily sorted by the document ID. This provides the basis for efficient query processing. This index (or inverted index) structure is the most efficient for supporting ad-hoc text search [MRS08].

During the indexing phase it is possible to perform some pre-processing techniques to reduce the final size of the index. The techniques most frequently used are the elimination of stop-words and stemming.

The stop-words are words that appear frequently in the collection however are not relevant, such as “a”, “are”, “is” and “the”. The elimination of these words from the index is a good way to reduce its size.

The collection can have variations of the user query terms such as the plural or one of its conjugations. To solve this issue, instead of indexing the word as stated in the document, a stemming is performed, i.e., the original word affixes are removed [MRS08]. This process will reduce the number of words in the vocabulary because several words will match a primitive term. We can consider as an example the word “legal”, which can be obtained through the stemming of words such as “illegal”, “legally”, “legalization”, among others.

2.1.2 Retrieval

The Boolean model is a classical Information Retrieval model and, at the same time, the first and most adopted one [MRS08]. This model is based on Boolean logic which considers both the documents to be searched and the user’s query are as sets of terms. Retrieval is based on whether or not the documents contain the query terms. This model is one of the most basic, easy to implement and with an intuitive concept, however, it only records term presence or absence, in most cases is preferred giving more weight to documents that have a term several times as opposed to ones that contain it only once. To be able to do this is necessary to use the term frequency information in the postings lists [MRS08]. The Boolean model retrieves a set of matching documents, but commonly the user wishes to have the returned results ordered (or “ranked”). This requires having a mechanism for determining a document score which encapsulates how good a match a document is for a query, this mechanism is called a weighting model.

TF-IDF (Term Frequency-Inverse Document Frequency) is an common weighting model used in information retrieval. This model determines the score of a document two key factors: (1) how frequently a term occurs in a document (TF) and (2) how rarely a term occurs in the document collection (IDF) [MRS08]. The TF-IDF weighting model assigns to term t a weight in document d given by:

$$TF\text{-}IDF_{t,d} = TF_{t,d} \times IDF_t \tag{2.1}$$

where $TF\text{-}IDF_{t,d}$ assigns to a given t term a weight in document d that is: (1) highest when t occurs many times within a small number of documents, (2) lower when the term occurs fewer times in a document, or occurs in many documents and (3) lowest when the term occurs in virtually all documents [MRS08]. The final score of a document is the sum of all the $TF\text{-}IDF_{t,d}$ for every query term (qt_i):

$$Score(d,qt) = \sum_{qt=0}^{i} TF\text{-}IDF_{qt_i,d} \tag{2.2}$$

2.1.3 Evaluation

One of the methods to evaluate an IR system is to analyse whether the documents returned for a particular question are relevant or not. This method compares, for a given question, the set of documents in the collection with relevant documents returned by the system.

The most basic formulas for assessing the returned documents quality are the precision and recall. If R is the set of documents that are relevant to a given query q in a collection I and, A is the set of documents returned by the system, and Ra the returned set of documents that are relevant to the query q. We can define the precision as the ratio between the relevant returned documents (Ra) and the set of returned documents (A), and recall as the ratio between the relevant documents returned (Ra) and the set of documents relevant (R) [MRS08]. (See Equations 2.3 e 2.4)

$$Precision = \frac{|Ra|}{|A|} \tag{2.3}$$

$$Recall = \frac{|Ra|}{|R|} \tag{2.4}$$

As these values alone can not be sufficient to evaluate an IR system (a system can have a recall of 1 if it returns all the documents), is possible to analyse a combination of both through an precision-recall curve (Figure 2.2). Through these curves it is concluded that the precision is inverse to the recall, meaning, increasing the a performance of one of them will decrease the performance of the other.

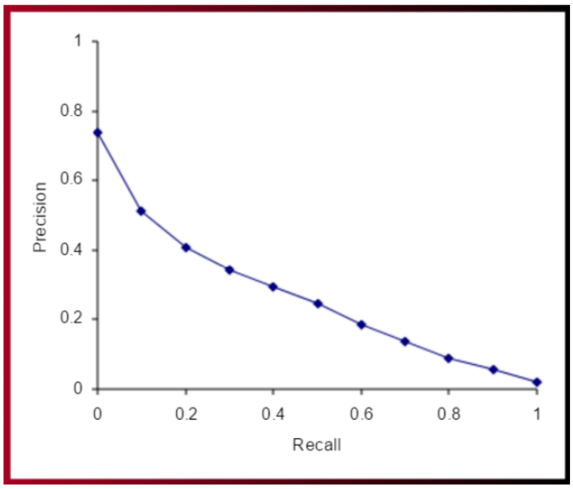


Figure 2.2: SabIR/Cornell 8A1 11pt precision do TREC 8 (1999) [MRS08]

Another evaluation measurement involves determining the precision of a number of documents, for example, determine the precision in the first five documents. This method is called precision at n ($P@n$) and represents the quality of a reply, since the user typically can only see the first n documents and not the whole set of returned documents.

According to Järvelin e Kekäläinen [JK00] one of the most popular measures for evaluating the IR is the nDCG (Normalized Discounted Cumulative Gain). The nDCG has two advantages over other methods. First, the nDCG is compatible with non-binary relevance assessments. While the precision only makes the distinction between "relevant" or "not relevant" documents, the nDCG can consider a document to be partially relevant. Second, the nDCG imposes a discount function on the rank position of the documents, while the precision uses a uniform value for all positions. This is a very important feature for the Web, where most attention is given to the top results on the list of retrieved documents. Like precision, nDCG can be used up to a given number of documents, nDCG@n.

2.1.4 Test Collection

Test collection are mostly used on IR researches to assess the effectiveness of the system. Because the use of test collections was greatly accepted by researchers that many conferences and meetings are devoted purely to their use, including several international conferences which have run since the early 1990s [San10].

A test collection is usually divided into three components: (1) a collection of documents; each document is given an unique identifier, a docid; (2) a set of topics (queries); each query is given an id (qid); and (3) a set of relevance judgements (qrels or query relevance set) composed of a list of qid/docid pairs, detailing the relevance of documents to a given topic [San10].

Having an appropriate test collection, an IR researcher indexes the document collection and then submits each the topic (query) into the system resulting in a list of docids known as a *run*. Then the content of each run the is compared with the qrels to asses which of the retrieved documents were given a relevance judgement [San10]. Finally, an evaluation measure, like P@n or nDCG@n, is used to quantify the effectiveness of that run. (See Figure 2.3)

Test collections allow researchers to locate points of failure in their retrieval system, but more commonly, these collections are used to assess the effectiveness of multiple retrieval systems [San10]. Using the same test collection it is possible to compare systems developed by different researchers or compare different configurations of the same system [San10].

2.2 Query Expansion in Information Retrieval

Query expansion (or term expansion) is the process of supplementing the original query with additional terms, and can be considered as a method to improve the retrieval performance. This method can be applied regardless the used retrieval technique. The initial query, provided by the user, may be an incomplete representation of the information need, by itself or in relation to the documents in the collection [Eft96].

The query expansion process can be divided in two stages: (1) initial query formulation and (2) query reformulation [Eft96]. At the initial query formulation stage, the user writes a query and submits it to the system. At the query reformulation stage, having some results from the first stage, the user manually or the, the system automatically, or both, adjust the initial query adding

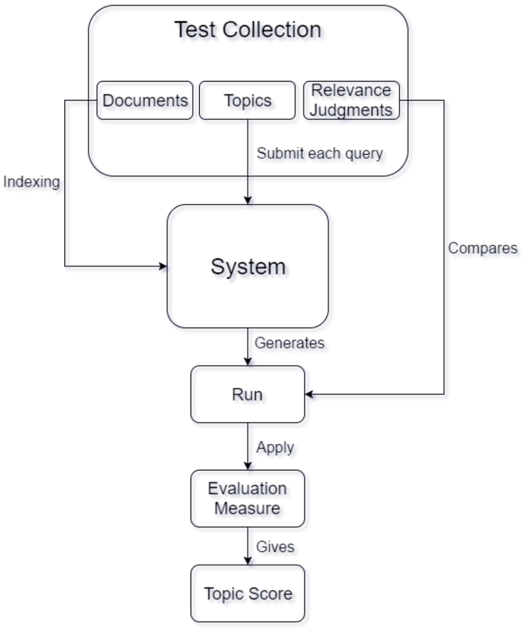


Figure 2.3: Use of a Test Collection

more terms with the goal of improving the final outcome [Eft96]. The query expansion can be performed manually, automatically or interactively (semi-automatic). (as depicted in Figure 2.4)

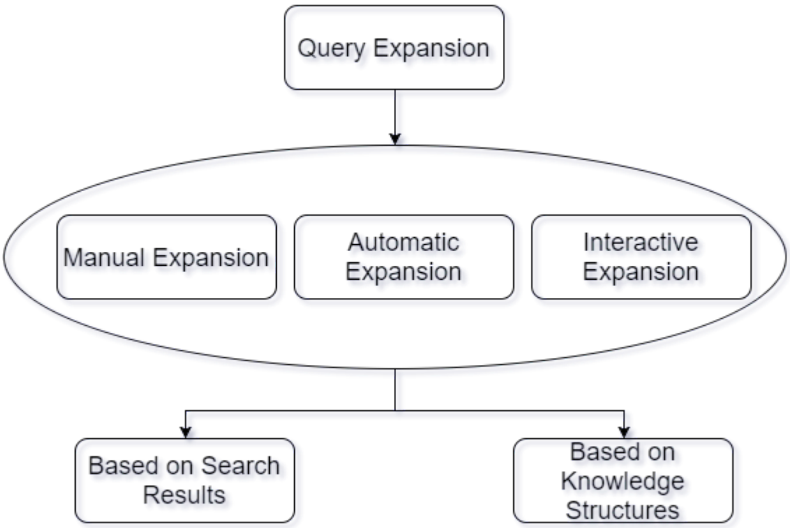


Figure 2.4: Query Expansion

Manual query expansion involves more than just a straightforward combination of terms. It is increasingly complicated, dynamic and its success varies considerably depending on the abilities of the individual searcher. The user has to learn how to use the existing systems and their query languages to develop ways of information seeking which are adjusted to his information needs [Eft96]. This requires a vast knowledge of the subject, which in most cases this kind of knowledge is the type of knowledge the user intends to find.

In automatic query expansion, the system itself is responsible for expanding the initial query based on some method that retrieves the new terms from a specific source like a thesauri or a dictionary [Eft96]. This method does not require a prior knowledge from the user to expand the query thereby facilitating the search process.

In interactive query expansion (or semi-automatic) there are two parties responsible for determining and selecting terms for the expansion [Eft96]. One is the retrieval system, which acts in the same manner as the system in the automatic query expansion, retrieving new terms from a specific source. The other is the user, that chooses the terms to be appended to the query from a ranked list of terms.

There are two key elements that should be considered when applying the query expansion process, which are the source that will provide the new terms and which method will be used to select those terms. One type, called relevance feedback, is based on search results. Documents that are returned in a previous iteration, and that were considered relevant, become the source of new terms to be added to the original query. The other type involves using knowledge structures that are independent of the retrieval process like thesauri, dictionaries and lexicons .

The query expansion process can be applied to any topic, this can be shown in articles submitted to the Text Retrieval Conference (TREC) which includes news [LF15], micro-blogs [YHM14], support to clinical decision [ZHF15] and many other tracks. TREC began in 1992 as part of the Tipster Text program [NIS16], with the sponsorship of the US Department of Defense and the NIST (National Institute of Standards and Technology). The purpose of TREC is to support and promote research in the field of Information Retrieval, providing the necessary infrastructure for large scale evaluation of text retrieval methodologies [TRE16b].

Lu and Fang [LF15] used Wikipedia pages to build expansion models that were related to events in the query. Their main objective is proving that the event related entities, which are the entities mentioned in the queries, are essential when defining the event. For example the query “Vauxhall helicopter crash”, without “Vauxhall” or “helicopter”, makes the event less defined and the results become more generic. They propose that both event type and event related entities should be considered when expanding a query.

You, Huang and Mu [YHM14] propose query expansion methodologies for microblogs because they have unique features different from traditional web pages or database documents. This methodology adjusts the ranking score of a document considering how close the document time stamp is to the event, using Google as an external data corpus.

2.3 Query Expansion in Health Information Retrieval

The Internet has been recognized as an important source of health related information [Ric06]. For this reason thousands of people have adopted the practice of searching the web for information related to their health and the health of their family and friends. Due to the vast amount of information on the Web this practice is not always efficient. One of the reasons is the difference in the knowledge that each user has about health topics. On one hand we have specialists such as doctors, requiring documents with a more scientific language, on the other hand we have lay users who need content to be less technical and easier to understand.

Big companies like Google and Microsoft have been investing in the area of Health Information Retrieval [Mic16, Bro16], however, not all systems that were developed were able to have the desired adherence as is the case of Google Health. This system intended to offer users information about their health and well-being, however, over the years they found out that the system did not achieve the expected impact and that the adherence of this new service was greatly reduced, for these reasons Google decided to discontinue Google Health [Bro16].

Adam Bosworth [Bos16] refers three main features that a user should expect from a Health Information Retrieval system for it to be considered a good health system. The first is the discovery, in which a user should be able to find the most relevant information. Then comes the action, which gives users access to a personalized service for the best possible support. Finally we have the community, users must learn from those who are in similar situations in order to make correct decisions.

Knowledge has an impact on the formulation of the queries and the kind of documents that the user wants. On average, a user writes between two or three terms in a query, regardless of their information need [SWJS01]. When searching for health-related information, limited knowledge of medical vocabulary makes users simplify their queries which in turn makes the system return more generic documents that are not relevant or don't meet the user needs. The query expansion process tries to diminish this by expanding the original query with medical related terms from sources like the UMLS metathesaurus, MedlinePlus pages or Medical Subject Headings (MeSH). The development of new approaches for query expansion in the field of medicine, has been the focus of several articles submitted to Information Retrieval workshops belonging to organizations like SIGIR or CLEF.

One of the workshops of the Special Interest Group on Information Retrieval conference (SIGIR) was about Health Information Retrieval called MedIR [Med16c], this workshop aimed to bring together researchers interested in medical information research in order to identify obstacles that need to be addressed to achieve advances in the state of the art and to stimulate partnerships to address these challenges [GJK^{+}14, GKJ^{+}14].

Even being submitted to the same workshop some articles were focused on very different themes, such as Koopman and the Zuccon [KZ14b] article that focused on the question "Why Assessing Relevance in Medical IR is Demanding" and concluded that the assessment of relevance is, in some cases, related to the ambiguity of the queries made. Deng, Stoehr and Denecke did a study on the attitudes of users in Information Retrieval in order to assess the decisions of users based on an analysis of feelings [DSD14].

The initiative of the Conference and Labs of the Evaluation Forum (CLEF) has as its main task to promote research, innovation and the development of access to information systems with an emphasis on multilingual and multimodal information [CLE16a]. CLEF is divided into several laboratories, each is focused on one subproblem of Information Retrieval. Professionals and researchers from any sector can access all the information related to these laboratories.

The CLEF has a workshop dedicated to the area of medicine, CLEF eHealth. The purpose of the CLEF eHealth is to facilitate and support patients and their families in understanding and access relevant medical information [CLE16b].

Table 2.1 depicts the most relevant information from articles in the CLEF eHealth 2015 such as the base methodology, techniques for the query expansion like Pseudo Relevance Feedback (PRF) or Unified Medical Language System (UMLS), if they used machine learning techniques like Explicit Semantic Analysis (ESA) or, cluster (CBEEM) or concept-based models (CBDC) , and the choice of search engines. Of the submitted articles to the CLEF eHealth 2015 who achieved the best results in P@10 and nDCG@10 were the teams ECNU [SHH+15], KISTI [OJK15] and CUNI [SBP15], who occupied the first 10 places.

Table 2.1: Approaches used in CLEF eHealth 2015

Article	**Baseline**	**Query Expansion**	**Machine Learning**	**Search Engine**
[Lu15]	*BM25*	*PRF*	*No*	*Not defined*
[OJK15]	*Dirichlet Smoothing*	*PRF*	*ESA; CBDC; CBEEM*	*Lucene*
[GH15]	*BM25*	*PRF*	*No*	*Terrier*
[HNH15]	*Dirichlet Smoothing*	*UMLS; Wikipedia*	*No*	*Lucene*
[TAM15]	*BM25*	*PRF*	*Learning to Rank*	*Terrier*
[SHH+15]	*TF-IDF*	*Google; MeSH*	*Learning to Rank*	*Terrier*
[DGZ15]	*Bag-of-words*	*UMLS; Wikipedia*	*No*	*Lucene*
[KTBG15]	*VSM*	*PRF*	*No*	*Terrier*
[LN15]	*Dirichlet Smoothing*	*UMLS; Wikipedia*	*No*	*Indri*
[SBP15]	*Dirichlet Smoothing*	*UMLS; PRF*	*No*	*Terrier*

The ECNU team [SHH+15] explored query expansion and machine learning. For the query expansion they used the titles of the first ten results of Google to add to the original query. Then the medical terms were extracted with the aid of MeSH. For machine learning the ECNU team used the method Learning to Rank combining the results and ranks obtained from BM25, PL2 and BB2 weighting models.

Team KISTI [OJK15] used Lucene to index the collection and used the Dirichlet Smoothing as the weighting model. This group focused on the re-ranking of documents exploring three methodologies: explicit semantic analysis (ESA), concept-based document centrality (CBDC) and, based cluster external expansion model (CBEEM).

The team CUNI [SBP15] used Terrier to index the collection and three different weighting models: Bayesian smoothing with Dirichlet prior, Per-field normalization (PL2F) and Loss Given Default (LGD). The query expansion used the UMLS thesaurus as a source of terms obtaining synonyms of the terms used in the original query.

2.4 Readability in Health Information Retrieval

Although relevance is known to be a multidimensional concept, traditional retrieval measures only consider one dimension of relevance: topicality [Zuc16]. Topicality is a measure that can determine if a document is relevant for a given information need. However, if a user has difficulties reading this document he won't understand it, causing it to not provide relevant information, thus becoming irrelevant.

When determining the relevance of health-related documents, readability should be involved in its definition [ZK14]. The field of medicine is filled with complex and specific terms that lay people have difficulty in understanding. If only topicality is considered to measure the document relevance users might not be able to understand a vast number of them.

Low health literacy has consequences on patients taking medicines improperly, missing appointments, and failing to grasp expectations due to misunderstood or complex instructions [B04]. Health documents present additional difficulties because they usually employ medico-scientific terminology.

Wiener and Wiener-Pla [WWP14] have investigated the readability (measured by the SMOG reading index) of Web pages concerning pregnancy and the periodontium as retrieved by Google, Bing and Yahoo!. The research hypothesis was that web articles written below the 8th grade level do not provide adequate health information concerning periodontal changes, consequences, and control during pregnancy as compared with those written at or above the 8th grade level. They proved that articles written below the 8th grade level were more likely to recommend brushing twice a day and using a soft-bristled brush and, articles at or above the 8th grade level were more likely to discuss preterm birth and periodontal disease.

Walsh and Volsko [WV08] have shown that most online information sampled from five US consumer health organizations and related to the top 5 medical related causes of death in US are presented at a readability level (measured by the SMOG, FOG and Flesch-Kincaid reading indexes [MP82]) that exceeds that of the average US citizen (7th grade level). Their findings support that Web-based medical information intended for consumer use is written above the United States Department of Health and Human Services (USDHHS) recommended reading levels and that compliance with these recommendations may increase the likelihood of consumer comprehension.

Considering this, Zuccon and Koopman [ZK14] propose two understandability-based variants of rank biased precision, characterized by an estimation of understandability based on document readability and by different models of how readability influences user understanding of document content. Their findings suggest that considering understandability along with topicality in the evaluation of information retrieval systems leads to different claims about systems effectiveness than considering topicality alone.

2.4.1 Readability Metrics

Readability metrics measure the difficulty of understanding a passage of text. Some of these metrics are based on features such as number of syllables and the number of words in a sentence. These features ignore concept difficulty and are based on assumptions about writing style that may not hold in all environments.

There are many readability metrics. SMOG, FOG and Flesch-Kincaid are three of the most widely used readability metrics [MP82]. They all estimate the educational grade level necessary to understand a document [SC01].

SMOG Readability Formula estimates the years of education a person needs to understand a piece of writing. McLaughlin created this formula as an improvement over other readability formulas [Lau69] and defined it as:

$$SMOG = 3 + \sqrt{Number\ Of\ Polysyllable\ Words\ In\ 30\ Sentences} \tag{2.5}$$

If the document is longer than 30 sentences, the first 10 sentences, the middle 10 sentences, and the last 10 sentences are used. If the document has fewer than 30 sentences only the number of polysyllabic words are counted and conversion table is used to calculate the grade level (See Table 2.2). The SMOG measure tends to give higher values than other readability metric [Lau69].

Table 2.2: SMOG Conversion Table.

Polysyllabic Word Count	Readability Score
1 - 6	5
7 - 12	6
13 - 20	7
21 - 30	8
31 - 42	9
43 - 56	10
57 - 72	11
73 - 90	12
91 - 110	13
111 - 132	14
133 - 156	15
157 - 182	16
183 - 210	17
>210	18

The Gunning Fog Index Readability Formula, or simply called FOG Index, is attributed to American textbook publisher, Robert Gunning. Gunning observed that most high school graduates were unable to read. Much of this reading problem was a writing problem. His opinion was that newspapers and business documents were full of “fog” and unnecessary complexity. In 1952, Gunning created an easy-to-use Fog Index defined as:

$$FOG = 0.4\,(ASL + PHW) \tag{2.6}$$

where ASL is the Average Sentence Length (number of words divided by the number of sentences) and PHW is the Percentage of Hard Words (words of three or more syllables).

The Flesch-Kincaid readability metric was developed under contract to the U.S. Navy in 1975 by Rudolph Flesch and John Peter Kincaid. This Flesch-Kincaid formula was first used by the U.S. Army for assessing the difficulty of technical manuals in 1978 and soon after became the Department of Defense military standard. This formula is used to assess several legal documents in the U.S. The state of Pennsylvania was the first to use this to assess automobile insurance policies, that were required to be lower than a ninth-grade level of reading difficulty [Med16a]. The Flesch-Kincaid formula is defined as:

$$FK = (0.39 * ASL) + (11.8 * ASW) - 15.59 \tag{2.7}$$

where ASL is the Average Sentence Length (the number of words divided by the number of sentences) and the ASW is the Average number of Syllable per Word (the number of syllables divided by the number of words).

Chapter 3

Methods

In this chapter we present all the methods that will be used in the different approaches. Starting with the baseline which indicates the weighting model that will be used to rank the retrieved documents to a given query. Continuing with the test collections that will be a source of documents and queries to be used in the indexing and retrieval phase. We also identify the chosen retrieval system and the reasons for this decision. In the indexing stage one of the collections needed to be restructured to be processed by the retrieval system. For all the approaches we re-rank the runs based on the documents readability using three different formulas to combine relevance and readability. Finally, we indicate which evaluation measures are used to evaluate the efficiency of the system.

3.1 Baseline

The baseline is an measurement of the process functionality before any change occurs. The baseline in Information Retrieval is a weighing model that counts as a run and it is applied in every approach. This allows a comparison between the baseline and one of the approaches to verify if an improvement was accomplished.

We used BM25 term weighting model to score and rank documents according to their relevance to a given query. It is based on the probabilistic retrieval framework developed in the 1970s and 1980s by Stephen E. Robertson, Karen Sparck Jones, and others [RJ76].

For a given query Q, the relevance score of a document D based on the BM25 term weighting model is expressed as:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \frac{TF(q_i,D).(k1+1)}{TF(q_i,D)+k1.(1-b+b.\frac{|D|}{avgdl})} \tag{3.1}$$

where TF is the number of occurrences of a given term q_i in the document D. |D| the size of the document in words, *avgdl* the average size of a document. $k1$ an b are free parameters, usually chosen, in absence of an advanced optimization, as [MRS08]:

$$k1 \in [1.2; 2.0] \quad b = 0.75 \tag{3.2}$$

3.2 Test Collections

For the purpose of assessing the effectiveness of the approaches we used two distinct collections. These collections were provided by the CLEF eHealth Lab from 2015 and 2016.

3.2.1 CLEF eHealth 2015

3.2.1.1 Documents

The CLEF eHealth 2015 collection is provided by the Khresmoi project [HM12] which obtained about one million documents through a web crawler. Web pages certified by the HON Foundation and adhering to the HONcode principles were the primary source for the crawled domains, as well as other commonly used health and medicine sites such as Drugbank, Diagnosia and Trip Answers [PZG+15]. These web pages have a broad range of health topics and are likely to target both laypeople and professionals. This collection as a size of 6.3GB when compressed and approximately 50GB when extracted.

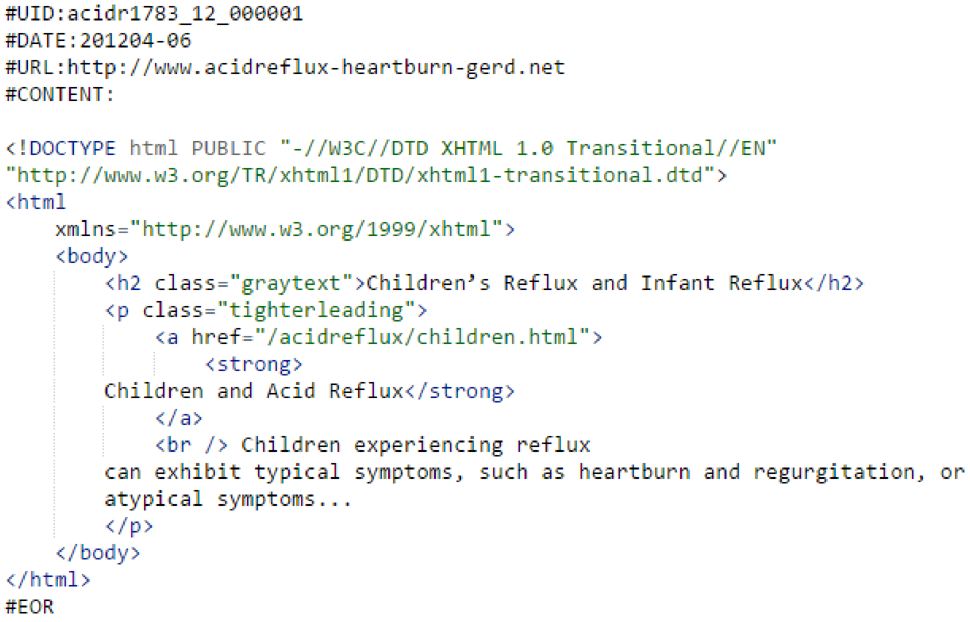
The documents in the collection were stored in .dat files with the following format (see Figure 3.1):

- #UID: Unique identifier for a document in the collection;
- #DATE: Date the document was obtained;
- #URL: URL for the source of the document;
- #CONTENT: Document Content.

3.2.1.2 Queries

To build the CLEF eHealth 2015 queries several volunteers were asked to generate queries after reviewing images and videos related to medical symptoms [PZG+15] (Figure 3.2).

This process tried to simulate a situation of when a health consumer has an information need regarding symptoms or conditions they may be affected by. This methodology for eliciting self-diagnosis queries was shown to be effective by Stanton [SIM14]. Each volunteer gave 3 queries for each condition they saw, generating a total of 266 queries. Then, for each condition, the CLEF organization randomly selected 3 queries, giving the 67 queries that will be used (Figure 3.3).



```
#UID:acidr1783_12_000001
#DATE:201204-06
#URL:http://www.acidreflux-heartburn-gerd.net
#CONTENT:

<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html
    xmlns="http://www.w3.org/1999/xhtml">
    <body>
        <h2 class="graytext">Children’s Reflux and Infant Reflux</h2>
        <p class="tighterleading">
            <a href="/acidreflux/children.html">
                <strong>
        Children and Acid Reflux</strong>
            </a>
            <br /> Children experiencing reflux
        can exhibit typical symptoms, such as heartburn and regurgitation, or
        atypical symptoms...
        </p>
    </body>
</html>
#EOR
```

Figure 3.1: Example of a Document from CLEF eHealth 2015

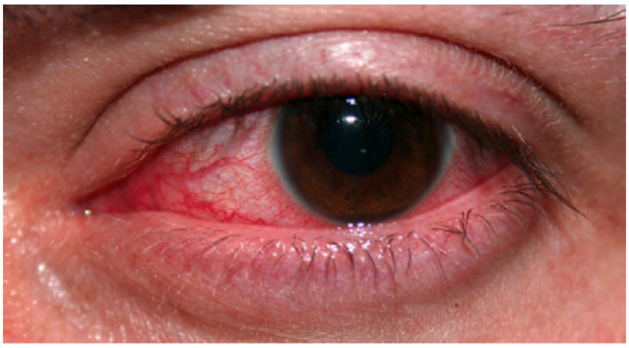
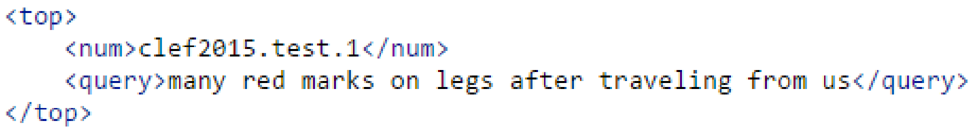


Figure 3.2: Example of an image provided to volunteers for generating potential search queries [PZG+15].



```
<top>
    <num>clef2015.test.1</num>
    <query>many red marks on legs after traveling from us</query>
</top>
```

Figure 3.3: Query Examples from CLEF eHealth 2015.

3.2.1.3 Relevance and Readability Assessments

Five medical students from Medizinische Universitat Graz (Austria) were employed to perform the relevance and readability assessments of the documents [PZG+15] using Relevation! [KZ14a]. To give a correct relevance assessment the students had access to the query the document was retrieved for and the target symptom or condition that generate that query [PZG+15]. Relevance assessments were provided on a three point scale: 0, “Not Relevant”; 1, “Somewhat Relevant”; 2, “Highly Relevant” [PZG+15].

To evaluate the documents readability each assessor was asked if he believed a patient would understand it [PZG+15]. Assessments were provided on a four point scale: 0, “It is very technical and difficult to read and understand”; 1, “It is somewhat technical and difficult to read and understand”; 2, “It is somewhat easy to read and understand”; 3, “It is very easy to read and understand” [PZG+15].

3.2.2 CLEF eHealth 2016

3.2.2.1 Documents

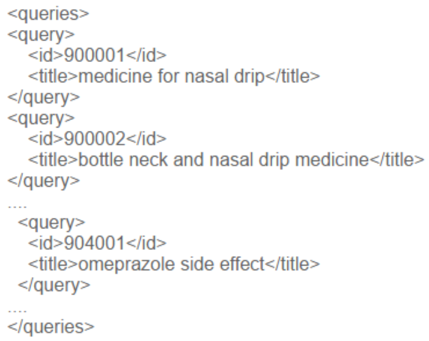
The collection for CLEF eHealth 2016 is the ClueWeb12 B13 Dataset [Pro16]. This collection was generated by taking the 10 million ClueWeb09 URLs that had the highest PageRank scores, and then removing any page that was not in the top 90% of pages least likely to be spam, according to the Waterloo spam scores. These URLs were used as a starting point for a crawl which excluded any page that appeared in the blacklist provided by a URL blacklist service [URL16]. As Table 3.1 shows, the size of the document collection increased tremendously from 2015 to 2016.

Table 3.1: Comparison between datasets from 2015 and 2016.

	Size	Number of documents
CLEF eHealth 2015	43.6GB	1,583,273
CLEF eHealth 2016	1,950GB	52,343,021

3.2.2.2 Queries

Queries from CLEF eHealth 2016 explore real health consumer posts from health web forums [CLE16b]. They were extracted from posts on the askDocs forum of Reddit, and presented to query generators. Query generators had to create queries based on what they read in the initial user post, making several variations for the same condition. These queries were assigned a unique six digit id, in which the first three numbers represent the post the original post, and the last three identify the variation of the query as showed on Figure 3.4. The Linux program aspell was used to correct some misspellings on the English queries.



```
<queries>
<query>
   <id>900001</id>
   <title>medicine for nasal drip</title>
</query>
<query>
   <id>900002</id>
   <title>bottle neck and nasal drip medicine</title>
</query>
....
  <query>
   <id>904001</id>
   <title>omeprazole side effect</title>
  </query>
....
</queries>
```

Figure 3.4: Query Examples from CLEF eHealth 2016.

3.2.2.3 Relevance and Readability Assessments

At the time of writing this document the relevance and readability assessments for this test collection are not available.

3.3 Retrieval System Selection

The choice of the search engine was made taking into account its popularity in research studies in the IR area and taking into account the assessment made by Christian Middleton and Ricardo Baeza-Yates [MBY07]. The most popular search engines in Information Retrieval researches are Lucene [Luc16] and Terrier [Ter16] .

The initial evaluation of Middleton and Baeza-Yates undergoes checking the last time that the search engines was updated [MBY07]. Lucene and Terrier have both been recently updated so the choice will depend on their performance.

As a first performance test Middleton and Baeza-Yates used 3 collections with 750MB, 1.6GB and 2.7GB to determine the indexing time by the search engines [MBY07]. In this test, Terrier had better results with the three collections, approximately 30% lower than Lucene (See Figure 3.5).

The size of the index created by the search engines was another point of evaluation by Middleton and Baeza-Yates [MBY07]. In this case it was Lucene that got the best results with an average of 25% of the size of the collection, while the Terrier got an average of 50% (See Table 3.2).

Table 3.2: Comparison of Lucene and Terrier index sizes [MBY07].

	700MB	1.6GB	2.7GB
Lucene	25%	23%	26%
Terrier	51%	47%	52%

During the evaluation of the index sizes it was found that several search engines, including Lucene, had a scaling problem. While Terrier had an expected indexing time when indexing collections with 10GB, Lucene took more than 7 times its expected indexing time [MBY07] (See Figure 3.6).

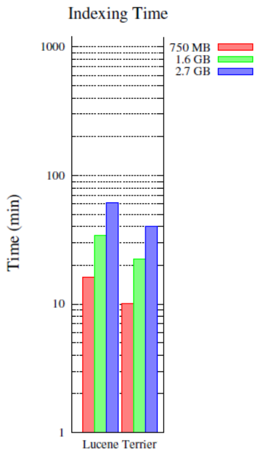


Figure 3.5: Lucene and Terrier indexing time on three different collections [MBY07].

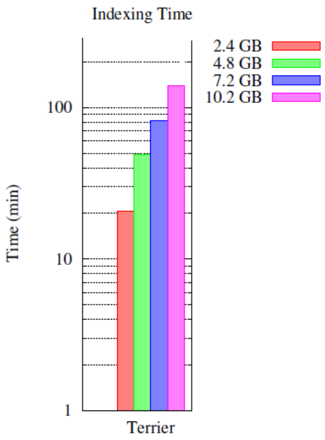
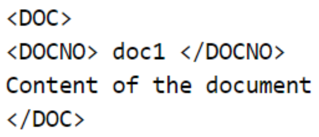


Figure 3.6: Terrier indexing time on larger collections [MBY07].

Knowing that the 2015 and 2016 document collections have approximately 50GB and 2000GB, respectively, Lucene ceases to be a viable candidate leaving Terrier as the best option. Furthermore, for the 2016 collection the organization of the CLEF eHealth Lab chose only to use Terrier and Indri.

3.4 Indexing Process

After choosing a search engine, it is necessary to index the document collection. Terrier doesn't have a parser for the document format in the CLEF collection so we decided to convert the documents to the TREC format (Figure 3.7) which is the default parser of Terrier. This was only made in the 2015 document collection because, in the 2016 collection the organization of the CLEF eHealth Lab provided an index on Terrier and Indri to all the participants through a virtual machine on the Azure platform.



```
<DOC>
<DOCNO> doc1 </DOCNO>
Content of the document
</DOC>
```

Figure 3.7: TREC document format.

Knowing that the documents in the collection are web pages, we assumed that extracting the text from the pages would improve indexing performance, because it would remove a great quantity of irrelevant content, like, HTML tags and scripts. To do this we used the Jsoup library. Jsoup is a Java library for working with HTML. It provides a very convenient API for extracting and manipulating data, using DOM, CSS, and jquery-like methods [Hed16].

After extracting the text from the web pages, the collection went from 50GB to 8GB and had an indexing time of around 30 minutes.

3.5 Re-Ranking

A re-rank method was developed to combine the readability metrics (SMOG, FOG and Flesch-Kincaid) with the relevance scores from Terrier. Three different formulas were used to combine these values. In formula 3.4 *MR* is the maximum readability to be considered, i.e., any document that has a readability score higher than *MR* will be considered as too hard for anyone to understand. The last formula was proposed by Zuccon and Koopman [ZK14], where a user is characterized by a readability threshold (*th*) and every document that has a readability score below *th* is considered readable, while documents with readability above *th* are considered unreadable. Figure 3.8 shows a graphic representation of the Formulas 3.3, 3.4 and 3.5.

$$Score = Relevance/Readability \tag{3.3}$$

$$Score = Relevance * log(\frac{MR}{Readability}) \tag{3.4}$$

$$Score = Relevance * (\frac{1}{2} - \frac{\arctan(Readability - th)}{\pi}) \tag{3.5}$$

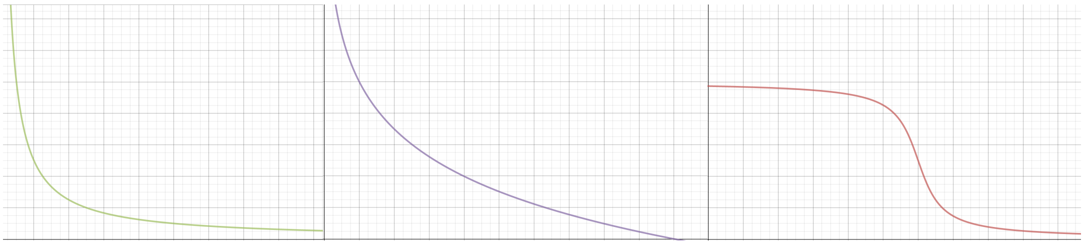


Figure 3.8: Graphic representation of the formulas used to combine relevance and readability.

In this method we re-rank each run with one of the readability metrics and one of the above combination formulas (3.3, 3.4 and 3.5), which generates a total of 9 different variants for each run.

3.5.1 Document Analysis

The CLEF eHealth 2015 and 2016 collections were analysed regarding their readability. For the 2015 collection, the analysis was made for all the documents. However, for the 2016 collection, due to its size and files compression, it was only possible to analyse a fraction of its documents within an reasonable time, that is, the first ten of each query for every approach.

Table 3.3: Number of documents analysed on the datasets from 2015 and 2016.

	Number of documents analysed
CLEF eHealth 2015	1,583,273
CLEF eHealth 2016	10,772

The documents from both collections are web pages, so we used the Jsoup library to extract the text from those pages to calculate the readability scores.

Because the text is extracted from the page without considering its location, it will, in most cases, decrease the number of sentences which, consequently, increases the readability score, e.g. a page with tables. The content within the tables is not punctuated so when it's extracted by the Jsoup it will be appended to another sentence increasing its size and number of polysyllabic words.

The used readability metrics consider that a document is easily understood when it has a grade between 7th and 9th and anything above 12th grade is too hard for most people to read [Lau69, Med16a].

In the analysis, we considered readability scores between 5 and 20 because, after reviewing the results, we found that these values allowed a better understanding of the documents distribution.

As shown in the Figures 3.9 and 3.10 most of the documents have a readability score higher than 12th grade. This demonstrates that a user will have difficulties when searching for health related information. The readability analysis of the 2016 collection is a good way to show this difficulty because it was only done to the first ten documents, i.e. the first documents that a user will read.

With this analysis we defined MR in formula 3.4 as 20 and the threshold (th) in formula 3.5 as 12.

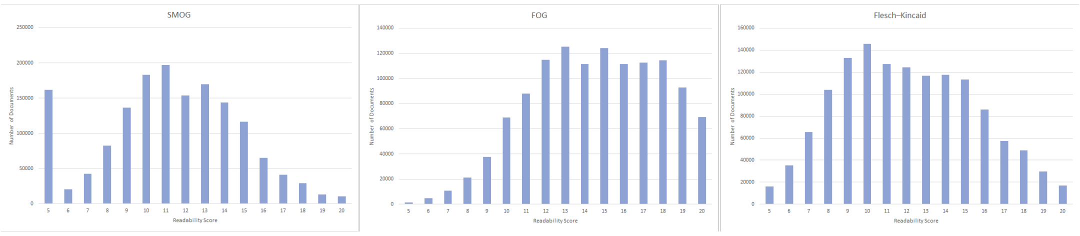


Figure 3.9: Readability Analysis for the 2015 collection

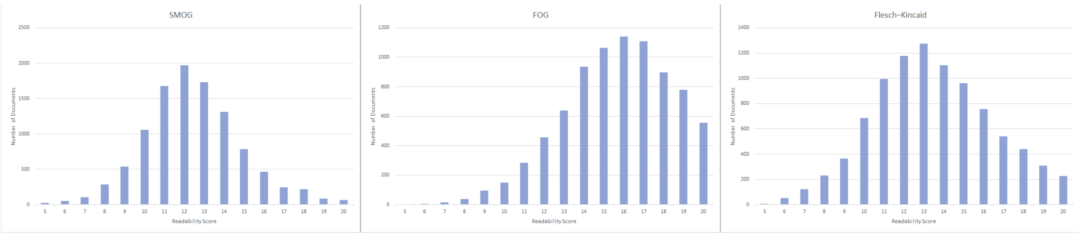


Figure 3.10: Readability Analysis for the 2016 collection

3.6 Evaluation

System evaluation was conducted using two relevance measures: (1) precision at 10 ($P@10$) and (2) normalized discounted cumulative gain at 10 ($nDCG@10$). Precision was computed using binary relevance assessments (relevant or not) and nDCG was computed using the graded assessments. These evaluation metrics were computed using trec eval [TRE16a].

Following the methods of Zuccon and Koopman article [ZK14] an evaluation using relevance and readability assessments was made. This evaluation was made using a readability-biased modification of the Rank Biased Precision (RBP) formula, $uRBP$ and its graded version $uRBPgr$.

RBP is designed with the idea that a user will start at the top of the retrieved document list and he will proceed to the next document with a probability p, or finish the search with probability $1-p$ [MZ08]. When the user reviews a relevant document the RBP score increases, therefore RBP is computed as the sum of the probability of examining each relevant document:

$$RBP(p) = (1-p)\sum_{i=1}^{\infty} r_i * p^{i-1} \tag{3.6}$$

where $r_i \in [0, 1]$ is the relevance judgement of the *i*th ranked document, and the $(1 - p)$ factor is used to scale the RBP within the range [0, 1]. The probability of a user examining the next document reflects the persistence of the user [MZ08].

The RBP parameter p (RBP persistence parameter) was set to 0.8 for all variations of this measure, following the findings of Park and Zhang [PZ07].

The readability-biased evaluation were performed using the ubire tool [Lab16].

Chapter 4

Approaches

In this section we present several approaches to identify which terms will be added to the query.

Starting with Pseudo Relevance Feedback that uses the documents in the test collection as a source of terms. Continuing with the Medical Text Indexer (MTI) that identifies medical concepts on the original query and appends them to it. Wikipedia is used in several approaches using the articles contents or the titles of similar articles as a source of terms for the query expansion. Wikipedia articles are also used to find references to medical web pages like MedlinePlus and ICD-10. Finally the UMLS Metathesaurus definitions of the MTI concepts are used as a source of terms.

It is important to mention that every approach excludes stop words.

Figure 4.1 shows a graphic representation of all the approaches with each path being a different approach.

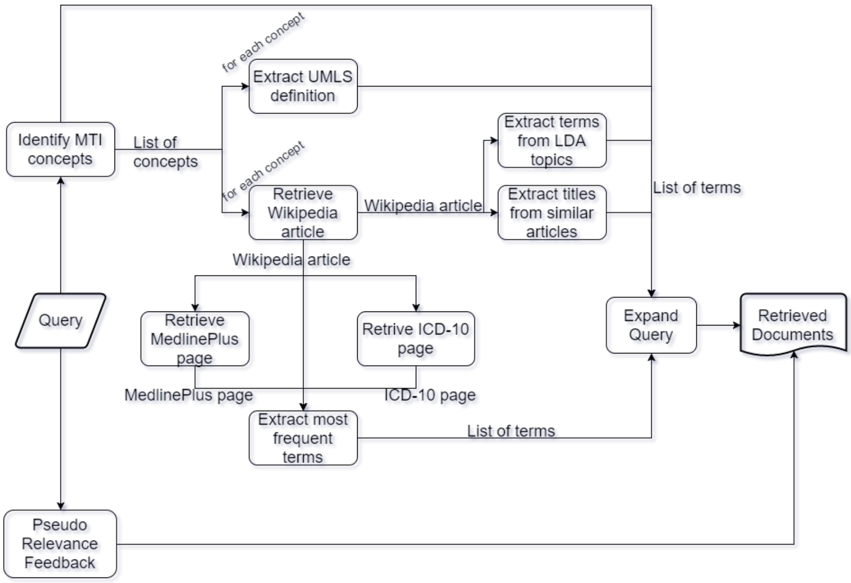


Figure 4.1: Query expansion approaches.

After an initial evaluation of the results it was discovered that most of the top documents in the different runs didn't have a relevance judgement. This means the documents would be considered non-relevant. This wouldn't be an issue if the documents were indeed non-relevant but, after manually opening and reading those documents, we confirmed that they were relevant and, for this reason, they needed to be considered as such in the relevance judgements. Unfortunately evaluating each document for all the queries would take too long, not to mention that we didn't had the proper knowledge to make a correct evaluation. So it was decided that, in the retrieval phase that uses the 2015 document collection, only documents with an relevance judgement would be valid. For the 2016 collection, because the index was provided in a virtual machine, we couldn't make any changes to the set of documents used in the retrieval phase.

In this section each approach will be followed by the results obtained in the evaluation phase. These results will be shown in two tables: one with the results from the query expansion alone and the other of the query expansion and readability re-rank. Because each query expansion run has nine different re-rank runs we decided to aggregate all the re-rank results into one table. To do this we chose only two evaluation measures to evaluate relevance and readablity: $P@10$ and $uRBP$. The complete tables for the re-rank runs are displayed in the Appendix A.

For simplification purposes the Formulas 3.3, 3.4 and 3.5 will be called Basic, Log and Arctan formulas.

4.1 Pseudo Relevance Feedback

Pseudo relevance feedback is a method of query expansion that uses the document collection in which it runs as the source for its terms [Eft96]. In this method the top documents returned by the baseline are used to modify the query by re-weighting the existing query terms and by adding terms that appear useful and by deleting terms that do not [Eft96].

Terrier provides two different models to apply the pseudo relevance feedback method: the Bose-Einstein and the Kullback-Leibler Divergence. The Bose-Einstein model calculates the weight of terms, as following [AVR02]:

$$w(t) = tf_x.log_2(\frac{1+P_n(t)}{P_n(t)}) + log_2(1+P_n(t)) \tag{4.1}$$

$$P_n(t) = \frac{tf_c}{N} \tag{4.2}$$

where tf_x is the frequency of the query term t in the top-ranked documents, tf_c is the frequency of term t in the collection, and N is the number of documents in the collection [Lu15]. The Kullback-Leibler Divergence computes the divergence between the probability distribution of terms in the whole collection and in the top ranked documents obtained using the original query [IS10]. The most likely terms to be appended are those in the top ranked documents with a low document frequency. For the term t this score is:

$$KLD(t) = [P_r(t) - P_c(t)].log \frac{\frac{f(t)}{NR}}{P_c(t)} \tag{4.3}$$

where $P_r(t)$ is the probability of t estimated from the top retrieved documents relative to a query (R). $P_c(t)$ is the probability of t estimated using the whole collection [Lu15].

For this approach two runs were created to identify which one of these models provides better results. We used the Terrier default values for the top-ranked documents (3) and terms (10) used for the query expansion.

4.1.1 Results

The Pseudo Relevance Feedback approach shows that even with an automatic query expansion process it is possible to improve over a simple retrieval. The results of both its runs (Table 4.1) do not differentiate by a significant amount suggesting that either one can be considered when applying an automatic query expansion process to a system. None of the re-rank runs (Table 4.2) outperformed the results of the baseline in both relevance and readability.

Table 4.1: Pseudo Relevance Feedback results for CLEF eHealth 2015 collection.

Run	P@10	nDCG@10	uRBP	uRBPgr
Baseline	0.3455	0.3027	0.3148	0.3033
Bose-Einstein	0.3545	0.3008	**0.3212**	**0.3110**
Kullback-Leibler	**0.3576**	**0.3021**	0.3202	0.3100

Table 4.2: Pseudo Relevance Feedback Re-Rank results for CLEF eHealth 2015 collection.

		SMOG		FOG		Flesch-Kincaid	
	Run	P@10	uRBP	P@10	uRBP	P@10	uRBP
	Baseline	0.3455	0.3148	0.3455	0.3148	0.3455	0.3148
Basic Formula (3.3)	Bose-Einstein	**0.3197**	**0.2779**	**0.3167**	**0.2809**	**0.3076**	**0.2704**
	Kullback-Leibler	0.3167	0.2712	0.3121	0.2800	0.3030	0.2647
Log Formula (3.4)	Bose-Einstein	**0.3045**	**0.2560**	**0.2970**	**0.2532**	**0.2939**	0.2638
	Kullback-Leibler	0.3000	0.2541	0.2864	0.2527	0.2894	**0.2648**
Arctan Formula (3.5)	Bose-Einstein	0.3227	0.2817	**0.2848**	0.2565	**0.3455**	0.3128
	Kullback-Leibler	**0.3318**	**0.2853**	0.2773	**0.2598**	0.3439	**0.3134**

4.2 Query expansion using the Medical Text Indexer

The National Library of Medicine (NLM) Medical Text Indexer (MTI) combines human expertise and Natural Language Processing technology to curate the biomedical literature more efficiently and consistently. Since 2002, MTI has been the main product of the Indexing Initiative project providing indexing recommendations based on the Medical Subject Headings (MeSH) [JGM13]. Every week MTI recommends approximately 4,000 new citations for indexing and processes a file of approximately 7,000 old and new records for both Cataloging and the History of Medicine Division (HMD) [JGM13]. Between 2002 and 2012, MTI was used to provide fully-automated indexing for NLM's Gateway abstract collection, which was not manually indexed [JGM13]. The designation of First-Line Indexer (MTIFL) was given to MTI in 2011 because of its success with several publications [JGM13].

Queries were processed by MTI which linked the text from the query to the MeSH vocabulary resulting in additional related concepts. The identified concepts are likely to be important for the retrieval process. However the MTI results are machine generated what, depending on the query, could result in irrelevant concepts.

In this approach, we appended all the concepts identified by the MTI to the original query.

4.2.1 Results

The results for the MTI approach (Table 4.3) achieved a significant improvement in relevance and readability compared with the baseline. Identifying medical concepts related to the query proved to be one of the best ways to improve the system performance. On the re-ranks runs (Table 4.4), few of them improved over the baseline. Even so, none of them got results higher than the ones using only the query expansion.

Table 4.3: MTI results for CLEF eHealth 2015 collection.

Run	P@10	nDCG@10	uRBP	uRBPgr
Baseline	0.3455	0.3027	0.3148	0.3033
MTI	**0.4061**	**0.3530**	**0.3381**	**0.3276**

Table 4.4: MTI Re-Rank results for CLEF eHealth 2015 collection.

		SMOG		FOG		Flesch-Kincaid	
	Run	P@10	uRBP	P@10	uRBP	P@10	uRBP
	Baseline	0.3455	0.3148	0.3455	0.3148	0.3455	0.3148
Basic Formula (3.3)	MTI	**0.3470**	**0.3087**	**0.3227**	**0.2953**	**0.3167**	**0.2768**
Log Formula (3.4)	MTI	**0.3227**	**0.2889**	**0.2985**	**0.2684**	**0.3136**	**0.2814**
Arctan Formula (3.5)	MTI	**0.3379**	**0.3050**	**0.2909**	**0.2697**	**0.3515**	**0.3288**

4.3 Query expansion using the Wikipedia

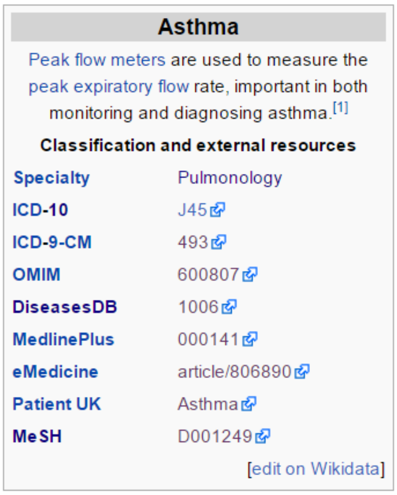
Wikipedia is a free encyclopedia, written collaboratively by the people who use it. Many people are constantly improving Wikipedia, making thousands of changes per hour [Wik16b]. This makes Wikipedia an enormous source of information likely to contain medical terms in lay language. As shown in the work of Laurent and Vickers [LV09], the English Wikipedia is a prominent source of online health information when compared to other online health information providers like MedlinePlus.

Using the Wikipedia as a base, we defined two methods to get terms for the query expansion process. One of the methods extracts the most frequent terms from Wikipedia articles. The other uses Wikipedia as a directed graph to identify similar articles and then extracts terms from the titles of these articles.

4.3.1 Term Frequency

The MediaWiki action API is a web service that provides a convenient access to wiki features, data, and meta-data over HTTP, via a URL [Med16b]. This API was used to find the articles that best match the concepts obtained through the MTI.

An analysis of the obtained Wikipedia pages allowed us to identify some that were not health-related. To minimize this, we decided to exclude the pages not containing an infobox similar to the one presented in Figure 4.2 [Wik16a] which contains information about the category of the page (e.g. anatomy, disease, drug).



Asthma

Peak flow meters are used to measure the peak expiratory flow rate, important in both monitoring and diagnosing asthma.[1]

Classification and external resources

Specialty	Pulmonology
ICD-10	J45
ICD-9-CM	493
OMIM	600807
DiseasesDB	1006
MedlinePlus	000141
eMedicine	article/806890
Patient UK	Asthma
MeSH	D001249

[edit on Wikidata]

Figure 4.2: Wikipedia Asthma Infobox.

This approach had several variants. We chose the 5, 10 and 15 most frequent terms of each article. In addition we considered (1) all articles found with the MTI concepts and (2) only the articles considered health-related using the strategy defined above.

4.3.1.1 Results

In the Wikipedia Term Frequency approach (Table 4.5) increasing the number of the most frequent terms used didn't improve its efficiency. However, when only using health-related Wikipedia articles the scores of relevance and readability improved over the runs where all articles were considered. Most of the re-rank runs (Table 4.6) that improved the relevance scores were the ones that only used health-related articles. Only the combination of the Arctan formula and the Flesch-Kincaid metric brought readability improvements over the baseline.

Table 4.5: Wikipedia Term-Frequency results for CLEF eHealth 2015 collection.

Run	P@10	nDCG@10	uRBP	uRBPgr
Baseline	0.3455	0.3027	0.3148	0.3033
Wiki TF 5	0.3636	0.3165	0.2761	0.2823
Wiki TF 10	0.3515	0.3142	0.2614	0.2726
Wiki TF 15	0.3545	0.3068	0.2410	0.2606
Wiki TF 5 Health	0.3864	0.3388	**0.3189**	**0.3190**
Wiki TF 10 Health	**0.3894**	**0.3455**	0.3077	0.3122
Wiki TF 15 Health	0.3848	0.3370	0.2966	0.3061

Table 4.6: Wikipedia Term-Frequency Re-Rank results for CLEF eHealth 2015 collection.

		SMOG		FOG		Flesch-Kincaid	
	Run	P@10	uRBP	P@10	uRBP	P@10	uRBP
	Baseline	0.3455	0.3148	0.3455	0.3148	0.3455	0.3148
Basic Formula (3.3)	Wiki TF 5	0.3379	0.2715	0.3136	0.2709	0.3061	0.2548
	Wiki TF 10	0.3242	0.2596	0.3106	0.2583	0.3136	0.2505
	Wiki TF 15	0.3364	0.2585	0.3333	0.2551	0.3182	0.2487
	Wiki TF 5 Health	0.3591	0.3048	0.3500	**0.3055**	0.3152	0.2790
	Wiki TF 10 Health	0.3652	**0.3077**	0.3576	0.3019	**0.3394**	**0.2861**
	Wiki TF 15 Health	**0.3667**	0.2957	**0.3591**	0.2942	0.3364	0.2770
Log Formula (3.4)	Wiki TF 5	0.3227	0.2646	0.3106	0.2619	0.3091	0.2660
	Wiki TF 10	0.3227	0.2666	0.3182	0.2690	0.3152	0.2650
	Wiki TF 15	0.3364	0.2693	**0.3318**	0.2710	0.3182	0.2599
	Wiki TF 5 Health	0.3455	0.2971	0.3152	0.2767	0.3167	0.2849
	Wiki TF 10 Health	**0.3515**	**0.3024**	0.3197	**0.2828**	0.3167	**0.2866**
	Wiki TF 15 Health	0.3470	0.2952	0.3227	0.2772	**0.3212**	0.2805
Arctan Formula (3.5)	Wiki TF 5	0.3455	0.2855	0.2864	0.2630	0.3379	0.3024
	Wiki TF 10	0.3455	0.2713	0.3000	0.2572	0.3455	0.2982
	Wiki TF 15	**0.3530**	0.2763	**0.3045**	0.2657	0.3515	0.2941
	Wiki TF 5 Health	**0.3530**	0.2965	0.2970	**0.2724**	**0.3621**	**0.3293**
	Wiki TF 10 Health	0.3515	**0.2970**	0.2970	0.2715	0.3576	0.3239
	Wiki TF 15 Health	0.3485	0.2935	0.3000	0.2692	0.3515	0.3176

4.3.2 Link Analysis

As shown in the work of Almasari [MA], Wikipedia is a hypertext network in which each article can refer to other Wikipedia article using hyperlinks. Considering only internal links, which are links that target an other Wikipedia article it is possible to represent Wikipedia articles as a directed graph $G(A;L)$ of articles A connected by links L. L is the set of all the Incoming and Outgoing Links from the article A.

Each concept from MTI was used to search for an Wikipedia article which served as a starting point. Using the Wikipedia directed graph it is possible to retrieve the articles which referred and are referred by the first article. This method returns thousands of articles that aren't relevant to the expansion process because even if they're referred by the first article they might not be in the same category. To solve this issue we used the Jaccard similarity coefficient. This coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets [Ley08]. The Incoming (I) and Outgoing (O) Links from two articles were used as the sets for the Jaccard coefficient.

$$J(I,O) = \frac{|I \bigcap O|}{|I \bigcup O|} \tag{4.4}$$

$$I = I_{article1} \bigcup I_{article2} \tag{4.5}$$

$$O = O_{article1} \bigcup O_{article2} \tag{4.6}$$

$I_{article1}$, $I_{article2}$, $O_{article1}$ and $O_{article2}$ are the Incoming and Outgoing Links from both the starting article and the article being compared (See Figure 4.3). This allows the Jaccard coefficient to calculate the similarity between these two articles.

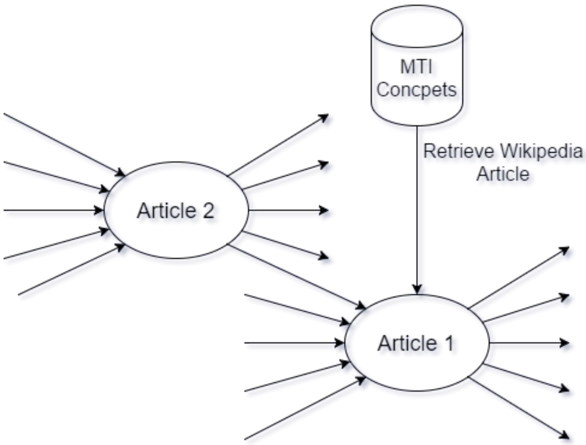


Figure 4.3: Example for the Link Analysis Approach.

We used the Java Wikipedia Library (JWPL) to compute this coefficient. The JWPL is a free, Java-based application programming interface that allows access to all the information in Wikipedia [ZMG08, DKP16]. This library uses a Wikipedia dump from March 2016.

In this approach we added to the original query all the titles of articles that had a Jaccard similarity coefficient greater than 0.25, 0.50 and 0.75. We considered one alternative using all articles found with the MTI concepts and another using only the articles that were considered health-related using the strategy defined above.

4.3.2.1 Results

In the Wikipedia Link Analysis approach (Table 4.7) the higher the Jaccard similarity coefficient the lower the relevance scores are. This approach is similar to the Term Frequency one, where using only health-related Wikipedia articles improved the relevance and readability scores. Most of the re-rank runs (Table 4.8) that improved the relevance scores were the ones that only used health-related articles. Only the combination of the Arctan formula and the Flesch-Kincaid metric brought readability improvements over the baseline.

Table 4.7: Wikipedia Link Analysis results for CLEF eHealth 2015 collection.

Run	P@10	nDCG@10	uRBP	uRBPgr
Baseline	0.3455	0.3027	0.3148	0.3033
Wiki Link 0.25	0.3788	0.3298	0.2743	0.2820
Wiki Link 0.50	0.3621	0.3138	0.2711	0.2761
Wiki Link 0.75	0.3500	0.3049	0.2768	0.2772
Wiki Link 0.25 Health	**0.3848**	**0.3466**	0.3061	0.3044
Wiki Link 0.50 Health	0.3788	0.3382	**0.3128**	**0.3093**
Wiki Link 0.75 Health	0.3727	0.3196	**0.3128**	0.3046

Table 4.8: Wikipedia Link Analysis Re-Rank results for CLEF eHealth 2015 collection.

		SMOG		FOG		Flesch-Kincaid	
	Run	P@10	uRBP	P@10	uRBP	P@10	uRBP
	Baseline	0.3455	0.3148	0.3455	0.3148	0.3455	0.3148
Basic Formula (3.3)	Wiki Link 0.25	0.3348	0.2575	0.3273	0.2594	0.3121	0.2460
	Wiki Link 0.50	0.3076	0.2545	0.3030	0.2534	0.2894	0.2417
	Wiki Link 0.75	0.3076	0.2500	0.3045	0.2541	0.2939	0.2402
	Wiki Link 0.25 Health	**0.3652**	0.2903	0.3530	0.3041	**0.3409**	0.2916
	Wiki Link 0.50 Health	0.3636	**0.2985**	**0.3621**	**0.3104**	0.3348	**0.2923**
	Wiki Link 0.75 Health	0.3500	0.2915	0.3515	0.3070	0.3227	0.2822
Log Formula (3.4)	Wiki Link 0.25	0.3242	0.2628	0.3061	0.2565	0.3152	0.2584
	Wiki Link 0.50	0.3076	0.2560	0.2909	0.2507	0.2909	0.2526
	Wiki Link 0.75	0.3076	0.2533	0.3000	0.2518	0.3061	0.2519
	Wiki Link 0.25 Health	0.3455	0.2863	0.3227	0.2848	**0.3394**	0.2943
	Wiki Link 0.50 Health	**0.3485**	**0.2961**	0.3273	**0.2860**	0.3364	**0.2958**
	Wiki Link 0.75 Health	0.3364	0.2820	**0.3303**	0.2833	0.3242	0.2885
Arctan Formula (3.5)	Wiki Link 0.25	0.3333	0.2820	0.2939	0.2632	**0.3621**	0.3116
	Wiki Link 0.50	0.3182	0.2773	0.2818	0.2532	0.3333	0.3048
	Wiki Link 0.75	0.3136	0.2675	0.2864	0.2442	0.3288	0.2967
	Wiki Link 0.25 Health	0.3333	0.2811	0.3015	0.2771	0.3591	**0.3314**
	Wiki Link 0.50 Health	**0.3455**	**0.2984**	**0.3091**	**0.2800**	0.3576	0.3289
	Wiki Link 0.75 Health	0.3409	0.2904	**0.3091**	0.2690	0.3545	0.3261

4.4 Query expansion using MedlinePlus

MedlinePlus is the National Institutes of Health Web site for patients, their families and friends. Produced by the National Library of Medicine, the world's largest medical library, it brings information about diseases, conditions, and wellness issues in lay language [Med16d].

Using the information on the infobox (Figure 4.2), obtained through the search of the MTI concepts on the Wikipedia, it is possible to access the corresponding MedlinePlus page. MedlinePlus pages are generally splitted in different sections with relevant information about the searched concept. The sections that were considered most relevant for the query expansion process were the Causes, Symptoms, Treatment, Possible Complications and Alternative Names sections.

We tested several variants of this method. We chose the top 5, 10 and 15 most frequent terms on each of the above mentioned sections including all the terms in the Alternative Names section. We also made a run with only the terms from the Alternative Names section.

4.4.1 Results

Using the different sections in a MedlinePlus page as a source of terms proved to be effective in improving the relevance scores (Table 4.9). However, using only the terms from the Alternative Names section didn't improve relevance as much, but it had the best readability score. Most of the re-rank runs (Table 4.10) didn't improve over the baseline. The best relevance score was obtained through the Basic formula using the SMOG metric, and the best readability score was obtained through the Arctan formula using the Flesch-Kincaid metric.

Table 4.9: MedlinePlus results for CLEF eHealth 2015 collection.

Run	P@10	nDCG@10	uRBP	uRBPgr
Baseline	0.3455	0.3027	0.3148	0.3033
Medline AltNames	0.3621	0.3138	**0.3180**	0.3124
Medline TF 5	0.3879	**0.3450**	0.3172	**0.3169**
Medline TF 10	**0.3894**	0.3420	0.3125	0.3147
Medline TF 15	0.3879	0.3437	0.3099	0.3143

Table 4.10: MedlinePlus Re-Rank results for CLEF eHealth 2015 collection.

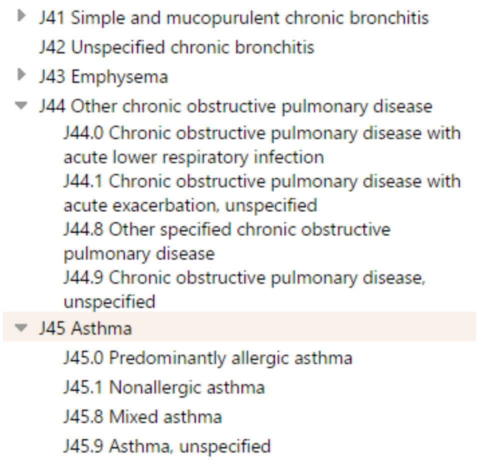
		SMOG		FOG		Flesch-Kincaid	
	Run	P@10	uRBP	P@10	uRBP	P@10	uRBP
	Baseline	0.3455	0.3148	0.3455	0.3148	0.3455	0.3148
Basic Formula (3.3)	Medline AltNames	0.3364	0.2912	0.3288	0.2891	0.3045	0.2663
	Medline TF 5	0.3500	0.3014	0.3439	**0.3019**	0.3091	**0.2709**
	Medline TF 10	**0.3621**	**0.3051**	**0.3455**	0.2944	**0.3121**	0.2662
	Medline TF 15	0.3515	0.2950	0.3379	0.2902	0.3106	0.2629
Log Formula (3.4)	Medline AltNames	0.3242	0.2785	0.3030	0.2692	**0.3106**	0.2744
	Medline TF 5	0.3318	0.2851	0.3182	0.2772	0.3076	**0.2801**
	Medline TF 10	**0.3364**	**0.2908**	**0.3197**	**0.2803**	0.3091	0.2755
	Medline TF 15	0.3318	0.2886	0.3091	0.2693	0.3076	0.2719
Arctan Formula (3.5)	Medline AltNames	0.3379	0.2931	0.2985	0.2739	0.3455	0.3165
	Medline TF 5	0.3409	0.2941	0.3030	0.2769	0.3545	0.3253
	Medline TF 10	**0.3439**	**0.2999**	**0.3106**	**0.2833**	**0.3591**	**0.3257**
	Medline TF 15	0.3364	0.2940	0.3015	0.2725	0.3530	0.3209

4.5 Query expansion using the ICD-10

ICD-10 is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list produced by the World Health Organization (WHO). This classification has information on several diseases, signs and symptoms, abnormal findings and external causes of injury [ICD16].

The approach using ICD-10 is similar to the one used in MedlinePlus taking advantage on the contents of the infobox from Wikipedia. The ICD-10 page contains not only information about the search concept but also about other diseases or symptoms related to the initial concept displayed in a hierarchy (See Figure 4.4). This related information is used as a source of terms to use on the query expansion process.

We chose the top 5, 10 and 15 most frequent terms of an ICD-10 page to append to the original query.



- J41 Simple and mucopurulent chronic bronchitis
- J42 Unspecified chronic bronchitis
- J43 Emphysema
- J44 Other chronic obstructive pulmonary disease
  - J44.0 Chronic obstructive pulmonary disease with acute lower respiratory infection
  - J44.1 Chronic obstructive pulmonary disease with acute exacerbation, unspecified
  - J44.8 Other specified chronic obstructive pulmonary disease
  - J44.9 Chronic obstructive pulmonary disease, unspecified
- J45 Asthma
  - J45.0 Predominantly allergic asthma
  - J45.1 Nonallergic asthma
  - J45.8 Mixed asthma
  - J45.9 Asthma, unspecified

Figure 4.4: ICD-10 hierarchy example when searching for Asthma.

4.5.1 Results

The ICD-10 approach (Table 4.11) shows that using related diseases or symptoms can improve the relevance of the retrieved documents. The change in the number of the most frequent terms used didn't show a significant difference in the results, but the more terms used the lower the relevance score. Only the run which used fewer terms obtained an improvement in readability. On the re-rank runs (Table 4.12) the best relevance score was obtained through the Basic formula using the SMOG metric, and the best readability score was obtained through the Arctan formula using the Flesch-Kincaid metric.

Table 4.11: ICD-10 results for CLEF eHealth 2015 collection.

Run	P@10	nDCG@10	uRBP	uRBPgr
Baseline	0.3455	0.3027	0.3148	0.3033
ICD-10 TF 5	**0.3970**	**0.3419**	**0.3242**	**0.3223**
ICD-10 TF 10	0.3939	0.3380	0.3125	0.3133
ICD-10 TF 15	0.3848	0.3355	0.3046	0.3070

Table 4.12: ICD-10 Re-Rank results for CLEF eHealth 2015 collection.

		SMOG		FOG		Flesch-Kincaid	
	Run	P@10	uRBP	P@10	uRBP	P@10	uRBP
	Baseline	0.3455	0.3148	0.3455	0.3148	0.3455	0.3148
Basic Formula (3.3)	ICD-10 TF 5	0.3561	**0.3065**	**0.3606**	**0.3176**	0.3333	**0.2962**
	ICD-10 TF 10	0.3591	0.2964	0.3576	0.3063	**0.3394**	0.2877
	ICD-10 TF 15	**0.3697**	0.2989	0.3591	0.2989	**0.3394**	0.2842
Log Formula (3.4)	ICD-10 TF 5	0.3364	0.2933	0.3152	**0.2810**	0.3242	**0.2925**
	ICD-10 TF 10	**0.3470**	**0.2960**	**0.3167**	0.2763	**0.3258**	0.2897
	ICD-10 TF 15	0.3455	**0.2960**	0.3136	0.2726	0.3242	0.2851
Arctan Formula (3.5)	ICD-10 TF 5	0.3364	0.3012	**0.2955**	**0.2698**	**0.3682**	**0.3419**
	ICD-10 TF 10	**0.3424**	**0.3023**	0.2924	0.2627	0.3621	0.3346
	ICD-10 TF 15	**0.3424**	0.2922	0.2939	0.2619	0.3652	0.3300

4.6 Query expansion using Latent Dirichlet Allocation over Wikipedia

Latent Dirichlet Allocation (LDA) is a generative probabilistic model of a corpus. The idea behind LDA is that every document can be deconstructed into sets of topics. These topics, in turn, are characterized by a distribution of words [BNJ03].

The modelling process of LDA can be described as finding a mixture of topics for each document, i.e., $P(z|d)$, with each topic described by terms following another probability distribution, i.e., P$(t|z)$ [BNJ03]. This can be formalized as:

$$P(t_i|d) = \sum_{j=1}^{Z} P(t_i|z_i = j)P(z_i = j|d) \tag{4.7}$$

where $P(t_i|d)$ is the probability of the ith term for a given document d and z_i is the latent topic. $P(t_i|z_i = j)$ is the probability of t_i within topic j. $P(z_i = j|d)$ is the probability of picking a term from topic j in the document [BNJ03]. It is possible to adjust the degree of specialization of the topics by specifying in advance the number of latent topics Z. Using a fixed number of topics and Dirichlet priors for the distributions LDA can estimate the topic-term distribution $P(t|z)$ and the document-topic distribution $P(z|d)$ from a document [BNJ03]. Gibbs Sampling [GS04] is one possible approach to this end. For each term t_i in a document d_i the Gibbs Sampling iterates several times, generating a new topic j for the term based on the probability $P(z_i = j|t_i, d_i, z_i)$ seen in Equation 4.8, until the LDA parameters converge.

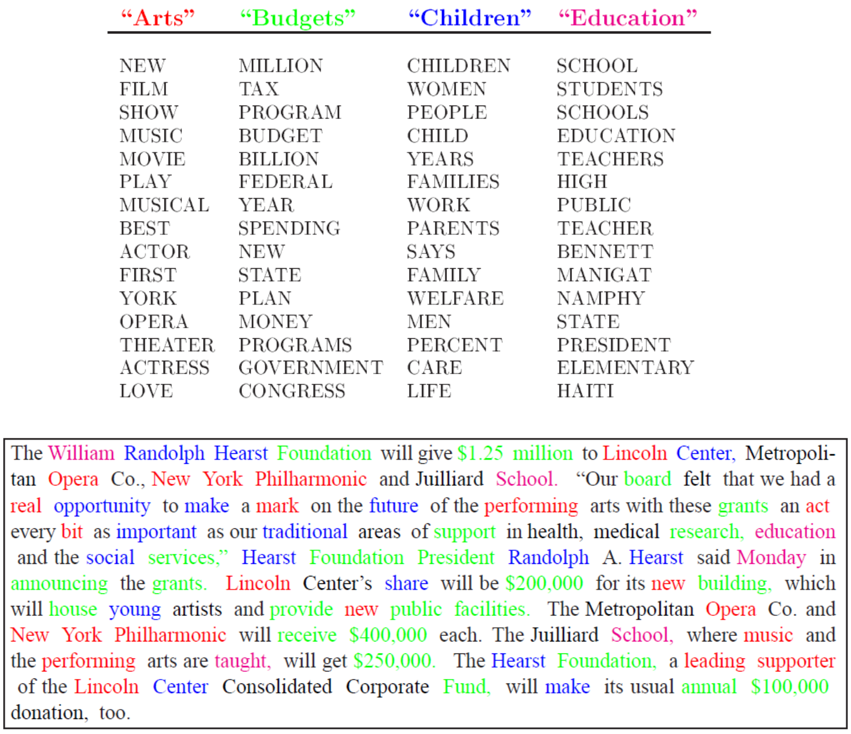
$$P(z_i = j|t_i, d_i, z_i) \propto \frac{C^{TZ}_{t_i j} + \beta}{\sum_t C^{TZ}_{tj} + T\beta} \frac{C^{DZ}_{d_i j} + \alpha}{\sum_z C^{DZ}_{d_i z} + Z\alpha} \tag{4.8}$$

C^{TZ} maintains a count of all topic-term assignments, C^{DZ} counts the document-topic assignments, all topic-term and document-topic assignments are represented by z_i and, α and β are the parameters for the Dirichlet priors, serving as smoothing parameters [GS04]. With this the probabilities in Equation 4.7 can be reformulated as [GS04]:

$$P(t_i|z_i = j) = \frac{C^{TZ}_{t_i j} + \beta}{\sum_t C^{TZ}_{t j} + T\beta} \tag{4.9}$$

$$P(z_i = j|d_i) = \frac{C^{DZ}_{d_i j} + \alpha}{\sum_z C^{DZ}_{d_i z} + Z\alpha} \tag{4.10}$$

Figure 4.5 shows an example of LDA applied to a text, generating four topics and a set of words representing each topic. These distributions seem to capture some of the underlying topics in the corpus. Each word on the text tends to peak towards one of the possible topic values, the words are color coded according to the topics they represent.



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 4.5: An example article from the TREC AP corpus. Each color codes a different factor from which the word is putatively generated [BNJ03].

One of LDA applications is the recommendation of tags for web documents. This can be shown in the works of Krestel, Fankhauser and Nejdl [KFN09] and, Choubey [Cho11].

Krestel, Fankhauser and Nejdl [KFN09] proved that using LDA to recommend topics achieved significantly better precision and recall than the use of association rules and also recommends more specific tags. Furthermore, extending documents with these tags significantly improves the search for new documents.

Choubey [Cho11] proposed two different approaches using LDA to address the tagging problem. The first approach was called topic words based approach, and it recommended the top words in the topics representing a document as tags for that particular document. The second approach, called topic distance based approach, used the tags of the most similar training documents to recommend tags for a test untagged document. Two different datasets were used to test these two approaches. One with only the descriptions corresponding to an URL, and another with crawled URL content. Choubey concluded that the topic distance based approach is better than the topic words based approach, when only the descriptions are used to construct documents, while the topic words based approach works better when the contents are used to construct documents.

For this approach JGibbLDA [PN16] was used to generate topics from texts related to each query. JGibbLDA [PN16] is a Java implementation of Latent Dirichlet Allocation (LDA) using Gibbs Sampling technique for parameter estimation and inference. The texts used were the Wikipedia articles obtained through the MediaWiki API using the MTI concepts as search terms.

In this approach we tried different combinations of number of topics and number of words to to identify which one contributes the most. We chose a combination of 3 topics with 1, 5 and 10 words, and 1, 5 and 10 topics with 5 words.

4.6.1 Results

The results for the LDA approach (Table 4.13) were the ones that differentiate the most between them regarding relevance, going from results worse than the baseline to one of significant improvement. The readability scores were always worse than the baseline. The best relevance scores on the re-rank runs (Table 4.14) were always the ones using 10 topics with 5 words each. The best readability scores were also brought these runs in the majority of cases.

Table 4.13: Latent Dirichlet Allocation results for CLEF eHealth 2015 collection.

Run	P@10	nDCG@10	uRBP	uRBPgr
Baseline	0.3455	0.3027	0.3148	0.3033
LDA 3T 1W	0.3591	0.3050	**0.2786**	0.2744
LDA 3T 5W	0.3333	0.2839	0.2480	0.2553
LDA 3T 10W	0.3545	0.3072	0.2345	0.2583
LDA 1T 5W	0.3455	0.3039	0.2717	**0.2754**
LDA 5T 5W	0.3394	0.3060	0.2352	0.2566
LDA 10T 5W	**0.3894**	**0.3353**	0.2453	0.2742

Table 4.14: Latent Dirichlet Allocation Re-Rank results for CLEF eHealth 2015 collection.

		SMOG		FOG		Flesch-Kincaid	
	Run	P@10	uRBP	P@10	uRBP	P@10	uRBP
	Baseline	0.3455	0.3148	0.3455	0.3148	0.3455	0.3148
Basic Formula (3.3)	LDA 3T 1W	0.3015	0.2640	0.3152	0.2641	0.2864	0.2410
	LDA 3T 5W	0.3015	0.2455	0.2955	0.2359	0.2742	0.2200
	LDA 3T 10W	0.3227	0.2434	0.3030	0.2343	0.2924	0.2200
	LDA 1T 5W	0.3242	**0.2765**	0.3136	0.2637	0.3015	0.2391
	LDA 5T 5W	0.3076	0.2289	0.3076	0.2350	0.2970	0.2224
	LDA 10T 5W	**0.3864**	0.2702	**0.3606**	**0.2720**	**0.3379**	**0.2574**
Log Formula (3.4)	LDA 3T 1W	0.2970	0.2492	0.2985	0.2391	0.2864	0.2426
	LDA 3T 5W	0.2712	0.2389	0.2667	0.2232	0.2727	0.2258
	LDA 3T 10W	0.2909	0.2389	0.2697	0.2228	0.2818	0.2231
	LDA 1T 5W	0.3121	0.2635	0.2985	0.2416	0.2985	0.2440
	LDA 5T 5W	0.2848	0.2378	0.2712	0.2235	0.2894	0.2235
	LDA 10T 5W	**0.3485**	**0.2704**	**0.3152**	**0.2647**	**0.3136**	**0.2497**
Arctan Formula (3.5)	LDA 3T 1W	0.3197	0.2733	**0.2833**	0.2459	0.3439	0.3091
	LDA 3T 5W	0.3091	0.2667	0.2485	0.2377	0.3515	0.2814
	LDA 3T 10W	0.3273	0.2720	0.2455	0.2382	0.3364	0.2695
	LDA 1T 5W	0.3379	0.2872	**0.2833**	0.2537	0.3621	**0.3171**
	LDA 5T 5W	0.3227	0.2622	0.2500	0.2343	0.3348	0.2710
	LDA 10T 5W	**0.3530**	**0.2895**	**0.2833**	**0.2617**	**0.3712**	0.3067

4.7 Query expansion using the Unified Medical Language System

The Unified Medical Language System (UMLS) [Bod04] is a repository of biomedical vocabularies developed by the US National Library of Medicine. UMLS more than 60 families of biomedical vocabularies with over 2 million names for 900,000 concepts that have relations between them.

The UMLS Metathesaurus has a REST API that provides access to its information. Using this API it is possible to extract terms using the definitions in UMLS related to the MTI concepts.

We chose the top 5, 10 and 15 most frequent terms on the UMLS definitions to be added to the original query.

4.7.1 Results

The use of definitions of medical concepts from the UMLS as a source of terms proved to be an effective way of improving the retrieval performance (Table 4.15). Changing the number of terms used didn't significantly affect the relevance scores. However, the best results in relevance and readability were brought using fewer terms. Few of the re-rank runs (Table 4.16) brought improvements in relevance, and only one brought improvements in readability when compared with the baseline.

Table 4.15: UMLS results for CLEF eHealth 2015 collection.

Run	P@10	nDCG@10	uRBP	uRBPgr
Baseline	0.3455	0.3027	0.3148	0.3033
UMLS TF 5	**0.4045**	**0.3716**	**0.3371**	**0.3383**
UMLS TF 10	0.3909	0.3397	0.2926	0.3062
UMLS TF 15	0.3985	0.3376	0.2872	0.3020

Table 4.16: UMLS Re-Rank results for CLEF eHealth 2015 collection.

		SMOG		FOG		Flesch-Kincaid	
	Run	P@10	uRBP	P@10	uRBP	P@10	uRBP
	Baseline	0.3455	0.3148	0.3455	0.3148	0.3455	0.3148
Basic Formula (3.3)	UMLS TF 5	**0.3636**	**0.2947**	0.3152	**0.2812**	0.3015	**0.2586**
	UMLS TF 10	0.3424	0.2763	0.3242	0.2684	**0.3136**	0.2582
	UMLS TF 15	0.3545	0.2818	**0.3258**	0.2708	0.3061	0.2553
Log Formula (3.4)	UMLS TF 5	0.3242	**0.2775**	0.3030	**0.2611**	0.3076	**0.2725**
	UMLS TF 10	0.3167	0.2648	**0.3061**	0.2605	**0.3152**	0.2660
	UMLS TF 15	**0.3258**	0.2673	0.2985	0.2580	0.3061	0.2677
Arctan Formula (3.5)	UMLS TF 5	**0.3652**	**0.3115**	**0.3106**	**0.2878**	**0.3667**	**0.3428**
	UMLS TF 10	0.3561	0.2884	0.2970	0.2717	0.3652	0.3185
	UMLS TF 15	0.3500	0.2827	0.3045	0.2725	0.3561	0.3071

Chapter 5

Results Discussion

In this chapter we will make several comparisons between the different approaches to evaluate their efficiency. We will separate these comparisons into three sections: (1) one comparing only the results form the query expansion runs; (2) one comparing the results of the re-rank runs; (3) and another comparing the best results of the query expansion runs and the re-rank runs.

5.1 Query Expansion Runs

In this section we will compare all the query expansion approaches based on their relevance and readability.

5.1.1 Relevance

Reviewing the query expansion runs is possible to conclude that every approach shows a relevance improvement when compared with the baseline. Table 5.1 shows the best run for each approach based on the $P@10$ score. Even though $P@10$ and $nDCG@10$ are both used to evaluate relevance, is possible to verify that the higher the score on $P@10$ it doesn't imply a high score on $nDCG@10$. Analysing the top runs of Table 5.1 we can conclude that the more scientific and heath-related sources brought better relevance scores.

Table 5.2 shows the best run for each approach, but now based on the $nDCG@10$ score. Comparing these two tables it is possible to verify the runs rearrangement, proving that even if a run has a high score on a binary evaluation that doesn't mean it will have a high score on a graded one. Furthermore, when evaluating with $nDCG@10$ the order of the documents has an impact on the score, this doesn't apply when evaluating with $P@10$.

Table 5.1: Best results for CLEF eHealth 2015 collection based on the *P*@10 score.

Run	P@10	nDCG@10
MTI	0.4061	0.3530
UMLS TF 5	0.4045	0.3716
ICD-10 TF 5	0.3970	0.3419
Wiki TF 10 Health	0.3894	0.3455
Medline TF 10	0.3894	0.3420
LDA 10T 5W	0.3894	0.3353
Wiki Link 0.25 Health	0.3848	0.3466
Kullback-Liebler	0.3576	0.3021
Baseline	0.3455	0.3027

Table 5.2: Best results for CLEF eHealth 2015 collection based on the *nDCG*@10 score.

Run	P@10	nDCG@10
UMLS TF 5	0.4045	0.3716
MTI	0.4061	0.3530
Wiki Link 0.25 Health	0.3848	0.3466
Wiki TF 10 Health	0.3894	0.3455
Medline TF 5	0.3879	0.3450
ICD-10 TF 5	0.3970	0.3419
LDA 10T 5W	0.3894	0.3353
Baseline	0.3455	0.3027
Kullback-Liebler	0.3576	0.3021

5.1.2 Readability

Based on the readability evaluations we can prove that even using only query expansion it is possible to improve its score. Analysing the top runs of Table 5.3 we can conclude that the more scientific and heath-related sources brought better readability scores. Comparing the Tables 5.3 and 5.4 we can conclude that even if a run has a high score on a binary evaluation that doesn't mean it will have a high score on a graded one.

Table 5.3: Best results for CLEF eHealth 2015 collection based on the *uRBP* score.

Run	uRBP	uRBPgr
MTI	0.3381	0.3276
UMLS TF 5	0.3371	0.3383
ICD-10 TF 5	0.3242	0.3223
Bose-Einstein	0.3212	0.3110
Wiki TF 5 Health	0.3189	0.3190
Medline AltNames	0.3180	0.3124
Baseline	0.3148	0.3033
Wiki Link 0.50 Health	0.3128	0.3093
LDA 3T 1W	0.2786	0.2744

Table 5.4: Best results for CLEF eHealth 2015 collection based on the *uRBPgr* score.

Run	uRBP	uRBPgr
UMLS TF 5	0.3371	0.3383
MTI	0.3381	0.3276
ICD-10 TF 5	0.3242	0.3223
Wiki TF 5 Health	0.3189	0.3190
Medline TF 5	0.3172	0.3169
Bose-Einstein	0.3212	0.3110
Wiki Link 0.50 Health	0.3128	0.3093
Baseline	0.3148	0.3033
LDA 1T 5W	0.2717	0.2754

5.2 Re-Rank Runs

In this section we will compare all the query expansion approaches after the readability re-rank, based on their relevance and readability. In addition we will make comparisons of the readability metrics and formulas.

For a better understanding of the different combinations of readability metrics and formulas, we will compare for each metric which formula is better and for each formula which metric brings better results.

5.2.1 Relevance

5.2.1.1 Metrics

Table 5.5, 5.6 and 5.7 shows the best relevance results using the SMOG, FOG and Flesch-Kincaid metrics for each approach, respectively. From all the runs shown in these tables only the one using PRF has a lower or equal score compared with the baseline in all the readability metrics. We can conclude that for the SMOG and FOG metric the Basic formula is the best one and, for the Flesch-Kincaid metric the Arctan formula is the one that gives better results.

Table 5.5: Best SMOG re-rank results for CLEF eHealth 2015 collection based on the *P*@10 score.

Run	Formula	P@10
LDA 10T 5W	Basic	0.3864
ICD-10 TF 15	Basic	0.3697
Wiki TF 15 Health	Basic	0.3667
UMLS TF 5	Arctan	0.3652
Wiki Link 0.25 Health	Basic	0.3652
Medline TF 10	Basic	0.3621
MTI	Basic	0.3470
Baseline	-	0.3455
Kullback-Liebler	Arctan	0.3318

Table 5.6: Best FOG re-rank results for CLEF eHealth 2015 collection based on the *P*@10 score.

Run	Formula	P@10
Wiki Link 0.50 Health	Basic	0.3621
LDA 10T 5W	Basic	0.3606
ICD-10 TF 5	Basic	0.3606
Wiki TF 15 Health	Basic	0.3591
Baseline	-	0.3455
Medline TF 10	Basic	0.3455
UMLS TF 15	Basic	0.3258
MTI	Basic	0.3227
Bose-Einstein	Basic	0.3167

Table 5.7: Best Flesch-Kincaid re-rank results for CLEF eHealth 2015 collection based on the *P*@10 score.

Run	Formula	P@10
LDA 10T 5W	Arctan	0.3712
ICD-10 TF 5	Arctan	0.3682
UMLS TF 5	Arctan	0.3667
Wiki Link 0.25	Arctan	0.3621
Wiki TF 5 Health	Arctan	0.3621
Medline TF 10	Arctan	0.3591
MTI	Arctan	0.3515
Baseline	-	0.3455
Bose-Einstein	Arctan	0.3455

5.2.1.2 Formulas

Table 5.8, 5.9 and 5.10 shows the best relevance results using the Basic, Log and Arctan formulas for each approach, respectively. From all the runs shown in these tables only the one using PRF has a lower or equal score compared with the baseline in all the readability metrics. We can conclude that for the Basic and Log formula the SMOG metric is the best one and, for the Arctan formula the Flesch-Kincaid metric is the one that gives better results.

Table 5.8: Best results for CLEF eHealth 2015 collection using the Basic formula based on the $P@10$ score.

Run	Metric	P@10
LDA 10T 5W	SMOG	0.3864
ICD-10 TF 15	SMOG	0.3697
Wiki TF 15 Health	SMOG	0.3667
Wiki Link 0.25 Health	SMOG	0.3652
UMLS TF 5	SMOG	0.3636
Medline TF 10	SMOG	0.3621
MTI	SMOG	0.3470
Baseline	-	0.3455
Bose-Einstein	SMOG	0.3197

Table 5.9: Best results for CLEF eHealth 2015 collection using the Log formula based on the $P@10$ score.

Run	Metric	P@10
Wiki TF 10 Health	SMOG	0.3515
LDA 10T 5W	SMOG	0.3485
Wiki Link 0.50 Health	SMOG	0.3485
ICD-10 TF 10	SMOG	0.3470
Baseline	-	0.3455
Medline TF 10	SMOG	0.3364
UMLS TF 15	SMOG	0.3258
MTI	SMOG	0.3227
Bose-Einstein	SMOG	0.3045

Table 5.10: Best results for CLEF eHealth 2015 collection using the Arctan formula based on the $P@10$ score.

Run	Metric	P@10
LDA 10T 5W	Flesch-Kincaid	0.3712
ICD-10 TF 5	Flesch-Kincaid	0.3682
UMLS TF 5	Flesch-Kincaid	0.3667
Wiki TF 5 Health	Flesch-Kincaid	0.3621
Wiki Link 0.25	Flesch-Kincaid	0.3621
Medline TF 10	Flesch-Kincaid	0.3621
MTI	Flesch-Kincaid	0.3515
Baseline	-	0.3455
Bose-Einstein	Flesch-Kincaid	0.3455

5.2.2 Readability

5.2.2.1 Metrics

Table 5.11, 5.12 and 5.13 shows the best readability results using the SMOG, FOG and Flesch-Kincaid metrics for each approach, respectively. Only the Flesch-Kincaid metric brought results that had a significant improvement over the baseline. We can conclude that for the FOG metric the Basic formula is the best one and, for the Flesch-Kincaid metric the Arctan formula is the one that gives better results. The SMOG metric even if the Arctan formula had the best result the most common formula in Table 5.11 is the Basic.

Table 5.11: Best SMOG re-rank results for CLEF eHealth 2015 collection based on the *uRBP* score.

Run	Formula	uRBP
Baseline	-	0.3148
UMLS TF 5	Arctan	0.3115
MTI	Basic	0.3087
Wiki TF 10 Health	Basic	0.3077
ICD-10 TF 5	Basic	0.3065
Medline TF 10	Basic	0.3051
Wiki Link 0.50 Health	Basic	0.2985
LDA 10T 5W	Arctan	0.2895
Kullback-Liebler	Arctan	0.2853

Table 5.12: Best FOG re-rank results for CLEF eHealth 2015 collection based on the *uRBP* score.

Run	Formula	uRBP
ICD-10 TF 5	Basic	0.3176
Baseline	-	0.3148
Wiki Link 0.50 Health	Basic	0.3104
Wiki TF 5 Health	Basic	0.3055
Medline TF 5	Basic	0.3019
MTI	Basic	0.2953
UMLS TF 5	Arctan	0.2878
Bose-Einstein	Basic	0.2809
LDA 10T 5W	Basic	0.2720

Table 5.13: Best Flesch-Kincaid re-rank results for CLEF eHealth 2015 collection based on the *uRBP* score.

Run	Formula	uRBP
UMLS TF 5	Arctan	0.3428
ICD-10 TF 5	Arctan	0.3419
Wiki Link 0.25 Health	Arctan	0.3314
Wiki TF 5 Health	Arctan	0.3293
MTI	Arctan	0.3288
Medline TF 10	Arctan	0.3257
LDA 1T 5W	Arctan	0.3171
Baseline	-	0.3148
Kullback-Liebler	Arctan	0.3134

5.2.2.2 Formulas

Table 5.14, 5.15 and 5.16 shows the best readability results using the Basic, Log and Arctan formulas for each approach, respectively. Only the Arctan formula using the Flesch-Kincaid metric brought results that had a significant improvement over the baseline. We can conclude that for the Basic formula the best metric is SMOG, for the Log formula is FOG and, for the Arctan formula is the Flesch-Kincaid metric.

Table 5.14: Best results for CLEF eHealth 2015 collection using the Basic formula based on the *uRBP* score.

Run	Metric	uRBP
ICD-10 TF 5	FOG	0.3176
Baseline	-	0.3148
Wiki Link 0.50 Health	FOG	0.3104
MTI	SMOG	0.3087
Wiki TF 10 Health	SMOG	0.3077
Medline TF 10	SMOG	0.3051
UMLS TF 5	SMOG	0.2947
Bose-Einstein	FOG	0.2809
LDA 1T 5W	SMOG	0.2765

Table 5.15: Best results for CLEF eHealth 2015 collection using the Log formula based on the *uRBP* score.

Run	Metric	uRBP
Baseline	-	0.3148
Wiki TF 10 Health	SMOG	0.3024
Wiki Link 0.50 Health	SMOG	0.2961
ICD-10 TF 10	SMOG	0.2960
Medline TF 10	SMOG	0.2908
MTI	SMOG	0.2889
UMLS TF 5	SMOG	0.2775
LDA 10T 5W	SMOG	0.2704
Kullback-Liebler	Flesch-Kincaid	0.2648

Table 5.16: Best results for CLEF eHealth 2015 collection using the Arctan formula based on the *uRBP* score.

Run	Metric	uRBP
UMLS TF 5	Flesch-Kincaid	0.3428
ICD-10 TF 5	Flesch-Kincaid	0.3419
Wiki Link 0.25 Health	Flesch-Kincaid	0.3314
Wiki TF 5 Health	Flesch-Kincaid	0.3293
MTI	Flesch-Kincaid	0.3288
Medline TF 10	Flesch-Kincaid	0.3257
LDA 1T 5W	Flesch-Kincaid	0.3171
Baseline	-	0.3148
Kullback-Liebler	Flesch-Kincaid	0.3134

5.3 Query Expansion and Re-Rank Runs

In this section we will compare the best results of the query expansion runs with the best of the re-rank runs for each approach. This will determine the effectiveness of the re-rank methods.

5.3.1 Relevance

Table 5.17 shows the best relevance scores of both the query expansion runs (QE Runs) and the re-rank runs (RR Runs). Analyzing these results we can see that for every approach the relevance scores on the re-rank runs are lower. Even the best score of the re-rank runs would be one of the worst on the query expansion runs. The combination of the SMOG metric and the Basic formula was the one that gave the best relevance scores on the re-rank runs.

Table 5.17: Comparison of the best relevance results of the Query Expansion runs and the Re-Rank runs for CLEF eHealth 2015 collection based on the $P@10$ score.

QE Runs		RR Runs			
Run	P@10	Run	P@10	Metric	Formula
MTI	0.4061	LDA 10T 5W	0.3864	SMOG	Basic
UMLS TF 5	0.4045	ICD-10 TF 15	0.3697	SMOG	Basic
ICD-10 TF 5	0.3970	UMLS TF 5	0.3667	Flesch-Kincaid	Arctan
LDA 10T 5W	0.3894	Wiki TF 15 Health	0.3667	SMOG	Basic
Wiki TF 10 Health	0.3894	Wiki Link 0.25 Health	0.3652	SMOG	Basic
Medline TF 10	0.3894	Medline TF 10	0.3621	SMOG	Basic
Wiki Link 0.25 Health	0.3848	MTI	0.3515	Flesch-Kincaid	Arctan
Kullback-Liebler	0.3576	Bose-Einstein	0.3455	Flesch-Kincaid	Arctan

5.3.2 Readability

Table 5.18 shows the best readability scores of both the query expansion runs (QE Runs) and the re-rank runs (RR Runs). Some of the re-rank runs improved the readability scores compared with the query expansion runs, even the higher and lower scores are better. The combination of the Flesch-Kincaid metric and the Arctan formula was the one that gave the best readability scores on the re-rank runs.

Table 5.18: Comparison of the best readability results of the Query Expansion runs and the Re-Rank runs for CLEF eHealth 2015 collection based on the *uRBP* score.

QE Runs		RR Runs			
Run	uRBP	Run	uRBP	Metric	Formula
MTI	0.3381	UMLS TF 5	0.3428	Flesch-Kincaid	Arctan
UMLS TF 5	0.3371	ICD-10 TF 5	0.3419	Flesch-Kincaid	Arctan
ICD-10 TF 5	0.3242	Wiki Link 0.25 Health	0.3314	Flesch-Kincaid	Arctan
Bose-Einstein	0.3212	Wiki TF 5 Health	0.3293	Flesch-Kincaid	Arctan
Wiki TF 5 Health	0.3189	MTI	0.3288	Flesch-Kincaid	Arctan
Medline AltNames	0.3180	Medline TF 10	0.3257	Flesch-Kincaid	Arctan
Wiki Link 0.50 Health	0.3128	LDA 1T 5W	0.3171	Flesch-Kincaid	Arctan
LDA 3T 1W	0.2786	Kullback-Liebler	0.3134	Flesch-Kincaid	Arctan

Chapter 6

Conclusion and Future Work

6.1 Conclusion

On the query expansion runs we can verify that every approach improves the relevance score and that almost all of them improves the readability. The results of the different approaches allow to determine that the more scientific and health-related the terms for the query expansion are, the best are the results. With this we can conclude that the main objective of this thesis was accomplished.

As for the re-rank runs some of them did improved over the baseline, in both relevance and readability. However, when compared with the query expansion runs of the same approach the relevance scores are always lower, and only a few increase the readability scores. For example the approach using MTI. This approach has the best scores on *P@*10 and *uRBP* of all the query expansion runs, yet in the re-rank runs these scores are significantly lower. These scores can be justified by two reasons: (1) the readability measure (uRBP) evaluates both the relevance and readability, so if the re-rank method improves the score of irrelevant documents the value of uRBP will decrease, this proves that relevant medical documents won't always be understood by an user; (2) the readability metrics (SMOG, FOG, Flesch-Kincaid) used are not well suited to evaluate documents of a specific area, these metrics give a readability score based on the number of polysyllabic words, but some words are complex not by its size but by its meaning, e.g., the word "shock" is a common word frequently used in everyday life, however, in medical and health materials, the meaning of "shock" could be when "not enough blood and oxygen can get to your organs and tissues causing low blood pressure".

Although the results for the re-rank methods did not improved significantly over the query expansion, the combinations of the SMOG metric and the Basic formula and, the Flesch-Kincaid metric and the Arctan formula, were the ones that brought the better results for relevance and readability, respectively.

6.2 Future Work

In future work, we will continue to explore new query expansion models to find an effective way of supporting patients to find useful medical information. Continuing to improve the current ones like the Latent Dirichlet Allocation that could be applied to other sources of texts such as the ones found through the Wikipedia Link Analysis or the MedlinePlus pages.

Machine learning was not an approach addressed in this thesis. However, it could be one that complements this work with an unique approach that uses new methods.

In addition, we would like to incorporate readability metrics that weren't based on sentence lengths or polysyllabic words but were based in concepts.

References

[AVR02] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, October 2002.

[B04] Vastag B. Low health literacy called a major problem. *JAMA*, 291(18):2181–2182, 2004.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[Bod04] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270, 2004.

[Bos16] Adam Bosworth. Putting health into the patient's hands. Disponível em `https://googleblog.blogspot.pt/2007/05/putting-health-into-patients-hands.html`, Janeiro 2016.

[Bro16] Aaron Brown. An update on google health and google powermeter. Disponível em `https://googleblog.blogspot.pt/2011/06/update-on-google-health-and-google.html`, Janeiro 2016.

[Cho11] Rahul Choubey. *Tag Recommendation Using Latent Dirichlet Allocation*. PhD thesis, Kansas State University, 2011.

[CLE16a] CLEF. Information access evaluation meets multilinguality, multimodality and interaction. Disponível em `http://clef2016.clef-initiative.eu/index.php?page=Pages/cfLabsParticipation.html`, Janeiro 2016.

[CLE16b] CLEF. Lab details clefehealth. Disponível em `http://clef2016.clef-initiative.eu/index.php?page=Pages/cfLabsParticipation.html#l1`, Janeiro 2016.

[CYTDP06] David Carmel, Elad Yom-Tov, Adam Darlow, and Dan Pelleg. What makes a query difficult? In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 390–397, New York, NY, USA, 2006. ACM.

[DGZ15] Eva D'hondt, Brigitte Grau, and Pierre Zweigenbaum. Limsi @ clef ehealth 2015 - task 2. 2015.

[DKP16] DKPro. Dkpro jwpl. Available in https://dkpro.github.io/dkpro-jwpl/, May 2016.

[DSD14] Yihan Deng, Matthaeus Stoehr, and Kerstin Denecke. Retrieving attitudes: Sentiment analysis from clinical narratives. In Goeuriot et al. [GJK+14], pages 12–15.

[Eft96] Efthimis N. Efthimiadis. Query expansion. *Annual Review of Information Systems and Technologys*, 31:121–187, 1996.

[FD13] Susannah Fox and Maeve Duggan. Health online 2013. Available in http://www.pewinternet.org/2013/01/15/health-online-2013/, 2013.

[Fra11] Massimo Franceschet. Pagerank: Standing on the shoulders of giants. *Commun. ACM*, 54(6):92–101, June 2011.

[GH15] Andia Ghoddousi and Jimmy Xiangji Huang. York university at clef ehealth 2015: Medical document retrieval. 2015.

[GJK+14] Lorraine Goeuriot, Gareth J. F. Jones, Liadh Kelly, Henning Müller, and Justin Zobel, editors. *Proceedings of the Proceedings of the MedIR workshop on Medical Information Retrieval (MedIR@SIGIR)*, number 1276 in CEUR Workshop Proceedings, Aachen, 2014.

[GKJ+14] Lorraine Goeuriot, Liadh Kelly, Gareth J.F. Jones, Henning Müller, and Justin Zobel. Report on the sigir 2014 workshop on medical information retrieval (medir). In Goeuriot et al. [GJK+14], pages 1–3.

[GS04] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April 2004.

[Hed16] Jonathan Hedley. jsoup: Java html parser. Available in https://jsoup.org/, June 2016.

[HM12] A. Hanbury and H. Muller. Khresmoi – multimodal multilingual medical information search. 2012.

[HNH15] Nghia Huynh, Thanh Tuan Nguyen, and Quoc Ho. Teamhcmus: A concept-based information retrieval approach for web medical documents. 2015.

[ICD16] ICD10. International classification of diseases. Available in http://www.who.int/classifications/icd/en/, May 2016.

[IS10] Hazra Imran and Aditi Sharan. Selecting effective expansion terms for better information retrieval. *IJCSA*, 7(2):52–64, 2010.

[JGM13] Alan R. Aronson James G. Mork, Antonio J. Jimeno Yepes. The nlm medical text indexer system for indexing biomedical literature. 2013.

[JK00] Kalervo Järvelin and Jaana Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’00, pages 41–48, New York, NY, USA, 2000. ACM.

[KFN09] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. Latent dirichlet allocation for tag recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys ’09, pages 61–68, New York, NY, USA, 2009. ACM.

[KTBG15] Nesrine KSENTINI, Mohamed TMAR, Mohand BOUGHANEM, and Faez GARGOURI. Miracl at clef 2015 : User-centred health information retrieval task. 2015.

[KZ14a] Bevan Koopman and Guido Zuccon. Relevation! : an open source system for information retrieval relevance assessment. In *ACM SIGIR 2014 : The 37th Annual ACM Special Interest Group on Information Retrieval*, pages 1243–1244, Gold Coast Convention and Exhibition Centre, Queensland, Australia, 2014.

[KZ14b] Bevan Koopman and Guido Zuccon. Why assessing relevance in medical ir is demanding. In Goeuriot et al. [GJK+14], pages 16–19.

[Lab16] Information Ecology Lab. Ubire. Available in https://github.com/ielab/ubire, June 2016.

[Lau69] Harry Mc Laughlin. SMOG Grading-a New Readability Formula. *Journal of Reading*, 12(8), 1969.

[Ley08] Loet Leydesdorff. On the normalization and visualization of author co-citation data: Salton's cosine versus the jaccard index. *Journal of the American Society for Information Science and Technology*, 59(1):77–85, 2008.

[LF15] Kuang Lu and Hui Fang. Event oriented query expansion for news event queries. 2015.

[LN15] Xiao Jie Liu and Jian-Yun Nie. Bridging layperson's queries with medical concepts-grium@clef2015 ehealth task 2. 2015.

[Lop08] Carla Teixeira Lopes. Health information retrieval state of the art report. Faculdade de Engenharia da Universidade do Porto, 2008.

[Lu15] Fangmei Lu. Employing query expansion models to help patients diagnose themselves. 2015.

[Luc16] Lucene. Apache lucene. Disponível em https://lucene.apache.org/core/, Janeiro 2016.

[LV01] P. Lyman and H. R. Varian. How much information. Retrieved from http://www.sims.berkeley.edu/how-much-info on November 29, 2001.

[LV09] Michaël R. Laurent and Tim J. Vickers. Seeking health information online: Does wikipedia matter? *Journal of the American Medical Informatics Association*, 16(4):471–479, 2009.

[MA] Catherine Berrut Mohannad Almasri, Jean-Pierre Chevallet. Exploiting wikipedia structure for short query expansion in cultural heritage.

[MBY07] Christian Middleton and Ricardo Baeza-Yates. A comparison of open source search engines. 2007.

[Med16a] Byline Media. The flesch grade level readability formula. Available in http://www.readabilityformulas.com/flesch-grade-level-readability-formula.php, May 2016.

[Med16b] MediaWiki. Welcome to mediawiki.org. Available in https://www.mediawiki.org/wiki/MediaWiki, May 2016.

[Med16c] MedIR. Medir'14: Medical information retrieval. Disponível em http://sigir.org/sigir2014/finalworkshops.php#MedIR, Janeiro 2016.

[Med16d] MedLinePlus. About medlineplus. Available in https://www.nlm.nih.gov/medlineplus/aboutmedlineplus.html, May 2016.

[Mic16] Microsoft. Healthvault. Disponível em https://www.healthvault.com/pt/en, Janeiro 2016.

[Moo51] Calvin N. Moores. Zatocoding applied to mechanical organization of knowledge. *American Documentation*, 2:20–32, 1951.

[MP82] Douglas R. McCallum and James L. Peterson. Computer-based readability indexes. In *Proceedings of the ACM '82 Conference*, ACM '82, pages 44–48, New York, NY, USA, 1982. ACM.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[MZ08] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, December 2008.

[NIS16] NIST. Tipster text program. Disponível em http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/, Janeiro 2016.

[OJK15] Heung-Seon Oh, Yuchul Jung, and Kwang-Young Kim. Kisti at clef ehealth 2015 task 2. 2015.

[PN16] Xuan-Hieu Phan and Cam-Tu Nguyen. A java implementation of latent dirichlet allocation (lda) using gibbs sampling for parameter estimation and inference. Available in http://jgibblda.sourceforge.net/, June 2016.

[Pro16] The Lemur Project. The clueweb12 dataset: Dataset details. Available in http://lemurproject.org/clueweb12/specs.php, May 2016.

[PZ07] Laurence Park and Yuye Zhang. On the distribution of user persistence for rank-biased precision. In Amanda Spink, Andrew Turpin, and Mingfang Wu, editors, *Proceedings of The Twelfth Australasian Document Computing Symposium*, pages 17–24, Melbourne, Australia, December 2007. RMIT University.

[PZG+15] João Palotti, Guido Zuccon, Lorraine Goeuriot, Liadh Kelly, Allan Hanbury, Gareth J.F. Jones, Mihai Lupu, and Pavel Pecina. Clef ehealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. 2015.

[Ric06] Ronald E Rice. Influences, usage, and outcomes of internet health information searching: Multivariate results from the pew surveys. *International Journal of Medical Informatics*, 75(1):8–28, 2006.

[RJ76] 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.

[San10] Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval*, 4(4):247–375, 2010.

[SBP15] Shadi Saleh, Feraena Bibyna, and Pavel Pecina. Cuni at the clef ehealth 2015 task 2. 2015.

[SC01] Luo Si and Jamie Callan. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 574–576, New York, NY, USA, 2001. ACM.

[SHH+15] Yang Song, Yun He, Qinmin Hu, Liang He, and E. Mark Haacke. Ecnu at 2015 ehealth task 2: User-centred health information retrieval. 2015.

[SIM14] Isabelle Stanton, Samuel Ieong, and Nina Mishra. Circumlocution in diagnostic medical queries. In *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, SIGIR '14, pages 133–142, New York, NY, USA, 2014. ACM.

[SWJS01] Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, and Tefko Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, 2001.

[TAM15] Edwin Thuma, George Anderson, and Gontlafetse Mosweunyane. Ubml participation to clef ehealth ir challenge 2015: Task 2. 2015.

[Ter16] Terrier. Welcome to the terrier ir platform. Disponível em http://terrier.org/, Janeiro 2016.

[TRE16a] TREC. trec eval. Available in http://trec.nist.gov/trec eval/, June 2016.

[TRE16b] TREC. Trec overview. Disponível em http://trec.nist.gov/overview.html, Janeiro 2016.

[URL16] URLBlacklist. About. Available in http://urlblacklist.com/, May 2016.

[Wik16a] Wikipedia. Help:infobox. Available in https://en.wikipedia.org/wiki/Help:Infobox, May 2016.

[Wik16b] Wikipedia. Wikipedia : Introduction. Available in https://en.wikipedia.org/wiki/Wikipedia:Introduction, May 2016.

[WV08] Tiffany M Walsh and Teresa A Volsko. Readability assessment of internet-based consumer health information. *Respiratory Care*, 53(10):1310–1315, 2008.

[WWP14] R. Constance Wiener and Regina Wiener-Pla. Literacy, pregnancy and potential oral health changes: The internet and readability levels. *Maternal and Child Health Journal*, 18(3):657–662, 2014.

[YHM14] Sukjin You, Wei Huang, and Xiangming Mu. Uwm-hbut at trec 2014 microblog track: Using query expansion (qe) and event identification algorithm (eia) to improve microblog retrieval effectiveness. 2014.

[ZHF15] Sihui Zhang, Bin He, and Weiguo Fan. Cbia vt at trec 2015 clinical decision support track - exploring relevance feedback and query expansion in biomedical information retrieval. 2015.

[Zic13] Kathryn Zickuhr. Who's not online and why. Available in http://www.pewinternet.org/2013/09/25/whos-not-online-and-why-2/, 2013.

[ZK14] Guido Zuccon and Bevan Koopman. Integrating understandability in the evaluation of consumer health search engines. In Goeuriot et al. [GJK+14], pages 32–35.

[ZKP$^+$04] Qing T. Zeng, Sandra Kogan, Robert M. Plovnick, Jonathan Crowell, Eve-Marie Lacroix, and Robert A. Greenes. Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. *I. J. Medical Informatics*, 73(1):45–55, 2004.

[ZMG08] Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May 2008. electronic proceedings.

[Zuc16] Guido Zuccon. *Understandability Biased Evaluation for Information Retrieval*, pages 280–292. Springer International Publishing, Cham, 2016.

Appendix A

Readability Re-Rank Results

A.1 Pseudo Relevance Feedback

Table A.1: Pseudo Relevance Feedback SMOG Re-Rank results for CLEF eHealth 2015 collection.

		SMOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	Bose-Einstein	0.3197	0.2530	0.3201	0.2779	0.2671
	Kullback-Liebler	0.3167	0.2458	0.3144	0.2712	0.2607
Log Formula (3.4)	Bose-Einstein	0.3045	0.2308	0.3003	0.2560	0.2487
	Kullback-Liebler	0.3000	0.2247	0.2994	0.2541	0.2468
Arctan Formula (3.5)	Bose-Einstein	0.3227	0.2537	0.3272	0.2817	0.2724
	Kullback-Liebler	0.3318	0.2565	0.3312	0.2853	0.2754

Table A.2: Pseudo Relevance Feedback FOG Re-Rank results for CLEF eHealth 2015 collection.

		FOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	Bose-Einstein	0.3167	0.2413	0.3162	0.2809	0.2699
	Kullback-Liebler	0.3121	0.2360	0.3157	0.2800	0.2684
Log Formula (3.4)	Bose-Einstein	0.2970	0.2200	0.2837	0.2532	0.2447
	Kullback-Liebler	0.2864	0.2117	0.2828	0.2527	0.2430
Arctan Formula (3.5)	Bose-Einstein	0.2848	0.2221	0.2844	0.2565	0.2472
	Kullback-Liebler	0.2773	0.2203	0.2880	0.2598	0.2498

Table A.3: Pseudo Relevance Feedback Flesch-Kincaid Re-Rank results for CLEF eHealth 2015 collection.

		Flesch-Kincaid				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	Bose-Einstein	0.3076	0.2313	0.3025	0.2704	0.2595
	Kullback-Liebler	0.3030	0.2247	0.2973	0.2647	0.2530
Log Formula (3.4)	Bose-Einstein	0.2939	0.2284	0.2974	0.2638	0.2547
	Kullback-Liebler	0.2894	0.2215	0.2981	0.2648	0.2546
Arctan Formula (3.5)	Bose-Einstein	0.3455	0.2842	0.3502	0.3128	0.3008
	Kullback-Liebler	0.3439	0.2811	0.3516	0.3134	0.3012

A.2 Medical Text Indexer

Table A.4: MTI SMOG Re-Rank results for CLEF eHealth 2015 collection.

		SMOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	MTI	0.3470	0.2864	0.3474	0.3087	0.2951
Log Formula (3.4)	MTI	0.3227	0.2619	0.3264	0.2889	0.2773
Arctan Formula (3.5)	MTI	0.3379	0.2725	0.3502	0.3050	0.2934

Table A.5: MTI FOG Re-Rank results for CLEF eHealth 2015 collection.

		FOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	MTI	0.3227	0.2610	0.3230	0.2953	0.2844
Log Formula (3.4)	MTI	0.2985	0.2276	0.2915	0.2684	0.2567
Arctan Formula (3.5)	MTI	0.2909	0.2278	0.2908	0.2697	0.2585

Table A.6: MTI Flesch-Kincaid Re-Rank results for CLEF eHealth 2015 collection.

		Flesch-Kincaid				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	MTI	0.3167	0.2394	0.3032	0.2768	0.2664
Log Formula (3.4)	MTI	0.3136	0.2390	0.3067	0.2814	0.2703
Arctan Formula (3.5)	MTI	0.3515	0.3053	0.3668	0.3288	0.3178

A.3 Wikipedia Term-Frequency

Table A.7: Wikipedia Term-Frequency SMOG Re-Rank results for CLEF eHealth 2015 collection.

		SMOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	Wiki TF 5	0.3379	0.2684	0.3240	0.2715	0.2658
	Wiki TF 10	0.3242	0.2707	0.3202	0.2596	0.2563
	Wiki TF 15	0.3364	0.2756	0.3398	0.2585	0.2625
	Wiki TF 5 Health	0.3591	0.2985	0.3585	0.3048	0.2981
	Wiki TF 10 Health	0.3652	0.3095	0.3653	0.3077	0.3026
	Wiki TF 15 Health	0.3667	0.3020	0.3619	0.2957	0.2951
Log Formula (3.4)	Wiki TF 5	0.3227	0.2577	0.3106	0.2646	0.2584
	Wiki TF 10	0.3227	0.2637	0.3188	0.2666	0.2611
	Wiki TF 15	0.3364	0.2719	0.3315	0.2693	0.2670
	Wiki TF 5 Health	0.3455	0.2870	0.3453	0.2971	0.2902
	Wiki TF 10 Health	0.3515	0.2948	0.3524	0.3024	0.2949
	Wiki TF 15 Health	0.3470	0.2871	0.3494	0.2952	0.2901
Arctan Formula (3.5)	Wiki TF 5	0.3455	0.2786	0.3413	0.2855	0.2788
	Wiki TF 10	0.3455	0.2713	0.3300	0.2713	0.2637
	Wiki TF 15	0.3530	0.2835	0.3446	0.2763	0.2734
	Wiki TF 5 Health	0.3530	0.2841	0.3503	0.2965	0.2913
	Wiki TF 10 Health	0.3515	0.2840	0.3495	0.2970	0.2894
	Wiki TF 15 Health	0.3485	0.2825	0.3486	0.2935	0.2878

Table A.8: Wikipedia Term-Frequency FOG Re-Rank results for CLEF eHealth 2015 collection.

		FOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	Wiki TF 5	0.3136	0.2574	0.3092	0.2709	0.2636
	Wiki TF 10	0.3106	0.2689	0.3089	0.2583	0.2573
	Wiki TF 15	0.3333	0.2845	0.3223	0.2551	0.2619
	Wiki TF 5 Health	0.3500	0.2964	0.3458	0.3055	0.2973
	Wiki TF 10 Health	0.3576	0.3097	0.3546	0.3019	0.2989
	Wiki TF 15 Health	0.3591	0.3109	0.3588	0.2942	0.2976
Log Formula (3.4)	Wiki TF 5	0.3106	0.2330	0.2898	0.2619	0.2512
	Wiki TF 10	0.3182	0.2554	0.3010	0.2690	0.2614
	Wiki TF 15	0.3318	0.2586	0.3071	0.2710	0.2653
	Wiki TF 5 Health	0.3152	0.2513	0.3098	0.2767	0.2672
	Wiki TF 10 Health	0.3197	0.2729	0.3178	0.2828	0.2749
	Wiki TF 15 Health	0.3227	0.2620	0.3149	0.2772	0.2715
Arctan Formula (3.5)	Wiki TF 5	0.2864	0.2207	0.2821	0.2630	0.2508
	Wiki TF 10	0.3000	0.2273	0.2781	0.2572	0.2462
	Wiki TF 15	0.3045	0.2371	0.2882	0.2657	0.2552
	Wiki TF 5 Health	0.2970	0.2399	0.2986	0.2724	0.2619
	Wiki TF 10 Health	0.2970	0.2438	0.2971	0.2715	0.2604
	Wiki TF 15 Health	0.3000	0.2419	0.2956	0.2692	0.2591

Table A.9: Wikipedia Term-Frequency Flesch-Kincaid Re-Rank results for CLEF eHealth 2015 collection.

		Flesch-Kincaid				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	Wiki TF 5	0.3061	0.2381	0.2915	0.2548	0.2475
	Wiki TF 10	0.3136	0.2528	0.2966	0.2505	0.2476
	Wiki TF 15	0.3182	0.2553	0.3093	0.2487	0.2528
	Wiki TF 5 Health	0.3152	0.2564	0.3166	0.2790	0.2705
	Wiki TF 10 Health	0.3394	0.2873	0.3356	0.2861	0.2816
	Wiki TF 15 Health	0.3364	0.2783	0.3396	0.2770	0.2796
Log Formula (3.4)	Wiki TF 5	0.3091	0.2421	0.2974	0.2660	0.2552
	Wiki TF 10	0.3152	0.2605	0.3041	0.2650	0.2584
	Wiki TF 15	0.3182	0.2533	0.3082	0.2599	0.2590
	Wiki TF 5 Health	0.3167	0.2611	0.3227	0.2849	0.2758
	Wiki TF 10 Health	0.3167	0.2738	0.3275	0.2866	0.2794
	Wiki TF 15 Health	0.3212	0.2653	0.3302	0.2805	0.2779
Arctan Formula (3.5)	Wiki TF 5	0.3379	0.2923	0.3515	0.3024	0.2939
	Wiki TF 10	0.3455	0.2943	0.3483	0.2982	0.2911
	Wiki TF 15	0.3515	0.2988	0.3528	0.2941	0.2908
	Wiki TF 5 Health	0.3621	0.3177	0.3734	0.3293	0.3189
	Wiki TF 10 Health	0.3576	0.3162	0.3706	0.3239	0.3149
	Wiki TF 15 Health	0.3515	0.3103	0.3672	0.3176	0.3105

A.4 Wikipedia Link Analysis

Table A.10: Wikipedia Link Analysis SMOG Re-Rank results for CLEF eHealth 2015 collection.

		SMOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	Wiki Link 0.25	0.3348	0.2839	0.3298	0.2575	0.2544
	Wiki Link 0.50	0.3076	0.2696	0.3087	0.2545	0.2468
	Wiki Link 0.75	0.3076	0.2572	0.2910	0.2500	0.2416
	Wiki Link 0.25 Health	0.3652	0.3112	0.3538	0.2903	0.2840
	Wiki Link 0.50 Health	0.3636	0.3040	0.3495	0.2985	0.2893
	Wiki Link 0.75 Health	0.3500	0.2843	0.3386	0.2915	0.2832
Log Formula (3.4)	Wiki Link 0.25	0.3242	0.2738	0.3213	0.2628	0.2555
	Wiki Link 0.50	0.3076	0.2647	0.3048	0.2560	0.2469
	Wiki Link 0.75	0.3076	0.2556	0.2943	0.2533	0.2453
	Wiki Link 0.25 Health	0.3455	0.2944	0.3442	0.2863	0.2795
	Wiki Link 0.50 Health	0.3485	0.2900	0.3441	0.2961	0.2869
	Wiki Link 0.75 Health	0.3364	0.2709	0.3300	0.2820	0.2759
Arctan Formula (3.5)	Wiki Link 0.25	0.3333	0.2780	0.3376	0.2820	0.2728
	Wiki Link 0.50	0.3182	0.2748	0.3286	0.2773	0.2679
	Wiki Link 0.75	0.3136	0.2593	0.3156	0.2675	0.2611
	Wiki Link 0.25 Health	0.3333	0.2807	0.3412	0.2811	0.2747
	Wiki Link 0.50 Health	0.3455	0.2913	0.3486	0.2984	0.2886
	Wiki Link 0.75 Health	0.3409	0.2747	0.3409	0.2904	0.2843

Table A.11: Wikipedia Link Analysis FOG Re-Rank results for CLEF eHealth 2015 collection.

		FOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	Wiki Link 0.25	0.3273	0.2852	0.3214	0.2594	0.2574
	Wiki Link 0.50	0.3030	0.2619	0.3001	0.2534	0.2463
	Wiki Link 0.75	0.3045	0.2571	0.2895	0.2541	0.2462
	Wiki Link 0.25 Health	0.3530	0.3089	0.3530	0.3041	0.2940
	Wiki Link 0.50 Health	0.3621	0.3086	0.3495	0.3104	0.2990
	Wiki Link 0.75 Health	0.3515	0.2913	0.3419	0.3070	0.2949
Log Formula (3.4)	Wiki Link 0.25	0.3061	0.2649	0.2961	0.2565	0.2484
	Wiki Link 0.50	0.2909	0.2511	0.2851	0.2507	0.2408
	Wiki Link 0.75	0.3000	0.2421	0.2829	0.2518	0.2428
	Wiki Link 0.25 Health	0.3227	0.2798	0.3226	0.2848	0.2752
	Wiki Link 0.50 Health	0.3273	0.2795	0.3216	0.2860	0.2762
	Wiki Link 0.75 Health	0.3303	0.2742	0.3193	0.2833	0.2742
Arctan Formula (3.5)	Wiki Link 0.25	0.2939	0.2493	0.2890	0.2632	0.2521
	Wiki Link 0.50	0.2818	0.2361	0.2762	0.2532	0.2419
	Wiki Link 0.75	0.2864	0.2331	0.2661	0.2442	0.2346
	Wiki Link 0.25 Health	0.3015	0.2638	0.3018	0.2771	0.2659
	Wiki Link 0.50 Health	0.3091	0.2627	0.3054	0.2800	0.2683
	Wiki Link 0.75 Health	0.3091	0.2502	0.2958	0.2690	0.2586

Table A.12: Wikipedia Link Analysis Flesch-Kincaid Re-Rank results for CLEF eHealth 2015 collection.

		Flesch-Kincaid				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	Wiki Link 0.25	0.3121	0.2653	0.3029	0.2460	0.2432
	Wiki Link 0.50	0.2894	0.2486	0.2850	0.2417	0.2348
	Wiki Link 0.75	0.2939	0.2420	0.2756	0.2402	0.2331
	Wiki Link 0.25 Health	0.3409	0.2928	0.3408	0.2916	0.2828
	Wiki Link 0.50 Health	0.3348	0.2877	0.3320	0.2923	0.2820
	Wiki Link 0.75 Health	0.3227	0.2641	0.3202	0.2822	0.2728
Log Formula (3.4)	Wiki Link 0.25	0.3152	0.2707	0.3081	0.2584	0.2533
	Wiki Link 0.50	0.2909	0.2503	0.2942	0.2526	0.2450
	Wiki Link 0.75	0.3061	0.2459	0.2857	0.2519	0.2439
	Wiki Link 0.25 Health	0.3394	0.2923	0.3379	0.2943	0.2854
	Wiki Link 0.50 Health	0.3364	0.2874	0.3344	0.2958	0.2858
	Wiki Link 0.75 Health	0.3242	0.2675	0.3260	0.2885	0.2792
Arctan Formula (3.5)	Wiki Link 0.25	0.3621	0.3209	0.3640	0.3116	0.3030
	Wiki Link 0.50	0.3333	0.3039	0.3507	0.3048	0.2953
	Wiki Link 0.75	0.3288	0.2898	0.3356	0.2967	0.2879
	Wiki Link 0.25 Health	0.3591	0.3223	0.3785	0.3314	0.3179
	Wiki Link 0.50 Health	0.3576	0.3160	0.3684	0.3289	0.3152
	Wiki Link 0.75 Health	0.3545	0.3015	0.3646	0.3261	0.3138

A.5 MedlinePlus

Table A.13: MedlinePlus SMOG Re-Rank results for CLEF eHealth 2015 collection.

		SMOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	Medline AltNames	0.3364	0.2614	0.3355	0.2912	0.2808
	Medline TF 5	0.3500	0.2856	0.3495	0.3014	0.2914
	Medline TF 10	0.3621	0.2937	0.3566	0.3051	0.2961
	Medline TF 15	0.3515	0.2906	0.3540	0.2950	0.2888
Log Formula (3.4)	Medline AltNames	0.3242	0.2527	0.3248	0.2785	0.2703
	Medline TF 5	0.3318	0.2675	0.3342	0.2851	0.2770
	Medline TF 10	0.3364	0.2740	0.3392	0.2908	0.2829
	Medline TF 15	0.3318	0.2763	0.3381	0.2886	0.2815
Arctan Formula (3.5)	Medline AltNames	0.3379	0.2649	0.3417	0.2931	0.2841
	Medline TF 5	0.3409	0.2734	0.3467	0.2941	0.2854
	Medline TF 10	0.3439	0.2779	0.3528	0.2999	0.2920
	Medline TF 15	0.3364	0.2761	0.3469	0.2940	0.2863

Table A.14: MedlinePlus FOG Re-Rank results for CLEF eHealth 2015 collection.

		FOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	Medline AltNames	0.3288	0.2569	0.3216	0.2891	0.2783
	Medline TF 5	0.3439	0.2850	0.3371	0.3019	0.2923
	Medline TF 10	0.3455	0.2845	0.3376	0.2944	0.2880
	Medline TF 15	0.3379	0.2822	0.3401	0.2902	0.2860
Log Formula (3.4)	Medline AltNames	0.3030	0.2350	0.2984	0.2692	0.2590
	Medline TF 5	0.3182	0.2574	0.3075	0.2772	0.2676
	Medline TF 10	0.3197	0.2565	0.3115	0.2803	0.2715
	Medline TF 15	0.3091	0.2506	0.3060	0.2693	0.2626
Arctan Formula (3.5)	Medline AltNames	0.2985	0.2358	0.2975	0.2739	0.2617
	Medline TF 5	0.3030	0.2460	0.3009	0.2769	0.2648
	Medline TF 10	0.3106	0.2492	0.3077	0.2833	0.2720
	Medline TF 15	0.3015	0.2441	0.2968	0.2725	0.2613

Table A.15: MedlinePlus Flesch-Kincaid Re-Rank results for CLEF eHealth 2015 collection.

		Flesch-Kincaid				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	Medline AltNames	0.3045	0.2354	0.2981	0.2663	0.2562
	Medline TF 5	0.3091	0.2477	0.3040	0.2709	0.2605
	Medline TF 10	0.3121	0.2472	0.3064	0.2662	0.2592
	Medline TF 15	0.3106	0.2487	0.3104	0.2629	0.2588
Log Formula (3.4)	Medline AltNames	0.3106	0.2414	0.3046	0.2744	0.2634
	Medline TF 5	0.3076	0.2501	0.3119	0.2801	0.2689
	Medline TF 10	0.3091	0.2468	0.3102	0.2755	0.2663
	Medline TF 15	0.3076	0.2471	0.3134	0.2719	0.2649
Arctan Formula (3.5)	Medline AltNames	0.3455	0.2883	0.3533	0.3165	0.3032
	Medline TF 5	0.3545	0.3098	0.3674	0.3253	0.3128
	Medline TF 10	0.3591	0.3143	0.3687	0.3257	0.3132
	Medline TF 15	0.3530	0.3101	0.3685	0.3209	0.3095

A.6 ICD-10

Table A.16: ICD-10 SMOG Re-Rank results for CLEF eHealth 2015 collection.

		SMOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	ICD-10 TF 5	0.3561	0.2967	0.3581	0.3065	0.2980
	ICD-10 TF 10	0.3591	0.2956	0.3548	0.2964	0.2921
	ICD-10 TF 15	0.3697	0.3082	0.3606	0.2989	0.2969
Log Formula (3.4)	ICD-10 TF 5	0.3364	0.2827	0.3416	0.2933	0.2853
	ICD-10 TF 10	0.3470	0.2923	0.3463	0.2960	0.2892
	ICD-10 TF 15	0.3455	0.2926	0.3476	0.2960	0.2903
Arctan Formula (3.5)	ICD-10 TF 5	0.3364	0.2821	0.3501	0.3012	0.2920
	ICD-10 TF 10	0.3424	0.2811	0.3503	0.3023	0.2934
	ICD-10 TF 15	0.3424	0.2798	0.3454	0.2922	0.2860

Table A.17: ICD-10 FOG Re-Rank results for CLEF eHealth 2015 collection.

		FOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	ICD-10 TF 5	0.3606	0.3013	0.3555	0.3176	0.3067
	ICD-10 TF 10	0.3576	0.3005	0.3516	0.3063	0.2996
	ICD-10 TF 15	0.3591	0.2980	0.3476	0.2989	0.2942
Log Formula (3.4)	ICD-10 TF 5	0.3152	0.2572	0.3099	0.2810	0.2703
	ICD-10 TF 10	0.3167	0.2532	0.3056	0.2763	0.2666
	ICD-10 TF 15	0.3136	0.2505	0.3024	0.2726	0.2631
Arctan Formula (3.5)	ICD-10 TF 5	0.2955	0.2398	0.2955	0.2698	0.2591
	ICD-10 TF 10	0.2924	0.2367	0.2891	0.2627	0.2529
	ICD-10 TF 15	0.2939	0.2370	0.2880	0.2619	0.2517

Table A.18: ICD-10 Flesch-Kincaid Re-Rank results for CLEF eHealth 2015 collection.

		Flesch-Kincaid				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	ICD-10 TF 5	0.3333	0.2797	0.3325	0.2962	0.2862
	ICD-10 TF 10	0.3394	0.2821	0.3304	0.2877	0.2819
	ICD-10 TF 15	0.3394	0.2808	0.3271	0.2842	0.2784
Log Formula (3.4)	ICD-10 TF 5	0.3242	0.2657	0.3261	0.2925	0.2826
	ICD-10 TF 10	0.3258	0.2702	0.3268	0.2897	0.2820
	ICD-10 TF 15	0.3242	0.2624	0.3221	0.2851	0.2777
Arctan Formula (3.5)	ICD-10 TF 5	0.3682	0.3205	0.3818	0.3419	0.3285
	ICD-10 TF 10	0.3621	0.3128	0.3785	0.3346	0.3244
	ICD-10 TF 15	0.3652	0.3157	0.3759	0.3300	0.3210

A.7 Latent Dirichlet Allocation

Table A.19: Latent Dirichlet Allocation SMOG Re-Rank results for CLEF eHealth 2015 collection.

		SMOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	LDA 3T 1W	0.3015	0.2336	0.3020	0.2640	0.2499
	LDA 3T 5W	0.3015	0.2423	0.2978	0.2455	0.2392
	LDA 3T 10W	0.3227	0.2631	0.3147	0.2434	0.2469
	LDA 1T 5W	0.3242	0.2616	0.3259	0.2765	0.2688
	LDA 5T 5W	0.3076	0.2541	0.3027	0.2289	0.2346
	LDA 10T 5W	0.3864	0.3256	0.3810	0.2702	0.2834
Log Formula (3.4)	LDA 3T 1W	0.2970	0.2245	0.2915	0.2492	0.2399
	LDA 3T 5W	0.2712	0.2168	0.2823	0.2389	0.2325
	LDA 3T 10W	0.2909	0.2321	0.2932	0.2389	0.2385
	LDA 1T 5W	0.3121	0.2411	0.3088	0.2635	0.2554
	LDA 5T 5W	0.2848	0.2334	0.2960	0.2378	0.2382
	LDA 10T 5W	0.3485	0.2789	0.3440	0.2704	0.2715
Arctan Formula (3.5)	LDA 3T 1W	0.3197	0.2407	0.3149	0.2733	0.2607
	LDA 3T 5W	0.3091	0.2440	0.3109	0.2667	0.2569
	LDA 3T 10W	0.3273	0.2626	0.3271	0.2720	0.2678
	LDA 1T 5W	0.3379	0.2596	0.3296	0.2872	0.2752
	LDA 5T 5W	0.3227	0.2603	0.3213	0.2622	0.2580
	LDA 10T 5W	0.3530	0.2897	0.3667	0.2895	0.2876

Table A.20: Latent Dirichlet Allocation FOG Re-Rank results for CLEF eHealth 2015 collection.

		FOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	LDA 3T 1W	0.3152	0.2406	0.2944	0.2641	0.2525
	LDA 3T 5W	0.2955	0.2373	0.2813	0.2359	0.2315
	LDA 3T 10W	0.3030	0.2513	0.3043	0.2343	0.2417
	LDA 1T 5W	0.3136	0.2489	0.3038	0.2637	0.2573
	LDA 5T 5W	0.3076	0.2523	0.3039	0.2350	0.2412
	LDA 10T 5W	0.3606	0.3033	0.3515	0.2720	0.2784
Log Formula (3.4)	LDA 3T 1W	0.2985	0.2108	0.2651	0.2391	0.2302
	LDA 3T 5W	0.2667	0.2040	0.2502	0.2232	0.2168
	LDA 3T 10W	0.2697	0.2047	0.2608	0.2228	0.2214
	LDA 1T 5W	0.2985	0.2151	0.2704	0.2416	0.2331
	LDA 5T 5W	0.2712	0.2019	0.2641	0.2235	0.2218
	LDA 10T 5W	0.3152	0.2520	0.3065	0.2647	0.2628
Arctan Formula (3.5)	LDA 3T 1W	0.2833	0.2108	0.2692	0.2459	0.2371
	LDA 3T 5W	0.2485	0.1960	0.2555	0.2377	0.2280
	LDA 3T 10W	0.2455	0.1932	0.2625	0.2382	0.2318
	LDA 1T 5W	0.2833	0.2177	0.2762	0.2537	0.2435
	LDA 5T 5W	0.2500	0.1935	0.2625	0.2343	0.2282
	LDA 10T 5W	0.2833	0.2250	0.2880	0.2617	0.2550

Table A.21: Latent Dirichlet Allocation Flesch-Kincaid Re-Rank results for CLEF eHealth 2015 collection.

		Flesch-Kincaid				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	LDA 3T 1W	0.2864	0.2091	0.2726	0.2410	0.2323
	LDA 3T 5W	0.2742	0.2022	0.2587	0.2200	0.2146
	LDA 3T 10W	0.2924	0.2258	0.2790	0.2200	0.2246
	LDA 1T 5W	0.3015	0.2249	0.2785	0.2391	0.2337
	LDA 5T 5W	0.2970	0.2314	0.2840	0.2224	0.2280
	LDA 10T 5W	0.3379	0.2791	0.3280	0.2574	0.2624
Log Formula (3.4)	LDA 3T 1W	0.2864	0.2093	0.2731	0.2426	0.2341
	LDA 3T 5W	0.2727	0.2023	0.2592	0.2258	0.2201
	LDA 3T 10W	0.2818	0.2117	0.2706	0.2231	0.2244
	LDA 1T 5W	0.2985	0.2200	0.2812	0.2440	0.2380
	LDA 5T 5W	0.2894	0.2138	0.2711	0.2235	0.2241
	LDA 10T 5W	0.3136	0.2453	0.3059	0.2497	0.2528
Arctan Formula (3.5)	LDA 3T 1W	0.3439	0.2849	0.3471	0.3091	0.2936
	LDA 3T 5W	0.3515	0.2810	0.3371	0.2814	0.2755
	LDA 3T 10W	0.3364	0.2690	0.3319	0.2695	0.2683
	LDA 1T 5W	0.3621	0.3034	0.3675	0.3171	0.3064
	LDA 5T 5W	0.3348	0.2750	0.3337	0.2710	0.2696
	LDA 10T 5W	0.3712	0.3016	0.3693	0.3067	0.3060

A.8 UMLS

Table A.22: UMLS SMOG Re-Rank results for CLEF eHealth 2015 collection.

		SMOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	UMLS TF 5	0.3636	0.2895	0.3594	0.2947	0.2911
	UMLS TF 10	0.3424	0.2807	0.3524	0.2763	0.2756
	UMLS TF 15	0.3545	0.2945	0.3562	0.2818	0.2801
Log Formula (3.4)	UMLS TF 5	0.3242	0.2571	0.3372	0.2775	0.2733
	UMLS TF 10	0.3167	0.2595	0.3295	0.2648	0.2613
	UMLS TF 15	0.3258	0.2706	0.3314	0.2673	0.2653
Arctan Formula (3.5)	UMLS TF 5	0.3652	0.2940	0.3775	0.3115	0.3057
	UMLS TF 10	0.3561	0.2858	0.3602	0.2884	0.2845
	UMLS TF 15	0.3500	0.2811	0.3563	0.2827	0.2821

Table A.23: UMLS FOG Re-Rank results for CLEF eHealth 2015 collection.

		FOG				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	UMLS TF 5	0.3152	0.2676	0.3284	0.2812	0.2774
	UMLS TF 10	0.3242	0.2622	0.3315	0.2684	0.2712
	UMLS TF 15	0.3258	0.2696	0.3356	0.2708	0.2764
Log Formula (3.4)	UMLS TF 5	0.3030	0.2421	0.3002	0.2611	0.2588
	UMLS TF 10	0.3061	0.2487	0.3025	0.2605	0.2581
	UMLS TF 15	0.2985	0.2435	0.3006	0.2580	0.2570
Arctan Formula (3.5)	UMLS TF 5	0.3106	0.2544	0.3217	0.2878	0.2816
	UMLS TF 10	0.2970	0.2467	0.3063	0.2717	0.2668
	UMLS TF 15	0.3045	0.2468	0.3049	0.2725	0.2677

Table A.24: UMLS Flesch-Kincaid Re-Rank results for CLEF eHealth 2015 collection.

		Flesch-Kincaid				
	Run	P@10	nDCG@10	RBP	uRBP	uRBPgr
	Baseline	0.3455	0.3027	0.3569	0.3148	0.3033
Basic Formula (3.3)	UMLS TF 5	0.3015	0.2402	0.3071	0.2586	0.2569
	UMLS TF 10	0.3136	0.2464	0.3150	0.2582	0.2605
	UMLS TF 15	0.3061	0.2447	0.3126	0.2553	0.2596
Log Formula (3.4)	UMLS TF 5	0.3076	0.2500	0.3186	0.2725	0.2708
	UMLS TF 10	0.3152	0.2522	0.3171	0.2660	0.2665
	UMLS TF 15	0.3061	0.2487	0.3171	0.2677	0.2689
Arctan Formula (3.5)	UMLS TF 5	0.3667	0.3222	0.3962	0.3428	0.3340
	UMLS TF 10	0.3652	0.3126	0.3829	0.3185	0.3173
	UMLS TF 15	0.3561	0.2960	0.3657	0.3071	0.3058