FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Health Suggestions: Moving Beyond the Beta Version

Paulo Miguel Pereira dos Santos



Master in Informatics and Computing Engineering

Supervisor: Carla Teixeira Lopes

July 23, 2019

© The present thesis has been developed within the project "NORTE-01-0145-FEDER-000016", financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

Health Suggestions: Moving Beyond the Beta Version

Paulo Miguel Pereira dos Santos

Master in Informatics and Computing Engineering

Approved in oral examination by the committee:

Chair: Alexandre Valle de Carvalho
External Examiner: Teresa Gonçalves
Supervisor: Carla Teixeira Lopes

July 23, 2019

Abstract

Nowadays, the Web is considered the main source of information worldwide, providing a large amount of health-related information accessible to users through search engines. Since the majority of users have low levels of health literacy, they may have difficulties in formulating appropriate queries or effectively select from the retrieved results, since a large amount of health-related information contains several medico-scientific concepts, and not everyone can understand them easily. Therefore, it becomes important to develop strategies that improve the relevance and understandability of the results that are retrieved to make consumers more informed.

To bridge the gap between the medico-scientific terminology and the understandability by lay users, an extension to Google Chrome has been previously developed, creating multilingual suggestions in lay and medico-scientific terminologies to help users find useful information while searching for health-related subjects. The main goal of this dissertation is to verify if it is possible to improve the queries that are suggested by this extension, so that they return more relevant and comprehensible results to the user. To do so, different strategies for generating suggestions based on the initial consumer query are proposed, evaluated and compared.

For this purpose, we used the MetaMap tool to identify the biomedical concepts in the query introduced by the user. With those concepts, different strategies were proposed, based on the Consumer Health Vocabulary (CHV) and the vocabularies included in the Unified Medical Language System (UMLS). Since the UMLS also provides several types of relationships between concepts, parent, child and same level relations were used to retrieve concepts associated with the ones from the initial query, in order to explore which types of relations would be interesting to explore.

Since it was necessary to generate suggestions in both lay and medico-scientific terminologies, and the concepts retrieved from the UMLS do not have this distinction, several methods to automatically classify those concepts in one of the terminologies were developed. These methods compute the cosine similarity between the expression to be classified and the lay and medical preferred strings in the CHV. The different methods were evaluated with a glossary that mapped lay to medical terms, as well as with a set of user queries about health topics, in both terminologies. The best method used all the expressions from CHV and a threshold, which represents the cosine similarity value from which the concept is considered to be lay.

With this method, we are able to distinguish terminologies and generate the respective suggestions. The methods that generate these suggestions were evaluated with an English and a Portuguese test collection, and the results show that the best method to generate suggestions, taking into account the relevance and understandability of the documents, is the one that retrieves the CHV-preferred expression as lay suggestion for each identified concept in the initial query, and the UMLS-preferred expression in the medico-scientific one.

Since the best method is also based on the Consumer Health Vocabulary, as the current version of the extension, the main conclusion that can be drawn, since the performance is relatively superior, is that the use of medical entity recognition is crucial to generate better suggestions. This method was then implemented in the extension.

Resumo

Atualmente, a Web é considerada a principal fonte de informação, disponibilizando uma grande quantidade desta relacionada com saúde acessível aos utilizadores através dos motores de pesquisa. Como a maioria dos utilizadores tem baixa literacia em saúde, estes podem ter dificuldades em formular *queries* apropriadas ou em selecionar da lista de resultados, já que muita informação de saúde contém vários conceitos médico-científicos, e nem todos conseguem compreendê-los facilmente. Portanto, torna-se importante desenvolver estratégias que melhorem a relevância e a compreensão dos resultados que são retornados para tornar os consumidores mais informados.

Para combater as diferenças entre a terminologia médico-científica e a compreensão por parte dos utilizadores leigos, uma extensão para o Google Chrome foi desenvolvida anteriormente, fornecendo sugestões multilíngue em terminologias leigas e médico-científicas para ajudar os utilizadores a encontrar informações úteis enquanto pesquisam assuntos sobre saúde. O principal objetivo desta dissertação é verificar se é possível melhorar as *queries* sugeridas por esta extensão, para que estas retornem resultados mais relevantes e compreensíveis para o utilizador. Para tal, diferentes estratégias para gerar sugestões são propostas, avaliadas e comparadas.

Para isso, utilizamos a ferramenta MetaMap para identificar os conceitos biomédicos na *query* introduzida pelo utilizador. Com esses conceitos, diferentes estratégias foram propostas, baseadas no Consumer Health Vocabulary (CHV) e nos vocabulários incluídos no Unified Medical Language System (UMLS). Como o UMLS também fornece vários tipos de relações entre conceitos, utilizaram-se relações pai, filho e do mesmo nível para recuperar conceitos associados aos da *query* inicial, para explorar quais seriam interessantes utilizar.

Como era necessário gerar sugestões nas terminologias leiga e médico-científica, e os conceitos recuperados pelo UMLS não têm essa distinção, vários métodos foram desenvolvidos para classificar automaticamente esses conceitos numa das terminologias anteriores. Estes métodos calculam a semelhança do cosseno entre a expressão a ser classificada e as expressões leigas e médicas preferidas do CHV. Os diferentes métodos foram avaliados com um glossário que mapeia termos leigos a termos médicos, assim como com um conjunto de queries de vários utilizadores em ambas as terminologias. O melhor método utiliza todas as expressões do CHV e um threshold, que representa o valor da semelhança do cosseno a partir do qual o conceito é considerado leigo.

Com esse método, somos capazes de distinguir terminologias e gerar as respectivas sugestões. Os métodos que geram essas sugestões foram avaliados com uma coleção de teste em inglês e outra em português, e os resultados mostram que o melhor método para gerar sugestões, tendo em consideração a relevância e compreensão dos documentos, é aquele que recupera a expressão CHV-preferred como sugestão leiga para cada conceito identificado na *query* inicial, e a expressão UMLS-preferred para sugestão médico-científica.

Como o melhor método também é baseado no CHV, como a versão atual da extensão, a principal conclusão que pode ser tirada, já que o desempenho é relativamente superior, é que o uso do reconhecimento de entidades médicas é crucial para gerar boas sugestões. Este método foi então implementado na extensão.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Carla Teixeira Lopes, for her support, motivation and guidance throughout this dissertation. Her promptitude to help in every situation, together with the share of her knowledge, facilitated the whole process.

I would like to acknowledge Professor João Pascoal Faria and Professor João Falcão e Cunha for helping me by funding my participation in the CISTI'2019 conference, where I presented a paper about part of the work carried out in this dissertation.

To my friends from FEUP, Sérgio, Cristiana, Tiago, Carolina, Ariana, Marta, Afonso, Beatriz e Rui, a big thank you for all the friendship, support and companionship throughout this entire journey. It was such a great experience, and all the good times will be remembered.

I am extremely grateful for all the love, support and encouragement that all my family gave me, especially my mom, for being a constant presence throughout the whole process and for continuously giving me love, strength and motivation. Everything I have accomplished I owe it to her.

Last, but certainly not least, I would like to dedicate this dissertation to my grandfather Arlindo, who unfortunately passed away during the process of realization of this dissertation. To him, who had always been an example of a hardworking and respected man, thank you for always being concerned and kind with me. I know this accomplishment will make him very proud.

Paulo Santos

"Today's accomplishments were yesterday's impossibilities."

Robert H. Schuller

Contents

1 Introduction **1**
- 1.1 Context . . . 1
- 1.2 Motivation and Goals . . . 2
- 1.3 Problem . . . 2
- 1.4 Solution . . . 3
- 1.5 Evaluation . . . 4
  - 1.5.1 Baseline . . . 4
  - 1.5.2 Test collections . . . 4
- 1.6 Contributions . . . 4
- 1.7 Dissertation structure . . . 5

2 Background and State of the Art **7**
- 2.1 Information retrieval . . . 7
- 2.2 Query formulation support . . . 8
  - 2.2.1 Query expansion in information retrieval . . . 9
  - 2.2.2 Query refinement in information retrieval . . . 10
  - 2.2.3 Query suggestion in information retrieval . . . 11
- 2.3 Query formulation support in health information retrieval . . . 13
- 2.4 Available Resources . . . 15
  - 2.4.1 Unified Medical Language System . . . 16
  - 2.4.2 Consumer Health Vocabularies . . . 18
- 2.5 Concept recognition in the health domain . . . 19
  - 2.5.1 Rule based . . . 20
  - 2.5.2 Dictionary based . . . 23
  - 2.5.3 Machine learning based . . . 23

3 Automatic classification of a concept in lay or medico-scientific terminology **25**
- 3.1 Methodology . . . 25
- 3.2 Evaluation . . . 28
  - 3.2.1 Multilingual Glossary . . . 28
  - 3.2.2 Queries submitted by users . . . 30
- 3.3 Conclusions . . . 30

4 Strategies to generate query suggestions **33**
- 4.1 Implementation details . . . 33
  - 4.1.1 MetaMap . . . 34
  - 4.1.2 Methods to generate query suggestions . . . 35

5 Evaluation strategy and Results **43**
5.1 Test collections . . . 43
5.1.1 English test collection - CLEF eHealth Lab 2018 . . . 43
5.1.2 Portuguese test collection . . . 45
5.2 MetaMap's Evaluation . . . 46
5.3 Methods' Evaluation . . . 47
5.3.1 Evaluation measures . . . 48
5.3.2 Results with the English test collection . . . 49
5.3.3 Results with the Portuguese test collection . . . 52
5.3.4 Results discussion . . . 55
5.4 Conclusions . . . 56

6 Implementation of the best strategy in the Health Suggestions extension **57**
6.1 Extension status before the update . . . 57
6.2 Extension status using the best strategy . . . 59

7 Conclusions and Future Work **61**
7.1 Conclusions . . . 61
7.2 Future work . . . 62

References **63**

A Queries used to evaluate the methods **69**
A.1 Queries from CLEF eHealth 2018 . . . 69
A.2 Queries from a past user study . . . 70

B Configuration files **75**
B.1 Elasticsearch indexing configuration file . . . 75
B.2 FSCrawler's configuration file . . . 76

List of Figures

2.1 Description of the IR process [CBB+13]. 8
2.2 Design of a information retrieval system based on text [CBB+13]. 9
2.3 Architecture of the suggestion tool proposed by Lopes and Ribeiro [LR16]. . . . 14
2.4 Suggestion engine architecture of Health Suggestions [LF16]. 15
2.5 Relation between the entities of the UMLS Metathesaurus. 17
2.6 System diagram showing how MetaMap works [ALA10]. 20
2.7 MetaMap's human-readable output for the input text "obstructive sleep apnea" [ALA10]. 22

3.1 Diagram showing the used vocabularies. 25

4.1 Diagram explaining the methodology followed to generate suggestions. 33
4.2 Results of running MetaMap for the query "emotional and mental disorders". Left image with default options and right image with Word-Sense Disambiguation Server running. 34
4.3 JSON containing the preferred atom for the CUI "C0020538". 37
4.4 JSON containing the list of atoms in different vocabularies for the CUI "C0020538". 38
4.5 JSON containing the list of relations for the atom "A21145030". 39
4.6 JSON containing the list of relations for the CUI "C0020538". 40

6.1 Suggestions' panel before the update. 58
6.2 Options page before the update. 58
6.3 Suggestions' panel before the update. 59
6.4 Options page before the update. 60

LIST OF FIGURES

List of Tables

2.1 Comparison of query suggestion systems 12
2.2 Example of entities from UMLS Metathesaurus [oMUM]. 17

3.1 Description of the different approaches. 26
3.2 Evaluation measures for each method in both languages. 29
3.3 Evaluation measures for the queries. 30

4.1 Description of the different methods. 35
4.2 Lay and medico-scientific suggestions given by the different methods for the query "caffeine high blood pressure". Each cell contains one suggestion. 41

5.1 MetaMap's performance using English and Portuguese queries. 47
5.2 Relevance evaluation of the methods using the English test collection. 50
5.3 Understandability evaluation of the methods using the English test collection. . . 51
5.4 Relevance and understandability evaluation combined of the methods using the English test collection. 51
5.5 Relevance evaluation of the methods using the Portuguese test collection. 53
5.6 Understandability evaluation of the methods using the Portuguese test collection. 53
5.7 Relevance and understandability evaluation combined of the methods using the Portuguese test collection. 54
5.8 Time used by the methods to generate suggestions using the 50 English queries. . 55
5.9 Time used by the methods to generate suggestions using the 200 Portuguese queries. 55

Abbreviations

ACC	Accuracy
API	Application Programming Interface
AUI	Atom Unique Identifier
BPref	Binary preference
CFD	Consumer-Friendly Display
CHD	Child relationship
CHV	Consumer Health Vocabulary
CRF	Conditional Random Field
CUI	Concept Unique Identifier
HTTP	Hypertext Transfer Protocol
IR	Information Retrieval
JSON	JavaScript Object Notation
LUI	Lexical Unique Identifier
MeSH	Medical Subject Headings
MM	MetaMap
nDCG@k	Normalized Discounted Cumulative Gain at k
NER	Named Entity Recognition
NLP	Natural Processing Language
PAR	Parent relationship
PAT	Patient-authored text
P@k	Precision at k
RB	Broader relationship
RBP	Rank-biased Precision
REST	Representational State Transfer
RL	Similar relationship
RN	Narrower relationship
ROCD	Receiver Operating Characteristic Distance
SEN	Sensitivity
SPC	Specificity
SOAP	Simple Object Access Protocol
SUI	String Unique Identifier
SVM	Support Vector Machine
SY	Designated synonym
UMLS	Unified Medical Language System
uRBP	Understandability-biased RBP
uRBPgr	uRBP graded
USMLE	United States Medical Licensing Examination
WSD	Word-sense Disambiguation
XML	Extensible Markup Language

Chapter 1

Introduction

The high technological development that has occurred in recent years has transformed the daily life of users, revolutionizing the way they search for information, including the one related to health care. With the growing amount of information on the Web, patients can easily learn about various topics, improving their health management. Although this sounds like an ideal situation, there are still some problems and difficulties that need to be addressed.

1.1 Context

With the evolving health needs of the population, there is a need to increase the availability and accessibility of health information [JAJ17].

Nowadays, the Web is considered the main source of information worldwide. Associated with this, search engines are commonly used by Internet users for seeking health information, which is considered the third most popular activity on the Internet [Art04].

A study conducted in 2013 in the United States of America (USA) stated that, from the 85% of USA adults that use the Internet, 72% searched online for several health information topics within the past year, including searches for general information, minor health problems and serious conditions. From this online health seekers, 77% say they began the hunt for health or medical information in search engines such as Google, Bing, or Yahoo! [FD12].

Another study, conducted in February and March 2018, states that 87% of USA people from 14 to 22 years old, search for health information online. These searches involve topics such as fitness (63%), nutrition (52%), stress (44%), anxiety (42%), and depression (39%) [RF18].

At the European Union (E.U). level, a similar study concluded that 60% of the Europeans search for health topics online. Of these, 75% say that the Internet is a reliable source for these searches [HSBD11], stating that 82% to 87% of health consumers start their search in search engines.

Despite the increasing use of the Web to search for health-related information and the availability of such information, some authors defend that there is a possibility that the Web is creating some inequalities in the access to health information [JAJ17]. As laypeople have low levels of health literacy, they struggle to meet their information needs because, on the Internet, there is a large amount of health-related information containing several medico-scientific concepts, and not everyone can understand them easily [ZKP+04]. This barrier could be minimized with the use of health technologies, that could help overcome the difficulties caused by this lack of health literacy by laypeople.

1.2 Motivation and Goals

Health information searches are particularly important to inform health consumers and encourage them to be more active in managing their health.

The health consumers believe that the Internet provides access to reliable content and that their searches are successful [HSBD11]. Despite this, recent studies show that there are problems with this type of searches. Lack of scientific vocabulary or lack of proficiency in a language strongly present on the Web might limit users' access to health information. The inaccurate use of keywords in a query also limit this access and can be assisted through query modification techniques such as query expansion, query suggestion and query refinement [OMQL15].

There is evidence that multilingual query suggestions in lay and medico-scientific terminologies improve health information retrieval by those who are not health professionals [LR16].

Health Suggestions, an extension for Google Chrome, was built to address the common difficulties that laypeople face when searching for health information on the Web, assisting them obtaining better results while searching for health-related topics [LF16]. Taking into account the user's original query, suggestions in English and Portuguese, using the lay and medico-scientific terminologies, are provided. The generation of suggestions is based on the Consumer Health Vocabulary (CHV), a vocabulary that connects the technical terms used by health professionals to a consumer-friendly language. The main goal of this dissertation is to explore and improve this extension, by providing different strategies to generate more accurate suggestions based on the initial user's input, focusing not only on the relevance of the retrieved results but also on how well these results are understood by the user. By doing so, it is expected that health consumers will become more autonomous in the management of their health.

1.3 Problem

Due to the existing differences between lay and medico-scientific terminologies, health consumers face difficulties when searching for health information on the Internet. The domain knowledge gap between these terminologies makes it hard for health consumers to formulate appropriate queries or effectively select from the retrieved results. This can cause poor information retrieval performance regarding health consumer searches.

The terminology issues are associated with another challenge: identifying what documents are good to fulfill the information need that the consumer has. To do so, the user needs to have a certain level of health literacy, in order to understand the retrieved content and select the relevant information. The language barrier can also limit the access to health information, since users that are not proficient in widely spread languages will not understand the retrieved information.

The query suggestion system previously defined tries to reduce the domain knowledge gap and language barrier by suggesting queries that are appropriate to lay and expert users, both in English and Portuguese. However, there are some issues associated with the method behind the generation of suggestions. Given a query introduced by a user, the extension maps the entire query to a single medical concept, which may cause important information in the query to be lost. Given this concept, it uses the Consumer Health Vocabulary to map the lay terms to medico-scientific ones, returning these as suggestions.

1.4 Solution

To improve the generation process of alternative queries, different strategies are proposed, evaluated and compared, and the best one is implemented in the extension. These strategies involve concept recognition and incorporation of information from other terminologies like the Unified Medical Language System (UMLS), which is a compendium of many controlled vocabularies in the biomedical sciences. It provides a mapping structure among these vocabularies, as well as ontologies about the concepts.

To develop the different strategies, several steps need to be followed. These are described below.

Firstly, considering the users' initial query, it is necessary to use concept recognition methods to identify the possible biomedical concepts introduced by the user. For concept recognition, and based on the research done in this area, it was concluded that using the MetaMap tool, described in Section 2.5.1.1, would be the best approach to follow.

Thus, for a given query, MetaMap is used to obtain the UMLS concept unique identifiers (CUIs) present on it, which represent different biomedical concepts.

After having the biomedical concepts from the user's query, different strategies make use of the UMLS vocabularies and the semantic relationships between the same concepts of the various vocabularies, in order to be able to construct an appropriate query suggestion. These strategies are described in detail in Chapter 4.

When a certain atom from the UMLS is used, and does not have the distinction to which terminology it belongs to, the computation of the degree of association of it with the lay or medico-scientific terminology is necessary, to later associate that atom to a lay or medico-scientific suggestion. This is done through vocabularies that distinguish concepts by terminology and with the computation of the cosine similarity between the expressions and the different vocabularies. This approach is described in detail in Chapter 3.

1.5 Evaluation

In order to evaluate the different strategies, the new generated query is used to retrieve documents from an English and Portuguese test collections. The strategies are evaluated taking into account the relevance of the documents and its understandability by lay users.

1.5.1 Baseline

The suggestions currently proposed by the Health Suggestions extension are considered as the baseline, which is a measurement of the process before any change occurs. This allows a comparison between the baseline and the approaches to verify if an improvement was accomplished.

For a given query introduced by the user, the extension identifies only one concept from the CHV and its corresponding CUI associated with the query and retrieves the CHV and UMLS preferred terms, both in English and in Portuguese.

1.5.2 Test collections

To assess the effectiveness of the approaches, two different collections are used: one in English and another in Portuguese.

The English test collection is the one provided by the CLEF eHealth Lab from 2018. This lab provides resources and evaluation methods to test and validate search systems. To a given query there are assessments of relevance and understandability for certain documents and, with that, it is possible to determine if the query suggestions are useful in improving the users' search for health information.

The Portuguese test collection was compiled within the scope of this dissertation, with the assessments of previously conducted user studies. For the existing documents of this collection, there are also assessments of relevance and understandability related with certain queries.

These test collections are fully described in Chapter 5.

1.6 Contributions

The main contributions of the developed work consist of a method to automatically classify a concept as lay or medico-scientific, as well as an extension for Google Chrome that provides suggestions in lay and medico-scientific terminology, both in English and Portuguese, given an initial query introduced by a user in a search engine.

The initial work to automatically classify a concept as lay or medico-scientific, at that time only considering the English language and without using the difficulty scores provided by the CHV, allowed the publication and presentation of the poster "Is it a lay or medico-scientific concept? Proposals for an automatic classification" in the 9th International Conference of the CLEF Association, CLEF 2018.

After some improvements in the previous work, which consisted in considering the difficulty scores and to enable the classification of concepts in Portuguese, it was published and presented the short paper "Is it a lay or medico-scientific concept? Automatic classification in two languages" in the 14th Iberian Conference on Information Systems and Technologies, CISTI 2019. This work is described in Chapter 3.

All the work that was developed in this dissertation, including the main results and conclusions for the suggestion of queries in lay and medico-scientific terminologies, in both English and Portuguese was described in the long paper "Suggesting lay and medico-scientific queries in two languages to improve health consumers comprehension" that will be submitted to the 42nd European Conference on Information Retrieval, ECIR 2020.

The compilation of the test collections used in the evaluation is also seen as a contribution that can be made available to the community, especially the Portuguese one, since it was created within the scope of this dissertation.

1.7 Dissertation structure

Besides the introduction, this dissertation contains five more chapters. Chapter 2 provides background information and related works about query formulation support, as well as techniques of concept recognition and the available resources necessary to implement the proposed solution. Chapter 3 presents an approach to automatically classify a concept as lay or medico-scientific, due to the necessity of identifying to which terminology a concept belongs when developing strategies to generate query suggestions. Chapter 4 describes all the implemented strategies to suggest queries in lay and medico-scientific terminologies, both in English and Portuguese. Having these suggestions in consideration, Chapter 5 presents the evaluation strategy followed to classify these suggestions, as well as the results of the evaluation measures. After the evaluation, the best method was implemented in the extension, and the details are described in Chapter 6. Lastly, Chapter 7 refers to the main conclusions that can be drawn from the implementation of these methods as well as the evaluation of results, and also refers to the future work that can be done to improve them.

Chapter 2

Background and State of the Art

This chapter provides a basis for what will be developed. In Section 2.1, the concept of Information Retrieval is introduced. Section 2.2 introduces techniques of query formulation support in Information Retrieval, as well as methods that systems use to implement them. Section 2.3 restricts the previous methods to those existing in the health domain and Section 2.4 describes the available resources that can be used to carry out this work. Finally, Section 2.5 presents techniques of Concept Recognition.

2.1 Information retrieval

The meaning of the term information retrieval can be very broad, but as an academic field of study it can be defined as:

> "Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)." [Man09]

Nowadays, information retrieval is becoming the dominant form of information access, overtaking traditional database-style searching [Man09].

IR is also used to facilitate "semistructured" search, like finding documents with certain titles or whose body contains certain words.

Figure 2.1 illustrates the information retrieval process. Here, the user enters a query *q* from a Web browser, for example; then, this query is transformed into a "machine-readable" query *q'* by the *query and result interaction* module. The new query *q'* is passed to the *search and query analysis module*, responsible for accessing the *content management* module directly, which is linked to the back-end. The search module is responsible for preparing a set of results *r*, returning it to the user via the result interaction module. If the user is not completely satisfied, this set of results can be modified or updated into a new one (*r'*) [CBB+13].

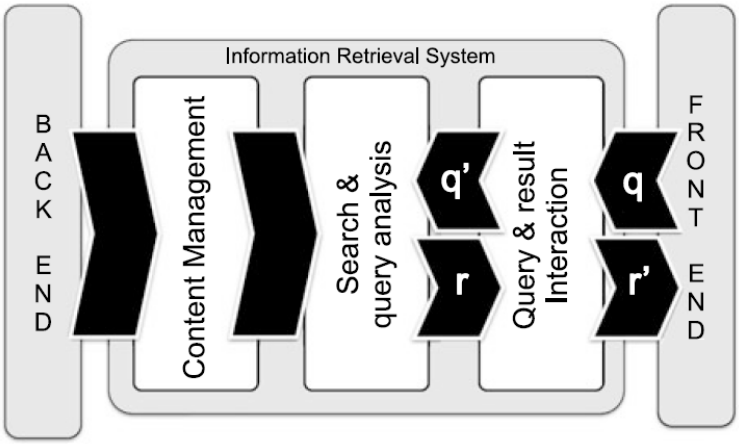


Figure 2.1: Description of the IR process [CBB+13].

Since most of the applications of IR deals with documents expressed in natural language, several textual operations might be necessary, in addition to the previous steps.

Figure 2.2 illustrates the steps that are usually followed by an IR engine to process textual queries. Via the user interface, the user expresses its need using a textual query q_U. This query q_U is then parsed and transformed by a set of textual operations, redefining the initial query to q'_U. This query is then transformed into a system-level representation, q_S, by a set of query operations. The query q_S is executed on top of a document collection D (e.g., a text database) to retrieve a set of relevant documents, R. The previous step assumes the existence of an index for the documents, thus allowing a fast query processing. The retrieved documents R are ranked taking into account the relevance that they have to the user's need. Finally, the user examines R for useful information [CBB+13].

Despite the benefits that an information retrieval system can bring to its users by facilitating access to information, there are still some problems that make them not considered to be fully efficient. The ineffectiveness of these systems is often caused by the inaccurate use of keywords in a query.

To achieve the best results possible when retrieving information from an IR system it may be necessary to use very specific keywords, which can be a problem when users are not sure about the nature of the content they need. To surpass this problem, several techniques for query formulation are used, such as query expansion, query refinement and query suggestion [OMQL15].

2.2 Query formulation support

As previously stated, in order to retrieve better results from an information retrieval system it is often necessary to use precise keywords from multiple fields. When users are trying to express their information need, they might use keywords that are too general or different than the ones

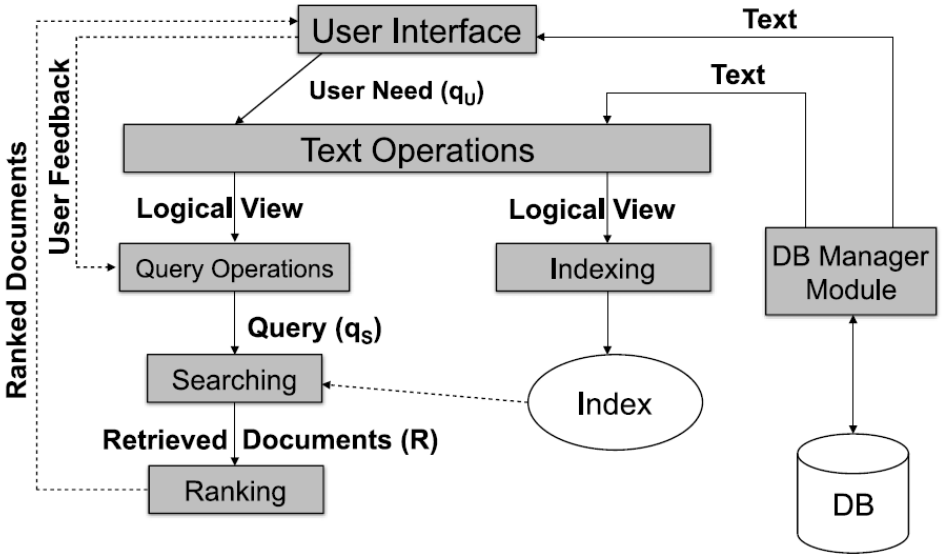


Figure 2.2: Design of a information retrieval system based on text [CBB+13].

included in documents, as well as an insufficient number of terms, making the query difficult to "be understood" by the system [EN16].

To solve this problem, several techniques have been proposed, including query expansion, query refinement as well as query suggestion. These approaches aim to improve the relevance and comprehension of the retrieved documents, taking into account the user's information need, by modifying the query. These modifications may consist of changes in the existing terms as well as the addition of new ones.

2.2.1 Query expansion in information retrieval

Some researchers firmly believe that queries formed by a small number of keywords are insufficient to express the user's information needs, therefore one of the main causes of the bad performance in information retrieval systems [CRB12].

The query expansion technique consists in adding meaningful terms to the initial query, causing it to express the information need in a more detailed way and removing possible ambiguities. By adding these terms, the possibility of matching with relevant documents is higher.

The first methods that addressed query expansion were based on the clustering of terms. Researchers believed that pairs of words that occur together in documents were usually associated with the same concept. Taking this into account, the terms were mapped into clusters, using terms from the same cluster to expand the original query [PW91].

Other techniques use global or local analysis of document collections in order to detect word relationships and use that to provide new terms. Vechtomova et al. [Vec03] proposed two approaches for query expansion with long-span collocates (combinations of two or more words that

normally occur within the same language and significantly co-occur with query terms). The global approach considered the entire document collection while the local analysis used a subset of the retrieved documents. The results show that the terms collected globally are too general and give worst results than the ones from the local analysis.

Using relevance information for query expansion is another approach followed by many researchers. In this technique, users are expected to make judgments about the retrieved documents and this information is then used to re-weight query terms or to expand the query with the most important terms from relevant documents. From these documents, Buckley et al. [BSAS] applied pseudo-relevance feedback by only considering the first initially retrieved documents as relevant, and selected the most frequent terms from those documents. The results show that this method is efficient on average, but for poor performing queries it is not, since it is more likely that the first documents are not relevant.

Cao et al. defended that many expansion terms identified with pseudo-relevance feedback methods may be unrelated to the query and give worst retrieval results. They also stated that to select good expansion terms it is not possible to rely only on the feedback documents and in the whole collection, because it is not possible to distinguish which ones are good or bad [CNR08].

To address this problem, they focused on the effectiveness of term proximity to select expansion terms. To select these terms, the authors created a term classification method based on Support Vector Machines (SVMs), where they predict the usefulness of candidates based on the term distribution, co-occurrence with query terms and the proximity from them. The results show that the retrieval effectiveness is improved when term classification is used.

More recently, Silva and Lopes [SL16] used different sources, like Wikipedia articles and definitions from the UMLS Metathesaurus, and methods like pseudo-relevance feedback to select the terms to be added to the original query.

2.2.2 Query refinement in information retrieval

Query refinement is also a technique for query formulation that aims to automatically transform the original query into a new one that represents the user's information need with higher accuracy, without adding new relevant terms [VWSG97]. Velez et al. [VWSG97] developed an algorithm called RMAP that uses new terms that are semantically related with each query term and combines them, refining the entire query.

Kraft and Zien proposed a method that uses anchor text from a large hypertext document collection as the source to the new related terms for a given query [KZ04]. The results show that the usage of the anchor texts produces refinements that return significantly better documents in terms of comprehension compared to refinements produced using only the document content.

Shen et al. [STZ05] combined previous user's queries and click-through information to create an algorithm that refined the initial query. The results show that the pseudo-relevance feedback improved the retrieval performance of the queries.

The most common approaches are based on replacing the terms in the original query by synonyms or to paraphrase the query. Zhu et al. stated that these approaches were not sufficient since

they were missing out the context of the queries by looking for each term individually. Therefore, they proposed a phrase-rewriting scheme by exploiting multiples online search engines, since they defend that with the use of phrases, the context was being preserved and outperformed the baselines [ZLXX16].

2.2.3 Query suggestion in information retrieval

To help in the formulation of queries in an information retrieval system, query suggestions can also be used. This technique suggests different alternatives to the initial query submitted by the user. The main goal of using such alternatives is to reformulate the initial query with a single click to allow an improvement in the efficiency of the user's web search. This can be done using simple modifications, such as adding or deleting terms, as well as replacing terms by synonyms or other related ones [JZH09]. According to previous studies, this is the preferred way for users to improve their original query [Men14], and differs from the previous techniques since it allows the users to choose an alternative for the initial query from a list of suggestions.

Normally, searching for information is an iterative process. Within the same search session, it is normal for users to formulate more than one query until they have satisfied their information needs. To help users on this process, the system can provide alternative suggestions, allowing the exploration of less familiar topics [KGB09].

Several implemented techniques that generate query suggestions take into account the user's initial query. Some of them are based on the user's click pattern, while others have the session as a base to generate the suggestions.

Some methods save the click patterns of each user in a search log, for each entered query. These patterns are later exploited to classify the queries into clusters, using those same clusters as the basis for generating the new suggestions, since it is assumed that queries from the same cluster are similar [OMQL15].

Leung, Ng and Lee have developed a method that provides suggestions based on user conceptual preferences. For this, two strategies were followed: one for the extraction of concepts from the results provided for a given query, using these concepts to suggest similar queries; another that uses data from the click patterns, creating clusters for them, thus understanding the conceptual preferences of the user, providing suggestions from them [LND08].

On the other hand, session-based methods assume that in the same search session, the queries submitted are related to each other in some way. This is due to the fact that users need to reformulate the query to meet their information need, thus creating a high number of similar queries that can be used as a basis in these systems [OMQL15].

Query logs, associated with information such as word translation relationships and word co-occurrence statistics that provide a cross-lingual query similarity value, were used to propose suggestions in a language different from the original one by Gao et al [GNN+10]. The authors concluded that combining these suggestions with pseudo-relevance feedback improved the effectiveness of cross-language information retrieval.

The work in the area of query suggestion use as a main method the suggestion of queries that are semantically related to the query that the user initially introduced. Because the user may have multiple search intentions, these strategies may not be able to handle queries with uncertain search aspects. To alleviate this problem, diversification has been injected into query suggestion systems with a probabilistic model or with bipartite graphs while personalization is often incorporated into query suggestion system by mining a user's past behavior [CCCdR18]. Taking this into account, the authors proposed a model that is based on the diversification and personalization of the suggestions. Diversification helps to generate queries that encompasses multiple topics, thereby increasing the likelihood of the suggested queries being clicked, while personalization ensures that the queries meet the user's search intent.

Without using query logs, Bhatia et al. [BMM11] developed a probabilistic mechanism for generating query suggestions which extracts candidate phrases of suggested queries from document corpus.

As can be seen, most of the methods responsible for suggesting queries use information existing in documents or in search logs, not taking into account the semantic relationships between the original query and the proposed suggestions. Besides that, existing query suggestion diversifying methods generally use a greedy algorithm, which has high complexity. To address this issue, Zheng et al. [ZZZ+14] proposed a method that generates semantically relevant queries using the WordNet ontology, evaluating it on a large-scale search log dataset of a commercial search engine.

Table 2.1 represents summarized information of the methods and data types that were used to develop the major query suggestion systems outlined above.

Table 2.1: Comparison of query suggestion systems

Authors	**Methods**	**Data**
Leung, Ng and Lee [LND08]	Suggestions based on user's conceptual preferences	Search logs with user's click patterns
Gao et al. [GNN+10]	Cross-lingual query similarity to generate multilingual suggestions	Query logs, word translation relationships and word co-ocurrence statistics
Chen et al. [CCCdR18]	Diversification and personalisation of suggestions	User's past behavior
Bhatia et al. [BMM11]	Probabilistic mechanism to generate query suggestions	Document corpus
Zheng et al. [ZZZ+14]	Semantically relevant queries	WordNet ontology

Machine learning algorithms are now being used in semantic query suggestions, using approaches based on Random Forests. This approach was compared with conventional classification algorithms and ensemble learning methods and the experimental results proved the superiority of the proposed scheme [Ona19].

2.3 Query formulation support in health information retrieval

Given the great difference between the terminologies used by health professionals and laypeople, different query expansion approaches have been proposed to overcome the difficulties that can emerge [Zie03]. Typically, information retrieval systems are evaluated taking into account the topical relevance of the retrieved results. However, in the health domain, it is also important to evaluate the knowledge obtained regarding its medical accuracy, which is a feature that is not directly associated with precision and recall.

Zeng et al. developed the Health Information Query Assistant (HIQuA) system, suggesting alternative or additional terms to the query, taking into account the semantic distance they have to the initial query. This distance is calculated using logs and the co-occurrence of concepts in medical documents, as well as the semantic relationships existing in medical vocabularies. The evaluation was made by conducting a study, composed of 213 individuals. These were divided into two groups, one that received suggestions and one that did not. This allowed the authors to conclude that recommendations based on semantic distance help consumers in the health information retrieval process by helping them formulate the queries [ZT06].

Liu and Wesley [LC07] proposed a query expansion method based on knowledge sources, exploiting the UMLS knowledge source, appending additional relevant terms to the original query. To evaluate the proposed method, this was compared with statistical expansion techniques, which simply uses terms that may not be relevant to the initial context of the query. The results in the scenario-specific expansion are 5% better than the ones with the statistical method.

An intelligent medical Web search engine called iMed was built by Luo and Tang to help users formulate queries based on medical knowledge and an interactive questionnaire. This system suggests symptoms based on the query introduced, recommends alternative answers taking into account their similarity of being the correct answers and suggests medical phrases that help to reformulate the query from the Medical Subject Headings (MeSH). This search engine was evaluated using real medical case records and medical questions from the United States Medical Licensing Examination (USMLE) exam and demonstrated effectiveness [ZCP+06].

When a medical information searcher is uncertain about the exact questions, he prefers to pose long queries describing his symptoms and expects to receive relevant information from search results [LGT+08]. MedSearch, also a medical Web search engine, addresses these challenges. This search engine allows users to enter long queries expressing their information needs, turning them into smaller queries containing only the words that are effectively important and representative of the user's need, returning diversified search results. Like the previous work, it uses MeSH to suggest medical phrases that help the user improve his query. In order to evaluate the search engine, the authors used medical questions, showing that this tool was able to handle comprehensive medical queries in an effective and efficient way [LGT+08].

MeSH, used in conjunction with social tagging, served as the basis for the creation of a search system developed by Zarro and Lin that provides consumers with terms from lay and medico-scientific terminologies so that they can adapt and expand the initial queries in order to explore

less familiar concepts in the health domain [ZL11].

To bridge the gap between the vocabularies that health professionals and laypeople use, Soldaini et al. investigated the utility of adding the most appropriate professional expression to queries initially submitted by users. To obtain this expression, they used three synonyms mappings and conducted two studies in which they asked users to answer health-related questions using interleaved results from a well-known search engine. The results of this evaluation led to the conclusion that the system was preferred by users as it helped them to answer correctly to the proposed medical questions [SYYT+16].

Stanton et al. studied how the use of more words than necessary to describe a health information need influenced the accuracy of the retrieved documents. Given a free-form colloquial health search query, the authors try to find the underlying professional medical term by using machine-learning algorithms [SIM14].

A query suggestion system was developed by Lopes and Ribeiro, combining the suggestion of multilingual alternatives (in Portuguese and English) with the use of lay and medico-scientific terminology as shown in Figure 2.3. They used the CHV that maps technical terms to consumer-friendly language. For each query, they identify the associated concept and then return its CHV and UMLS preferred names in English and Portuguese [LR16].

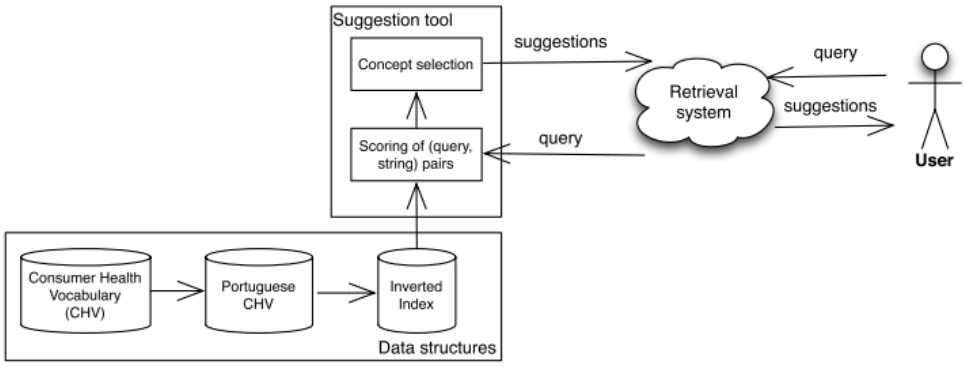


Figure 2.3: Architecture of the suggestion tool proposed by Lopes and Ribeiro [LR16].

A study with 40 subjects was conducted to evaluate how the language and terminology had an impact on the use of the suggestions provided by the previous system. For this, the levels of English proficiency, health literacy and topical familiarity were considered for each subject. The results of this study showed that users tend to prefer the suggestions in English, that the medico-scientific suggestions tend to be preferred by those who have higher levels of health literacy and that the suggestions are most used at the beginning of the search sessions, when users are less aware of their information needs [LR16].

Using the same suggestion tool described before, the same authors together with a medical doctor studied the effects that language and terminology of query suggestions have on medical accuracy, considering different user characteristics, by evaluating the system's impact on the medical accuracy of the knowledge acquired during the search [TPR17]. This evaluation showed that

the provided suggestions were important to reduce the number of incorrect contents. Even though some suggestions might not be clicked, they are useful either for subsequent queries' formulation or for interpreting search results. They also concluded that clicking on suggestions leads to answers with more correct content and that the benefits of certain languages and terminologies are more perceptible in users with certain levels of English proficiency and health literacy.

Lopes and Ribeiro also analyzed how the previous suggestion tools affected the precision of search sessions [LR18]. They conducted a user study with 40 participants and the results showed that a system providing these suggestions tends to perform better than a system without them and that on specific groups of users, clicking on suggestions has positive effects on precision while using them as sources of new terms has the opposite effect.

With the same objective as the previous works, Lopes and Fernandes created HealthSuggestions, an extension for Google Chrome to assist users obtaining high-quality search results in the health domain [LF16]. The implementation of the system as an extension for Google Chrome was motivated by the fact of reaching all the main search engines and due to its current dominance as users' preferred browser. It is composed by two distinct modules: suggestion engine and logging engine.

The suggestion engine is responsible for generating the suggestions, using the Consumer Health Vocabulary. The architecture is very similar to the previous work [LR16], using inverted indexes to associate each stemmed term to an inverse string frequency and a postings lists, that is, the list of CHV strings where it appears. To generate the suggestions, the steps presented in Figure 2.4 were followed.

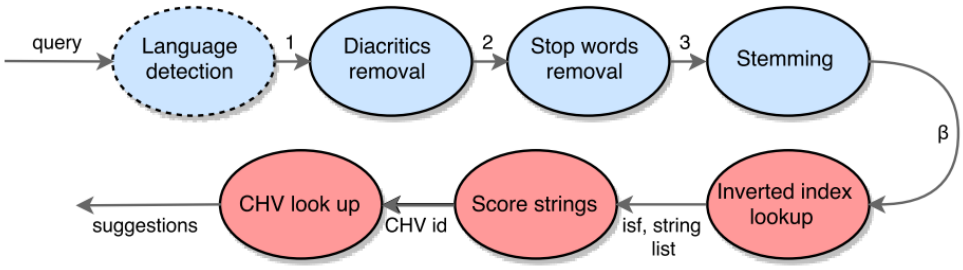


Figure 2.4: Suggestion engine architecture of Health Suggestions [LF16].

The logging system is important to understand users' search process and to improve the extension, by tracking most of users' actions while performing health-related searches.

2.4 Available Resources

Having knowledge bases, like databases and dictionaries, make the process of query suggestions much simpler, allowing these systems to behave as if they "understand" the meaning of the language of biomedicine and health.

2.4.1 Unified Medical Language System

Unified Medical Language System is a knowledge base that aggregates multiple dictionaries and vocabularies of the medical domain, as well as software tools to be used by system developers [Bod04]. The main goal of this knowledge base is to facilitate the development of computer systems that behave as if they "understand" the language of biomedicine and health. It is composed by three main knowledge sources: Metathesaurus, Semantic Network and SPECIALIST Lexicon.

2.4.1.1 Metathesaurus

The Metathesaurus consists of a large database of health-related, multipurpose and multilingual vocabularies. These vocabularies are composed of concepts related to health, their various names and the relationships that exist between them. It holds over 1 million biomedical concepts, 5 million concept names and nearly 200 controlled vocabularies and classification systems. This database was constructed from different thesauri, classifications, code sets and lists of controlled terms that are used in the health domain. The main objective is to provide, in a centralized way, different names and views for the same concept, as well as existing relations between different concepts, since each of these concepts is associated with at least one semantic type of the semantic network [oMUM].

The Metathesaurus is organized by concept, and the main goal is to relate the different names that exist in the various vocabularies to the same concept. Each concept and the concept names it contains have several types of unique, permanent identifiers.

A **concept** is a meaning, which in turn can have several different names. The idea is to associate all the names from the different vocabularies that mean the same thing with the concept that represents their meaning. Each concept has a unique and permanent concept identifier (CUI).

Each unique name associated with a concept (**string**) in each language of Metathesaurus has a unique and permanent string identifier (SUI). Any variation on the character set, upper-lower case, punctuation or language implies a new SUI. If a string has more than one meaning, it will be associated with different concepts (CUIs).

The **atoms** are the basis for the construction of the Metathesaurus, representing the concept names or strings for each of the source vocabularies. Every occurrence of a string in each source vocabulary has a unique atom identifier (AUI). For example, if the same string appears twice in the same vocabulary, one containing the long name and other the short name for the same concept, a unique AUI is assigned for each occurrence. The same string in different vocabularies also implies different AUIs. All these identifiers are associated with only one SUI, since all the occurrences represent different variations of the same string. Opposed to SUIs, an AUI is always linked to a single CUI, since an atom can only have one meaning.

Only considering the strings in the English language from the Metathesaurus, they are linked to all of its lexical variants or minor variations through a common term identifier (LUI). Like the SUI, the LUI can be linked to more than one concept. In contrast, each string identifier and each atom identifier can only be linked to a single LUI.

The relations between the entities described above are detailed in Figure 2.5. This diagram should not be seen as an integrated conceptual model since it serves to show relationships individually.

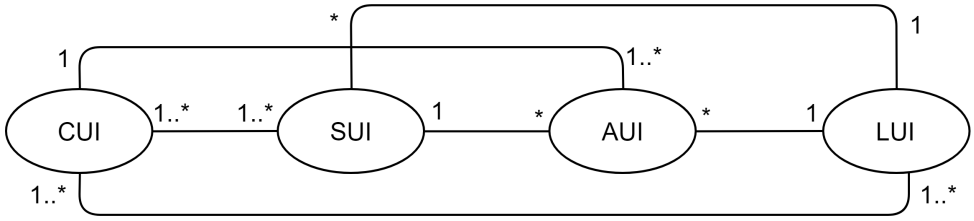


Figure 2.5: Relation between the entities of the UMLS Metathesaurus.

The example represented in Table 2.2 is a case where "Atrial Fibrillation" appears as an atom in two different source vocabularies. Therefore, different identifiers (AUI) are assigned to each one. Since they have an identical string or concept name, they are linked to a single SUI. "Atrial Fibrillations", the plural of "Atrial Fibrillation", has a different string identifier. Since these two are lexical variants of each other, both are linked to the same LUI. There is a different LUI and different SUIs and AUIs for "Auricular Fibrillation" and its plural. Since Atrial Fibrillation and Auricular Fibrillation have been judged to have the same meaning, they are linked to the same CUI.

Table 2.2: Example of entities from UMLS Metathesaurus [oMUM].

Concept (CUI)	**Terms (LUIs)**	**Strings (SUIs)**	**Atoms (AUIs)**
C0004238 Atrial Fibrillation (preferred) Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	**L0004238** Atrial Fibrillation (preferred) Atrial Fibrillations	**S0016668** Atrial Fibrillation (preferred)	**A0027665** Atrial Fibrillation (from MSH) **A0027667** Atrial Fibrillation (from PSY)
		S0016669 (plural variant) Atrial Fibrillations	**A0027668** Atrial Fibrillations (from MSH)
	L0004327 (synonym) Auricular Fibrillation Auricular Fibrillations	**S0016899** Auricular Fibrillation (preferred)	**A0027930** Auricular Fibrillation (from PSY)
		S0016900 (plural variant) Auricular Fibrillations	**A0027932** Auricular Fibrillations (from MSH)

The CUIs are the entities that gather all information in the Metathesaurus related to a certain concept, containing concept names, relationships and attributes. As can be seen in Table 2.2, one string from the concept is designated as the default preferred name for that concept.

Besides the synonymous relationships described above, the Metathesaurus also include other relationships between different concepts. These relationships can be between concepts from the

same vocabulary (**intra-source vocabulary relationships**) or between concepts from different vocabularies (**inter-source vocabulary relationships**), that are closely related. Examples of intra-source relationships are immediate parents/child/sibling relationships. The primary inter-source relationships in the Metathesaurus are the synonymous relationships represented in the Metathesaurus concept structure, as well as "like" or "similar" relationships between concepts from different vocabularies.

All relationships have a general label (REL), describing their basic nature, such as "Broader", "Narrower", "Child of", among others. About a quarter of the relationships also carry an additional label (RELA) that explains the nature of the relationship more exactly, such as "is_a", "branch_of", "component_of", among others.

2.4.1.2 Semantic Network

The main goal of the Semantic Network is to provide consistent categories to which the concepts represented in the above-mentioned vocabularies can be associated with. In addition, it also provides useful relationships between the various concepts. This network contains about 135 semantic types and 54 relationships [oMUM].

The categories, also known as Semantic Types, are defined in the network, containing textual descriptions and information inherent in hierarchies. These semantic types are the nodes in the network and are assigned to concepts. The Semantic Relations are the links of the network between the types.

2.4.1.3 SPECIALIST Lexicon and Lexical Tools

The SPECIALIST Lexicon provides the lexical information needed for the SPECIALIST Natural Language Processing System (NLP). It is intended to be a general English lexicon that includes many biomedical terms.

The UMLS contains some errors and inconsistencies, like ambiguity and redundancy, since the same concepts can be defined in multiple vocabularies, not consistent relationships and lack of ancestors [GMX+09].

The content of UMLS can be accessed in several ways. One of them is the Metathesaurus browser, a web interface that allows users to query the different vocabularies and search specific concepts, navigate through all the sources and retrieve concepts, semantic types and relations between other concepts. Another way of accessing is by downloading the entire UMLS databases. Lastly, UMLS can be accessed through HTTP APIs, one using REST and another using SOAP.

2.4.2 Consumer Health Vocabularies

Health consumers have difficulties finding and understanding health information, since a lot of this information requires health literacy. To improve patient's comprehension, it is becoming increasingly important to develop consumer health vocabularies mapping the medico-scientific and lay terminologies or, at least, to predict the lay-friendliness of medical concepts. Ideally, these

vocabularies should reflect the different ways health consumers express and think about health topics, helping to bridge this vocabulary gap [ZT06].

As stated before, several term identification methods are explored in the literature. Some of these methods include collaborative human review and automatic term recognition methods, stating that these were effective for identifying terms for the Consumer Health Vocabulary [ZTD+07].

One of the first approaches identified Consumer-Friendly Display (CFD) names by analyzing queries from a specialized site (MedlinePlus) in health information directed for laypersons [ZTC+05].

Another alternative used Wikipedia as a basis to identify pairs of lay and professional terms using a pattern-based text-mining approach. Since Wikipedia has a large text corpus and is maintained by the community, it is useful for this approach because community-generated text is a valuable resource for extracting consumer health vocabulary [VMHZ14].

A similar approach applies a Conditional Random Field (CRF) model, named ADEPT [MH13], to recognize patient-authored text (PAT) from text written by consumers in medical forums. This approach seems to have better results when compared to existent solutions.

The Consumer Health Vocabulary (CHV)[1] began to be built in 2006 and connects informal, common words and phrases about health to technical terms used by health care professionals, containing information about the concept and if it is a CHV-preferred or UMLS-preferred concept. Currently, the CHV is a part of the UMLS. Since it includes jargon, slang, ambiguous, and misspelled words as used by consumers and health care professionals, these are concepts that are not represented in other source vocabularies within the Metathesaurus.

To improve the readability of health texts, it is essential to be able to compute the difficulty of medical concepts. To do that, a model to predict the familiarity level of CHV strings among health consumers was developed [KTC+07], based on surveys completed by 52 consumers for assessing surface-level familiarity and concept-level familiarity. These scores are part of the CHV and are used to estimate the difficulty of a term. The *Frequency score* gives a measure of how likely it is that an average reader will be familiar with or understand a given term. *Context score* estimates the difficulty considering the context of the term while the *CUI score* determines how closely related the concept is to known examples of easy and difficult concepts. *Combo score* is a combination of frequency, context and CUI scores and *Combo score no "top words"* is a slight modification to *Combo score* that ignores top words, i.e., words that are easy to understand by laypeople.

Medline Plus[2] is also a vocabulary with lay terms which contains health information about more than 750 topics including current news and related information.

2.5 Concept recognition in the health domain

To provide consistent query suggestions it is necessary to apply techniques of concept recognition in the medical domain.

[1] https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV/
[2] https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MEDLINEPLUS/

Natural Processing Language (NLP) is the field with roots from computer science, artificial intelligence and computational linguistics concerned with human language. Tasks such as information retrieval, sentiment analysis, machine translation or Named Entity Recognition (NER) are commonly researched in NLP.

The last one (NER) identifies parts of unstructured text that relate to specific concepts of interest. To do that, it is important to properly prepare the data to be processed, by breaking the text into phrases, then splitting the phrases into its meaningful units (tokens), reducing the token to its stem or root and remove the words that are common to the language.

NER can be done based on rules, dictionary matching or machine learning.

2.5.1 Rule based

These systems rely on domain experts to define complex rules, based on orthographic characteristics combined with word syntactic and semantic properties. Since the rules are usually too specific, the overall performance when applied to a different context drops significantly.

2.5.1.1 MetaMap - a tool for recognizing UMLS concepts in text

MetaMap is a rule based system, mapping texts from the health domain to the UMLS Metathesaurus, or, equivalently, to discover Metathesaurus concepts referred to in free text [Aro01].

A system diagram showing how the tool identifies the concepts in the text and retrieves the corresponding UMLS concepts is represented in Figure 2.6.

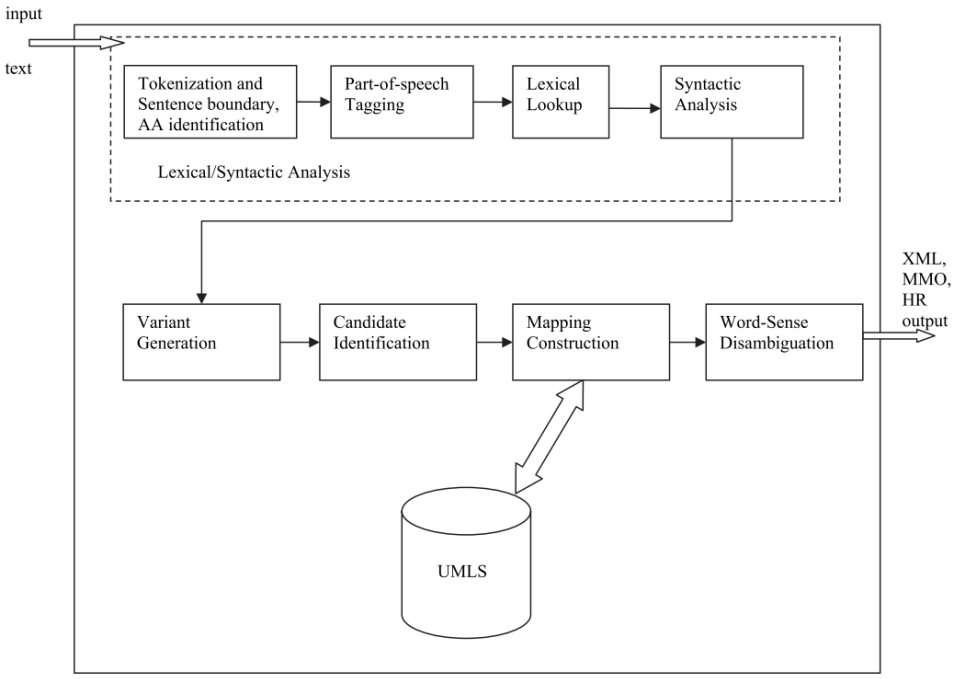


Figure 2.6: System diagram showing how MetaMap works [ALA10].

Firstly, the input text goes through a process of lexical and syntactic analysis, with several steps [ALA10], including:

- Tokenization, in which the input text is split into several tokens representing the words existing in the input; Sentence boundary, in which the sentences existing in the input are identified; Acronym/Abbreviation identification.
- Part-of-speech tagging, that consists in marking up a word as corresponding to a particular part of speech, e.g., as a noun, verb, adjective, among others.
- Use of the SPECIALIST Lexicon to do a lexical lookup of input words.
- Final syntactic analysis in which phrases and their lexical heads are identified.

The phrases identified in the previous steps are then analyzed by the following process:

- Variant generation, in which, by doing a table lookup, variants of all phrases' words are determined.
- Candidate identification, in which strings from Metathesaurus (candidates) that match some phrase text are computed and evaluated as to how well they match the input text.
- Mapping construction, combining and evaluating the candidates from the previous step with the information from UMLS to produce a final result that best matches the phrase text.
- Word-sense disambiguation, an optional step, used to deal with ambiguity in certain concepts. To surpass this ambiguity, concepts that are semantically consistent with surrounding text are favored.

To evaluate the candidates and the final mappings comparing to the input text, as stated above, the tool linearly combines four linguistically measures: centrality; variation; coverage; and cohesiveness [ALA10].

- Centrality consists in verifying if the linguistic head of the input text is associated with any of the candidates. It is a Boolean value, one if it is associated, zero if not.
- Variation is a measure representing the average of the variation between all the words existing in the input text and the words from the candidates.
- Coverage measures how much of the input text is present in the corresponding mapping.
- Cohesiveness measures how many chunks of contiguous text exist in the mapping.

The result of this evaluation is normalized to a value between 0 and 1000, in which coverage and cohesiveness have twice the weight of centrality and variation.

The MetaMap tool allows the user to create his configurations, according to his preferences, allowing to change dimensions such as:

- The vocabularies from the UMLS that will be used, as well as the source version.
- Output options, which determine the format of the output generated, and can be Extensible Markup Language (XML), Human Readable Output as well as MetaMap Machine Output.
- Processing options, which control the algorithmic computations to be performed.

For each phrase that MetaMap is able to identify in an input text, it produces a human-readable output as illustrated in Figure 2.7. This output is normally formed by three parts [ALA10]:

- The identified phrase.
- A list of candidates, composed by strings from UMLS that match some or all of the input text. If the preferred name for that concept differs from the candidate, it is shown in parentheses, as well as its semantic type.
- The final mappings, corresponding to different combinations of candidates that match as much of the phrase as possible.

```
Phrase: "obstructive sleep apnea"
Meta Candidates (11):
  1000 Obstructive sleep apnoea (Sleep Apnea, Obstructive) [Disease or
       Syndrome]
   901 Apnea, Sleep (Sleep Apnea Syndromes) [Disease or Syndrome]
   827 APNOEA (Apnea) [Pathologic Function]
   827 Sleep [Organism Function]
   827 Obstructive (Obstructed) [Functional Concept]
   827 Apnea (Apnea Adverse Event) [Finding]
   793 Sleeping (Asleep) [Finding]
   755 Obstruction [Individual Behavior,Pathologic Function]
   755 Sleepy [Finding]
   755 Sleeplessness [Sign or Symptom]
   755 Obstruction (Obstruction within Medical Device) [Phenomenon or
       Process]
Meta Mapping (1000):
  1000 Obstructive sleep apnoea (Sleep Apnea, Obstructive) [Disease or
       Syndrome]
```

Figure 2.7: MetaMap's human-readable output for the input text "obstructive sleep apnea" [ALA10].

In Figure 2.7 it is possible to see that the tool identifies 11 Metathesaurus candidates, the best of which received the maximum score of 1000, and by itself formed the top-scoring mapping. This is the preferred way for a human to analyze the mappings to concepts from input texts, although more structured formats exist, so that the results could be analyzed in a more automatic way. In

addition to the information that is shown in the figure, it is possible to obtain the CUI of the identified concepts, their semantic types, among others, in order to gain more information.

The main advantage of the tool is its thoroughness, characterized by its aggressive generation of word variants, the fact that lexical and syntactic analyses follow a linguistically approach as well as the metrics used for the evaluation of mappings. It also has the capability of constructing compound mappings when a single concept is insufficient to characterize input text phrases. Since it is highly configurable, the user can easily adapt it depending on the task that is being addressed [ALA10].

On the other hand, it only supports English texts, since all its implementation is based on that language. Its thoroughness can also be seen as a disadvantage, since it can take quite a long time to analyze longer inputs.

This tool can be used interactively through an available online platform, where users can easily configure several options. A Web API is also available, as well as a Java API, allowing developers to easily include this tool in their programs.

2.5.2 Dictionary based

Systems based on dictionaries recognize concepts in texts by making matches against the dictionary entries. The dictionary is a collection of words, containing concepts of one or multiple types. There are different techniques to deal with the dictionary string matching, like exact matching, approximate matching and Soundex algorithm.

Concepts with short names can produce a large number of false positives, which can be avoided by removing those names from the dictionary. Another common problem is the missing of spelling variations of the concepts in the dictionary, which can be surpassed by applying approximate string matching techniques.

Systems that are dictionary based normally have faster processing. IndexFinder is an algorithm that generates UMLS concepts by permuting words from an input text and then filtering out the irrelevant concepts via syntactic and semantic filtering [ZCM+03]. As the data is structured, the dictionary lookup time is reduced.

ConceptMapper [TCS10] is a similar solution that allows a highly configurable system, loading the entire dictionary into memory, not having any performance issues.

A more recent solution, NOBLE Coder [TML+16] implements a general algorithm for matching terms to concepts from an arbitrary vocabulary set. The good results achieved are due to the caching of the 0.2% most occurring words in the dictionary and the use of a NoSQL to persist the data structures on disk.

2.5.3 Machine learning based

Machine learning based approaches apply algorithms to learn how to recognize specific concept types. Comparing to the previous approach, it has the advantage of recognizing a new concept that

is not yet specified in the dictionary. Since these algorithms do not provide a mapping of concepts to resources, the dictionaries are still necessary.

These methods start to train models based on annotated texts and then apply these models to non-annotated texts, predicting concepts.

Takeuchi et al. [TC03] applied Support Vector Machines in a sample of 100 journal abstract texts taken from Medline topics to identify scientific and technical concepts of molecular biology, with 74% F-score on cross-validation tests.

Gimli [CMO13] is a tool for named-entity recognition, providing trained and optimized models for recognition of biomedical entities from scientific text by applying machine-learning methods, more specifically Conditional Random Fields (CRFs).

Chapter 3

Automatic classification of a concept in lay or medico-scientific terminology

To develop systems that suggest queries in lay or medico-scientific terminology requires knowing to which terminology concepts belongs to or, at least, are more similar to.

In the next sections, several approaches to compute the degree of association of a concept to a lay and medico-scientific terminology are proposed and compared.

Different thesauri that distinguish concepts by terminology are used for this purpose, as well as the cosine similarity to measure the closeness of concepts with subsets of those thesauri.

The evaluation of the approaches uses an existing glossary containing concepts in both terminologies in English and Portuguese and a set of queries submitted by users and classified by health professionals as lay or medico-scientific.

3.1 Methodology

To determine the degree of association of a medical expression to lay or medico-scientific terminology, several variants of a method that computes the cosine similarity of the expression with different vocabularies or their subsets were developed. These variants are detailed in Table 3.1 and a diagram representing the used vocabularies is presented in Figure 3.1.

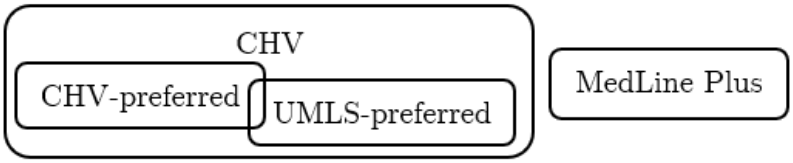


Figure 3.1: Diagram showing the used vocabularies.

Table 3.1: Description of the different approaches.

Name	Vocabulary	# expressions	Classification	Similarity method
CHV-single	CHV	146,334	Threshold	Single document
CHVpref-single	CHV-preferred	57,796		
UMLSpref-single	UMLS-preferred	30,060		
MedPlus-single	Medline Plus	3,329		
CHVUMLS pref-single	CHV-preferred and UMLS-preferred	87,856	max(simCHV, simUMLS)	
CHV-multi	CHV	146,334	Threshold	Several documents
CHVpref-multi	CHV-preferred	57,796		
UMLSpref-multi	UMLS-preferred	30,060		
MedPlus-multi	Medline Plus	3,329		
CHVUMLS pref-multi	CHV-preferred and UMLS-preferred	87,856	max(simCHV, simUMLS)	

In these variants, the queries and the documents representing the different vocabularies are seen as vectors. The cosine similarity between these vectors can be used as a measure of the score of the document for that query. Manning et al. [Man09] defined this score as it follows, in which the numerator represents the dot product between the vector of the query and the document, while the denominator is the product of their Euclidean lengths:

$$score(q,d) = \frac{\vec{V}(q) \cdot \vec{V}(d)}{|\vec{V}(q)||\vec{V}(d)|}$$

Besides the considered vocabulary, differences between these variants include the classification method, and the way vocabularies are considered in the cosine similarity computation. The classification method may be the comparison with a threshold or the identification of the most similar terminology in situations where both lay and medico-scientific vocabularies are used. Regarding the similarity computation method, a vocabulary may be considered a single document/vector or each of its expressions may be considered a different document/vector. In the latter, an average of all the similarity scores is computed.

Every method is based on a tf-idf weighting scheme. This value combines the definitions of term frequency (number of occurrences of the term in the document) and inverse document frequency (inverse of the document frequency, used to decrease the weight of frequently occurring terms and increase those that occur rarely), to produce a composite weight for each term in each document [Man09].

To reflect how important a word is to a document, the *tf-idf* weighting scheme assigns to term t a weight in document d given by:

$$tf - idf_{t,d} = tf_{t,d} \times idf_t$$

This weight is high when t occurs many times within a small number of documents; lower when the term occurs fewer times in a document, or many documents; and lowest when the term occurs in virtually all documents.

Based on these notions, three-term weighting schemes were used in every method: tf-idf, tf-idf*(1+frequencyScore) and tf-idf*(1+comboScore). The frequency and combo scores are difficulty scores associated with the expressions of the CHV, as explained in Section 2.4.2, and were used to include this information in the computation of the tf-idf values of the terms of each vocabulary. These values range from 0 to 1 and are equal to -1 when they were not calculated. The existence of each term in the CHV was verified. If the term is equal to a CHV expression, its frequency and combo scores are retrieved, otherwise, it is retrieved the average of the frequency and combo scores of expressions where the term appears. If the term does not appear at all, the frequency and combo scores are equal to 1. To this value, we add 1 in order to multiply these coefficients in the tf-idf matrix. By doing so, we benefit the most familiar terms, harm the less familiar and do not affect those of whom nothing is known.

Using the Consumer Health Vocabulary and the MedlinePlus, the approaches are divided into 2 different categories according to how each vocabulary is used: as a single document or as several documents.

Consider, as an example, the following vocabulary:

$$vocabulary = (..., acid, acne, acne\ vulgaris, adenoiditis, adrenal\ tumor, ...)$$

In the single-document approach, the cosine similarity is computed between the original medical expression (tumor, for example) and the document containing all the expressions of the considered vocabulary.

$$sim(medical\ concept, all\ vocabulary\ expressions) = cos\ \theta$$

In the multiple-document approach, instead of computing the similarity with a single document containing all the terms of a specific vocabulary, each expression is considered a different document and the similarity is computed with all of them. In the end, an average of these similarities is made.

$$\frac{\sum_{i=1}^{N} sim(medical\ concept, vocabulary\ expression) = cos\ \theta}{\#\ vocabulary\ expressions}$$

In CHV-single, the single document is composed of all the terms included in expressions of the CHV ("acid acne acne vulgaris adenoiditis adrenal tumor"). In CHVpref-single, only the terms of the CHV-preferred expressions are used ("acid acne adenoiditis adrenal tumor"). In UMLSpref-single, the document is formed by terms of UMLS-preferred expressions of the CHV ("acne acne vulgaris"). MedPlus-single uses terms of Medline Plus expressions of the UMLS ("acne"). CHVUMLSpref-single uses the maximum value between the similarity with CHV-preferred and

the similarity with UMLS-preferred, assigning the concept to the terminology to which it has a higher similarity.

Considering the previous vocabulary, CHV-multi uses several documents (“acid”, “acne”, “acne vulgaris”, “adenoiditis”, “adrenal tumor”), CHVpref-multi (“acid”, “acne”, “adenoiditis”, “adrenal tumor”), UMLSpref-multi (“acne”, “acne vulgaris”) and MedPlus-multi (“acne”).

Since the Consumer Health Vocabulary and the MedlinePlus terms are only available in English, the Google Translator API was used to translate those terms to Portuguese. By doing so, it is possible to compute the degree of association of a concept in Portuguese with the lay or medico-scientific terminology.

Both CHV and MedlinePlus vocabularies were stored in a MySQL database. For CHV we stored the CUIs, the CHV-preferred and UMLS-preferred expression (in English and Portuguese), as well as the difficulty scores. For MedlinePlus, we stored the CUIs and the expression, both in English and Portuguese.

After having all this information gathered and to implement the cosine similarity, we used Scikit-learn [PVG+11], a simple and efficient tool for data mining and data analysis that is available as a package to be used in Python. The documents and queries are transformed into the corresponding tf-idf matrix using the Sklearn TF-IDF Vectorizer and the cosine similarity is computed between them.

3.2 Evaluation

To evaluate this approach, two resources were used: the Multilingual Glossary and a set of queries submitted by users.

3.2.1 Multilingual Glossary

The Multilingual Glossary of technical and popular medical terms, available in nine European Languages [ooPN95], was used as the ground truth in this evaluation. This glossary has 1,830 medical expressions and, for each technical expression, has its corresponding popular expression.

Measures such as sensitivity (SEN), specificity (SPC), accuracy (ACC) and Receiver Operating Characteristic Distance (ROCD) were computed. Sensitivity measures the proportion of actual positives that are correctly identified as such, while specificity measures the proportion of actual negatives that are correctly identified as such. Accuracy is the tax of correct classifications and ROCD associates sensitivity and specificity, depicting relative tradeoffs between benefits and costs [Man09]. In the approaches where classification is done through threshold comparison, these measures were computed for several thresholds (]0, 1[).

After the computation of these measures, the goal is to maximize the value, for each method, of the ACC, SEN, and SPC and minimize the value of the ROCD, since this is an indicator of the best performance. Since the range of threshold values varies between 0 and 1, SEN and SPC can take the maximum value of 1. Therefore, to determine the best threshold for each method it is given higher importance to the ACC and ROCD, since they are aggregated measures.

To compare all methods, the values for the best threshold configuration are registered in Table 3.2 using the English and Portuguese terms (for each method, each line contains the term weighting schemes used: tf-idf, tf-idf*(1+frequencyScore), and tf-idf*(1+comboScore)).

Table 3.2: Evaluation measures for each method in both languages.

	English					Portuguese				
Method	**Threshold**	**ACC**	**SEN**	**SPC**	**ROCD**	**Threshold**	**ACC**	**SEN**	**SPC**	**ROCD**
CHV-single	0.0052	0.8454	0.7759	0.9056	0.2431	0.0011	0.8349	0.7889	0.8782	0.2468
	0.0043	**0.8549**	0.7899	0.9082	**0.2293**	0.0015	**0.8387**	0.7789	0.8932	**0.2455**
	0.0051	0.8539	0.7759	0.9209	0.2376	0.0015	0.8380	0.7765	0.8959	0.2465
CHVpref-single	0.0045	0.8428	0.7760	0.9015	0.2447	0.0014	0.8343	0.7789	0.8865	0.2485
	0.0043	0.8496	0.7863	0.9055	0.2337	0.0020	0.8357	0.7661	0.9015	0.2538
	0.0038	0.8393	0.7985	0.8763	0.2364	0.0010	0.8349	0.7824	0.8843	0.2464
UMLSpref-single	0.0007	0.4880	0.7321	0.1862	0.8567	0.0001	0.3117	0.4696	0.1569	0.9960
	0.0004	0.4880	0.7321	0.1862	0.8567	0.0001	0.3117	0.4696	0.1569	0.9960
	0.0004	0.4880	0.7321	0.1862	0.8567	0.0001	0.3117	0.4696	0.1569	0.9960
MedPlus-single	0.0030	0.8069	0.7875	0.8239	0.2759	0.0036	0.8263	0.7095	0.9344	0.2978
	0.0025	0.8321	0.7614	0.8928	0.2616	0.0020	0.8183	0.7484	0.8834	0.2774
	0.0026	0.8305	0.7664	0.8859	0.2599	0.0021	0.8203	0.7454	0.8899	0.2773
CHVUMLSpref-single	-	0.5732	0.6173	0.5296	0.6064	-	0.4270	0.5969	0.2484	0.8529
		0.5888	0.6617	0.5156	0.5908		0.4216	0.5846	0.2501	0.8572
		0.5279	0.5798	0.4748	0.6728		0.4015	0.4398	0.3620	0.8490
CHV-multi	0.0002	0.8377	0.7636	0.9019	0.2559	0.0002	0.8303	0.7712	0.8859	0.2556
	0.0002	0.8428	0.7659	0.9094	0.2510	0.0002	0.8317	0.7699	0.8899	0.2551
	0.0002	0.8425	0.7654	0.9093	0.2516	0.0002	0.8309	0.7700	0.8882	0.2557
CHVpref-multi	0.0002	0.8379	0.7554	0.9094	0.2608	0.0002	0.8294	0.7708	0.8848	0.2565
	0.0002	0.8422	0.7626	0.9107	0.2536	0.0002	0.8277	0.7656	0.8864	0.2605
	0.0002	0.8412	0.7605	0.9107	0.2556	0.0002	0.8279	0.7679	0.8847	0.2591
UMLSpref-multi	0.0044	0.5888	0.0008	0.9929	0.9992	0.0349	0.5200	0.0012	**0.9923**	0.9989
	0.0055	0.5927	0.0008	**0.9995**	0.9992	0.0272	0.5200	0.0012	**0.9923**	0.9989
	0.0053	0.5927	0.0008	**0.9995**	0.9992	0.0302	0.5200	0.0012	**0.9923**	0.9989
MedPlus-multi	0.0001	0.8069	0.7875	0.8239	0.2759	0.0007	0.8179	0.6812	0.9444	0.3236
	0.0002	0.8095	0.7671	0.8462	0.2791	0.0005	0.8125	0.7015	0.9156	0.3102
	0.0001	0.8062	0.7862	0.8239	0.2770	0.0007	0.8185	0.6806	0.9460	0.3239
CHVUMLSpref-multi	-	0.6478	0.8663	0.3983	0.6164	-	0.5175	0.8214	0.1939	0.8257
		0.6484	0.8664	0.3989	0.6158		0.5195	**0.8222**	0.1978	0.8217
		0.6484	**0.8672**	0.3974	0.6171		0.5126	0.8118	0.1943	0.8274

Comparing the methods that used lay vocabularies in the subset (CHV-single, CHVpref-single, MedPlus-single, CHV-multi, CHVpref-multi, MedPlus-multi), it is possible to say that CHV-single, CHVpref-single, CHV-multi, and CHVpref-multi have similar performance, with the first two having slightly better results. Since these four methods use vocabularies with more terms than the Medline Plus (used in MedPlus-single and MedPlus-multi), this was already expected.

UMLSpref-single and UMLSpref-multi had worse performance in distinguishing between lay and medico-scientific terms probably because, despite having a more medico-scientific character, it had the origin in a vocabulary with terminology for consumers, which could mean that the expressions are still consumer-friendly. Since the CHVUMLSpref-single and CHVUMLSpref-multi are methods based on the previous ones, they also had a bad performance.

Since the CHV-single is the vocabulary with a higher number of expressions, it was expected that it was the one retrieving the best results.

It is possible to observe that using the frequency and combo scores slightly improve the results, mainly in the English approaches.

Comparing the English and Portuguese approaches, it is possible to confirm that the methods returned slight better results by using the original vocabulary, which is normal, since the

Portuguese terms were obtained using the Google Translator API. The CHVUMLSpref-single, CHVUMLSpref-multi and UMLSpref-single had worse results compared with the corresponding methods in English, while those with lay vocabulary as subset had similar results to the ones that used English terms.

3.2.2 Queries submitted by users

As a complement to this evaluation, 492 queries submitted by users were also used. These queries had origin in a study done with users and were later classified as lay or medical expressions by three researchers, two of whom were health professionals. Of these three judgments, we picked the most frequent classification as the query's terminology.

These queries were used to evaluate the performance of the previously chosen method (CHV-single with frequency scores) and threshold in determining whether an expression is lay or medico-scientific.

A Python library that ported Google's language detection algorithm [Shu10] was used to determine whether the expression is in English or Portuguese, using the corresponding method based on that result. The results of the method were compared with the evaluations from health professionals and the same evaluation measures were calculated and are presented in Table 3.3.

Table 3.3: Evaluation measures for the queries.

Language	**ACC**	**SEN**	**SPC**	**ROCD**
English	0.7397	0.9034	0.2745	0.7319
Portuguese	0.7128	0.9649	0.3243	0.6766

As expected, the results are slightly worse than the ones before, since we used the thresholds that maximized the accuracy for the Multilingual Glossary.

For 22% of the queries, the method returned the value of 0, which did not allow to draw a conclusion as to which terminology the expression belongs to. These were not considered for the previous evaluation. This high value can be associated with the fact that the same query appears several times and, for that, it is not possible to compute the similarity.

3.3 Conclusions

Comparing CHV-single and CHV-multi, which belong to different categories, it is possible to state that the first performs better than the latter in all the measures. So, in this case, in the computation of the cosine similarity, it is better to use a subset where all terms are considered as one document, instead of having several documents with the different terms of the vocabulary.

The other methods had a very similar performance regardless of how the cosine similarity is computed, except for the UMLSpref approach. For the UMLSpref-single, the sensitivity is better, and the specificity is worse, compared with the UMLSpref-multi.

Besides, the methods that used CHV outperformed most of the other methods. So, the best method to assign a terminology to a concept in both languages is the method that uses CHV terms as a single vocabulary and that takes advantage of the frequency scores, with a threshold of 0.0043 for English and 0.0015 for Portuguese, value that represents the cosine similarity from which the concept is considered to be lay.

As expected, the methods that used the original expressions in English outperformed the ones that used the corresponding translation into Portuguese. This may be caused by the fact that the translation that the Google Translator API provides is not entirely reliable. When evaluating new expressions, the method needs to be improved, since in some cases the cosine similarity value is 0, not allowing assigning a terminology to an expression.

Chapter 4

Strategies to generate query suggestions

This chapter presents the details of the strategies followed to generate suggestions in lay and medico-scientific terminologies, both in English and in Portuguese, given an initial user query. To do this, we also used the strategy explained in the previous chapter, whenever it is necessary to assign one of the terminologies to a concept.

4.1 Implementation details

To generate the query suggestions, several approaches were followed. The main idea is explained in Figure 4.1.

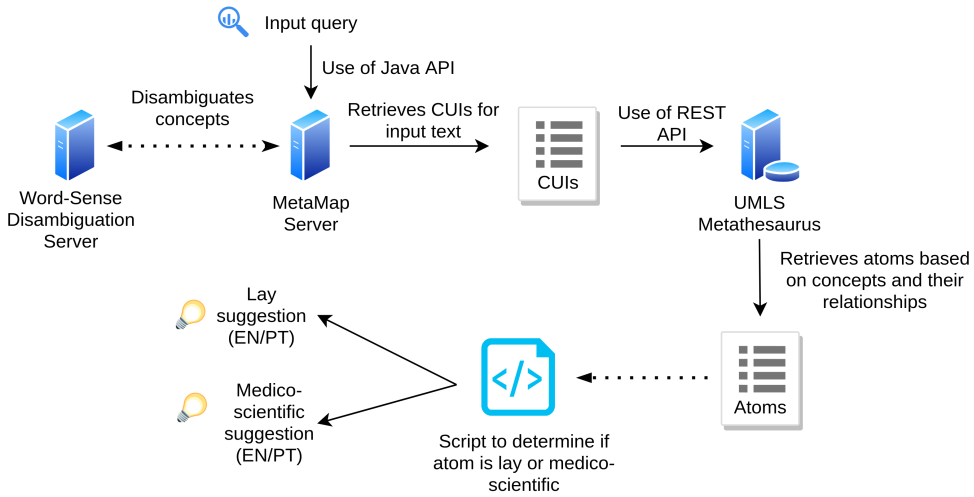


Figure 4.1: Diagram explaining the methodology followed to generate suggestions.

All of these approaches use the MetaMap Java API to communicate with the MetaMap Server. Depending on the approach, this one can communicate with the Word-Sense Disambiguation

Server to remove the ambiguity between certain concepts. After obtaining the list of CUIs for a given query, we used the REST API to communicate with the UMLS Metathesaurus and obtain the atoms related to a certain concept, as well as to make use of the existing relationships to get new atoms for a given CUI. When necessary, it is determined if an atom is more associated with the lay or medico-scientific terminology and the final atoms are used to create the query suggestions.

4.1.1 MetaMap

Given an initial query, the MetaMap tool, as described in Section 2.5.1.1, is used to analyze the input text and retrieve the UMLS concepts that are present in it.

Using the query "emotional and mental disorders" as an example, it is possible to observe, in Figure 4.2, that the tool identifies three phrases. From these three, only two are considered, since the phrase "and" is considered a stop-word and does not have any match in the UMLS. For each one of the considered phrases, the corresponding list of mappings is also retrieved, ordered by score. The left image is the result of running the MetaMap with the default options and the right image is the result of running the MetaMap together with the Word-Sense Disambiguation Server, invoking MetaMap with the option "-y" to eliminate possible ambiguities.

```
input text:
  emotional and mental disorders
Phrase:
 text: emotional
Mappings:
 Map Score: -1000
   Score: -1000
   Concept Id: C0013987
   Concept Name: Emotional
   Preferred Name: Emotions
   Matched Words: [emotional]
 Map Score: -1000
   Score: -1000
   Concept Id: C0849912
   Concept Name: Emotional
   Preferred Name: Emotional
   Matched Words: [emotional]
Phrase:
 text: and
Mappings:
Phrase:
 text: mental disorders
Mappings:
 Map Score: -1000
   Score: -1000
   Concept Id: C0004936
   Concept Name: MENTAL DISORDERS
   Preferred Name: Mental disorders
   Matched Words: [mental, disorders]
```

```
input text:
  emotional and mental disorders
Phrase:
 text: emotional
Mappings:
 Map Score: -1000
   Score: -1000
   Concept Id: C0849912
   Concept Name: Emotional
   Preferred Name: Emotional
   Matched Words: [emotional]
Phrase:
 text: and
Mappings:
Phrase:
 text: mental disorders
Mappings:
 Map Score: -1000
   Score: -1000
   Concept Id: C0004936
   Concept Name: MENTAL DISORDERS
   Preferred Name: Mental disorders
   Matched Words: [mental, disorders]
```

Figure 4.2: Results of running MetaMap for the query "emotional and mental disorders". Left image with default options and right image with Word-Sense Disambiguation Server running.

The MetaMap version that was used is from 2018, which uses the 2018AB release of UMLS.

In some cases, there is more than one mapping in each phrase. Normally, the approach is to choose the first mapping from the list of mappings for each phrase, since it is the one with the

highest score. In the left example in Figure 4.2, for the phrase "emotional", the chosen concept would be the one with concept id equal to C0013987. In this specific case, the two mappings have the same score, but with another example, different scores could occur, and with this approach it is guaranteed that the best possible mapping is chosen.

Another approach to chose the concept associated with a certain phrase is to use the Word-Sense Disambiguation (WSD) Server that MetaMap includes. In cases where MetaMap maps two or more concepts to a recognized entity in text, the WSD Server determines which concept is the best choice for the entity using the context in which the entity occurs. The invocation of WSD Server adds some overhead to processing, making the approach slightly slower when compared to the MetaMap by itself [ALA10]. In this case, for the phrase "emotional", the chosen concept is the one with concept id equal to C0849912.

For both approaches, after the identification of the concepts by the MetaMap, the CUIs and the concept names are stored, to be used as the source of information to generate the suggestions.

Therefore, for each of the methods used to generate the query suggestions described below, both approaches are used and compared.

4.1.2 Methods to generate query suggestions

After obtaining the CUIs included in the query with the help of MetaMap as previously described, several methods were implemented. All make use of the information in the UMLS Metathesaurus, like the preferred names for concepts as well as the various relationships between them. These methods are briefly described in Table 4.1.

Table 4.1: Description of the different methods.

Name	Used Vocabulary	Considered atoms	Relationships	Type of relationships
MM + CHV-preferred/ UMLS-preferred	CHV	All	-	-
MM + Preferred Atoms	All vocabularies from UMLS	Preferred atoms	-	-
MM + All preferred/ synonym atoms	All vocabularies from UMLS	All English preferred/ synonym atoms	-	-
MM + All Atoms	All vocabularies from UMLS	All English atoms	-	-
MM + All Atoms + Child/Parent/Same Relations	All vocabularies from UMLS	All English atoms	Relationships from atoms	Child/Parent/Same
MM + Broader/ Narrower Concepts	All vocabularies from UMLS	All English atoms	Relationships from concepts	Broader/Narrower

The main differences between these methods consist in the chosen vocabulary, since one only uses a subset of the UMLS called CHV, while others use all vocabularies from UMLS; the considered atoms, since all atoms from a vocabulary can be considered, but sometimes only the preferred or the English preferred atoms are used; the relationships used, that can be the relationships of the atoms associated to the identified concept, as well as the relationships directly associated with the concepts; and the type of relationship, that can be a relationship from the same level of the concept/atom, a child or parent relationship, as well as a broader/narrower relationship.

All the implementation details necessary to create these methods are described in the following sections. In all the methods, MM stands for MetaMap and represents the use of it in each of them.

The query "caffeine high blood pressure" is used as input to show the differences between the suggestions that are created in all the different methods. This query has two associated CUIs, which are the "C0006644" and "C0020538". In all sections, the corresponding English suggestions for both terminologies are shown. To generate the Portuguese ones we used the Google Translator API. The suggestions are also available in Table 4.2, to be easily compared.

4.1.2.1 MM + CHV-preferred/UMLS-preferred

Considering the CUIs identified as described in Section 4.1.1 and the MySQL database containing the CHV described in Section 3.1, we used the CUI to retrieve its corresponding CHV-preferred and UMLS-preferred expression. When there are no expressions associated with that CUI in the database, we used the concept name that MetaMap provides.

The new query is formed by the concatenation of all the expressions retrieved. The CHV-preferred expressions associated with the CUIs are "caffeine" and "high blood pressure disorder" and the UMLS-preferred expressions are "caffeine" and "hypertension".

The lay suggestion contains the CHV-preferred expressions (MM + CHV-preferred), while the medico-scientific suggestion contains the UMLS-preferred expressions (MM + UMLS-preferred).

4.1.2.2 MM + Preferred Atoms

In this method, for each CUI identified before, an HTTP request to the UMLS Metathesaurus is made through its REST API, to obtain the default preferred atom associated with that CUI.

For example, for the CUI "C0020538" whose concept name is "Hypertensive disease", the result of the HTTP request is the JSON available in Figure 4.3.

From this JSON, it is used the "result" JSONArray to obtain the "name" of the atom that corresponds to the default preferred name of the concept, as can be seen in the "termType" that is equal to "PN", which stands for Metathesaurus preferred name.

Since there is always a default preferred atom associated with a certain CUI, the set of all the preferred atoms associated with the CUIs from the initial query form the new suggestion.

One of the disadvantages of this method is that it only provides one atom per CUI and, therefore, it is not possible to make suggestions in the two terminologies. Thus, the new query is "Caffeine Hypertensive disease".

Using the preferred terms, we also thought to investigate the parent, child and same level relationships associated with this atom, as a way of creating a list of where the more lay or medico-scientific atoms would be chosen, depending on the terminology of the suggestion. After some investigation, we discovered that none of these atoms designated as preferred in Metathesaurus has this type of relationships, making this new method equal to the one currently implemented, thus not worth to implement.

```
{
    "pageSize": 25,
    "pageNumber": 1,
    "pageCount": 1,
    "result": {
        "classType": "Atom",
        "ui": "A8299096",
        "suppressible": "false",
        "obsolete": "false",
        "rootSource": "MTH",
        "termType": "PN",
        "code": "https://uts-ws.nlm.nih.gov/rest/content/2019AA/source/MTH/NOCODE",
        "concept": "https://uts-ws.nlm.nih.gov/rest/content/2019AA/CUI/C0020538",
        "sourceConcept": "NONE",
        "sourceDescriptor": "NONE",
        "attributes": "NONE",
        "parents": "NONE",
        "ancestors": null,
        "children": "NONE",
        "descendants": null,
        "relations": "NONE",
        "name": "Hypertensive disease",
        "language": "ENG",
        "contentViewMemberships": []
    }
}
```


Figure 4.3: JSON containing the preferred atom for the CUI “C0020538”.

4.1.2.3 MM + All preferred/synonym atoms

In this method, instead of retrieving the default preferred atom for each identified CUI, we retrieved the list of all English atoms that are the preferred names or a synonym in the various vocabularies of the UMLS. To do so, an HTTP request to the UMLS Metathesaurus is also made through its REST API.

For example, for the CUI “C0020538” whose concept name is “Hypertensive disease”, the result of the HTTP request is the JSON available in Figure 4.4. Some attributes that are not relevant to the method have been omitted.

In this JSON, the “result” JSONArray contains several objects representing the atoms that are preferred names or synonyms in a certain vocabulary, as can be seen by their “termType”, which is “PT” or “SY”. For each one, the “name” field is stored. Therefore, for a certain CUI, all the atoms associated with it are saved. If, for some reason, a concept does not have atoms associated, the concept name is saved.

Since it is not known if these atoms are more lay or medico-scientific, the approach described in Chapter 3 is used to classify each one in these terminologies, to be possible to create the suggestions in both terminologies. Therefore, all the atoms of a certain CUI are classified as lay or medico-scientific, with a value that represents its similarity with those terminologies.

For each CUI we chose the atom that has the highest value of similarity with the lay terminology, to generate the lay suggestion, and with the medico-scientific terminology, to generate the medico-scientific suggestion.

After completing this process for all CUIs, the generated lay suggestion for the example is

```
{
    "pageSize": 25,
    "pageNumber": 1,
    "pageCount": 4,
    "result": [
     (...)
        {
            "classType": "Atom",
            "ui": "A3258270",
             (...)
            "rootSource": "SNOMEDCT_US",
            "termType": "SY",
            "code": "https://uts-ws.nlm.nih.gov/rest/content/2019AA/source/SNOMEDCT_US/38341003",
            "concept": "https://uts-ws.nlm.nih.gov/rest/content/2019AA/CUI/C0020538",
            "sourceConcept": "https://uts-ws.nlm.nih.gov/rest/content/2019AA/source/SNOMEDCT_US/38341003",
             (...)
            "name": "HBP - High blood pressure",
            "language": "ENG",
             (...)
        },
        (...)
        {
            "classType": "Atom",
            "ui": "A21145030",
             (...)
            "rootSource": "MEDLINEPLUS",
            "termType": "PT",
            "code": "https://uts-ws.nlm.nih.gov/rest/content/2019AA/source/MEDLINEPLUS/34",
            "concept": "https://uts-ws.nlm.nih.gov/rest/content/2019AA/CUI/C0020538",
            "sourceConcept": "NONE",
             (...)
            "name": "High Blood Pressure",
            "language": "ENG",
             (...)
        },
        (...)
        }
    ]
}
```


Figure 4.4: JSON containing the list of atoms in different vocabularies for the CUI “C0020538”.

“Caffeine, NOS Hypertensive cardiovascular disease or syndrome” and the medico-scientific suggestion is “CAFFEINE HTN”.

4.1.2.4 MM + All Atoms

This method is equal to the previous one, with one difference: it retrieves all the English atoms that are associated with a concept, instead of retrieving only the preferred and synonym ones. Therefore, a result similar to the one showed in Figure 4.4 is returned.

With those atoms, we follow the same process defined before. Each concept has associated a list of atoms, in which each one is classified as lay or medico-scientific, with a similarity value associated. The most lay or medico-scientific atoms are selected for each concept, depending on the terminology, and this process is done for all concepts of all queries.

The lay suggestion for the example is “caffeine (medication) Hypertensive cardiovascular disease or syndrome” and the medico-scientific one is “CAFFEINE HTN”.

4.1.2.5 MM + All Atoms + Child/Parent/Same Relations

This method is based on the previous one, since it also identifies all English atoms that are associated with a concept, but with the extent of making use of its relations with other atoms.

For each atom identified we used its “ui”, to obtain the relations that this specific atom has with other atoms. An example of the existing relations of the atom “A21145030”, whose name is “High Blood Pressure”, is shown in Figure 4.5.


```
{
    "pageSize": 1000,
    "pageNumber": 1,
    "pageCount": 1,
    "result": [
        {
            "classType": "AtomRelation",
            "ui": "R136030263",
            "suppressible": false,
            "sourceUi": "NONE",
            "obsolete": false,
            "sourceOriginated": false,
            "rootSource": "MEDLINEPLUS",
            "relationLabel": "CHD",
            "additionalRelationLabel": "",
            "groupId": "NONE",
            "attributeCount": 0,
            "relatedId": "https://uts-ws.nlm.nih.gov/rest/content/2019AA/AUI/A21144434",
            "relatedIdName": "Seniors"
        },
        {
            "classType": "AtomRelation",
            "ui": "R136030335",
            "suppressible": false,
            "sourceUi": "NONE",
            "obsolete": false,
            "sourceOriginated": false,
            "rootSource": "MEDLINEPLUS",
            "relationLabel": "CHD",
            "additionalRelationLabel": "",
            "groupId": "NONE",
            "attributeCount": 0,
            "relatedId": "https://uts-ws.nlm.nih.gov/rest/content/2019AA/AUI/A21144435",
            "relatedIdName": "Blood, Heart and Circulation"
        },
        (...)
    ]
}
```


Figure 4.5: JSON containing the list of relations for the atom “A21145030”.

For a given atom it is intended to explore other atoms which are situated at a higher level (parent relations), at the same level, and at a lower level (child relations). To select this type of relations from the result list we use the attribute “relationLabel”. The parent relations are associated with a “PAR” label, the same level relations are associated with a “RL”/“SY” (alike/synonym) label, and the child relations are associated with a “CHD” label.

After filtering these types of relations, the “relatedIdName” is stored for each one, which corresponds to the atom name that is related with the initial one. Thus, for each of the initial atoms, three lists containing the atoms names from the three relation’s types are created. If an atom does not have any type of relation, its name is preserved.

Since these atoms are all associated with the initial concept, they are all gathered together and classified as lay and medico-scientific for each relationship, like in the previous method, choosing the ones with higher value, depending on the terminology.

Considering the child relationship, the generated lay suggestion for the example is “Heart Rate Affecting Drugs Blood Pressure Disorders” and the medico-scientific one is “Xanthine product

sympathomimetic". For the parent relationship, "Caffeine solution Hypertensive heart disease" is the lay suggestion, while "Caffeine solution Hypertensive episode" is the medico-scientific one. Finally, for the same level relationship, "Caffeine (substance) High Blood Pressure" is the lay suggestion and the medico-scientific suggestion is "Caffeine HTN".

4.1.2.6 MM + Broader/Narrower Concepts

This method is similar to the previous one, since it makes use of relations to retrieve new atoms to use in the suggestions. The main difference consists in using the relations existing between concepts, instead of using the relations of atoms associated with concepts.

So, for a certain concept identified in an initial query, an HTTP request to the UMLS Metathesaurus is made through its REST API to obtain its relations. For example, for the CUI "C0020538" whose concept name is "Hypertensive disease", the result of the HTTP request is the JSON available in Figure 4.6.


```
{
    "pageSize": 25,
    "pageNumber": 1,
    "pageCount": 4,
    "result": [
        {
            "classType": "ConceptRelation",
            "ui": "R19384035",
            "suppressible": false,
            "sourceUi": "NONE",
            "obsolete": false,
            "sourceOriginated": false,
            "rootSource": "MTH",
            "relationLabel": "RB",
            "additionalRelationLabel": "",
            "groupId": "NONE",
            "attributeCount": 0,
            "relatedId": "https://uts-ws.nlm.nih.gov/rest/content/2019AA/CUI/C0020539",
            "relatedIdName": "Hypertension due to oral contraceptive"
        },
        {
            "classType": "ConceptRelation",
            "ui": "R19383983",
            "suppressible": false,
            "sourceUi": "NONE",
            "obsolete": false,
            "sourceOriginated": false,
            "rootSource": "MTH",
            "relationLabel": "RB",
            "additionalRelationLabel": "",
            "groupId": "NONE",
            "attributeCount": 0,
            "relatedId": "https://uts-ws.nlm.nih.gov/rest/content/2019AA/CUI/C0020540",
            "relatedIdName": "Malignant Hypertension"
        },
        (...)
    ]
}
```


Figure 4.6: JSON containing the list of relations for the CUI "C0020538".

In this case, we intend to explore broader concepts, that can have an ascendant relation with the concept, as well as narrower concepts, that can have a descendant relation with it. These

relations are selected from the previous result list using the "relationLabel". The broader relations are associated with a "RB" label and the narrower ones with a "RN" label.

For each concept, two lists are created, one containing the concept names (obtained through the "relatedIdName" attribute) that are related with the initial one through a broader relation and other with the narrower ones. If a certain concept does not have any type of these relations, its name is added to the list.

The concept names from these lists are also classified as lay or medico-scientific, choosing those that have a higher value, depending on the terminology.

Considering the broader relation, the generated lay suggestion for the example is "Genace brand of acetaminophen/aspirin/caffeine Hypertensive renal disease, unspecified, without mention of renal failure" and the medico-scientific suggestion is "Excedrin Migraine Hypertensive renal disease, unspecified, without mention of renal failure". For the narrower relation, "CNS stimulants of abuse abnormalities of blood pressure regulation" is the lay suggestion, while "Xanthine or xanthine derivative abnormalities of blood pressure regulation" is the medico-scientific one.

Table 4.2: Lay and medico-scientific suggestions given by the different methods for the query "caffeine high blood pressure". Each cell contains one suggestion.

Method	**Lay suggestion**	**Medico-scientific suggestion**
MM + CHV/UMLS-preferred	caffeine high blood pressure disorder	caffeine hypertension
MM + Preferred Atoms	Caffeine Hypertensive disease	Caffeine Hypertensive disease
MM + All preferred/synonym atoms	Caffeine, NOS Hypertensive cardiovascular disease or syndrome	CAFFEINE HTN
MM + All Atoms	caffeine (medication) Hypertensive cardiovascular disease or syndrome	CAFFEINE HTN
MM + All Atoms + Child Relations	Heart Rate Affecting Drugs Blood Pressure Disorders	Xanthine product sympathomimetic
MM + All Atoms + Parent Relations	Caffeine solution Hypertensive heart disease	Caffeine solution Hypertensive episode
MM + All Atoms + Same Relations	Caffeine (substance) High Blood Pressure	Caffeine HTN
MM + Broader Concepts	Genace brand of acetaminophen/aspirin/caffeine Hypertensive renal disease, unspecified, without mentionof renal failure	Excedrin Migraine Hypertensive renal disease, unspecified, without mention of renal failure
MM + Narrower Concepts	CNS stimulants of abuse abnormalities of blood pressure regulation	Xanthine or xanthine derivative abnormalities of blood pressure regulation

Chapter 5

Evaluation strategy and Results

This chapter presents the necessary steps that have been followed to evaluate the suggestions given by all the previous methods. Since the proposed suggestions are both in English and Portuguese language, two collections have been used and are described below.

5.1 Test collections

To assess the effectiveness of the previously developed methods, two different test collections are used: one in English and another in Portuguese. For each test collection, there is a set of queries, documents and assessments of relevance and understandability, which will be used for the evaluation. The English test collection is the one provided by the CLEF eHealth Lab from 2018 and the Portuguese one was compiled within the scope of this dissertation.

5.1.1 English test collection - CLEF eHealth Lab 2018

The CLEF 2018 eHealth Evaluation Lab, Consumer Health Search task investigates the effectiveness of search engines in providing access to correct, trustworthy and understandable medical information present on the Web for laypeople. To do so, it provides resources and evaluation methods to test and validate search systems, such as a shared collection of documents and queries, as well as relevance assessments [JZP+18].

5.1.1.1 Queries

The queries used in this task consist of a set of 50 queries, which were issued by the general public to the HON[1] search service. These queries are available in Appendix A.1. Queries are formatted one per line in a tab-separated query file, with the first string being the query id, and the second string being the query text.

[1] https://hon.ch/en/

A sample of raw queries submitted to this search service and collected over a period of six months was used as the source of the final queries, which were manually selected by a domain expert. By doing so, it was possible to have a representative subset of the type of queries submitted to this search engine. Queries that have capitalized text, with complex medical terms and in languages other than English were not considered. Since the queries were not preprocessed they might have spelling mistakes.

5.1.1.2 Dataset

This dataset was created by compiling Web pages of selected domains acquired from the CommonCrawl[2]. These Web pages correspond to the result pages of submitting the CLEF 2018 queries to the Microsoft Bing APIs repeatedly over a period of few weeks. These results were further augmented by including several known reliable health Websites and other known unreliable health Websites, from lists previously compiled by health institutions and agencies.

The corpus was divided into folders, by domain name, containing a file for each Webpage from the domain available in the CommonCrawl dump. In total, 1903 were successfully acquired, but 84 were not available in the CommonCrawl dump, and for these, an empty folder in the corpus exists and represents the requested domain, although without any Web pages retrieved.

The full collection contains 5,535,120 Web pages and its uncompressed size is about 480GB. Given the size of the complete collection, the index provided by the organization was used to return the documents to the introduced queries. The index (available in Appendix B.1) was developed to the Elasticsearch v.5.1.1. and has not been optimized, beyond converting all text to lowercase, applying a standard stopword list and a standard rule-based stemming algorithm. Elasticsearch[3] is a full-text search engine, used to retrieved all kind of information from the indexed documents. The retrieval model has been set to the standard Elasticsearch BM25 (b=0.75 and k_1=1.2).

5.1.1.3 Assessments

To create the relevance and understandability assessments used in the evaluation, the organization of CLEF 2018 collected all unique topic-document pairs from all submitted runs from the different teams. After weighting each document according to their overall contribution to the effectiveness evaluation, they selected the top 500 weighted documents to form an assessment pool of 25,000 topic-document pairs.

After the assessment pool was formed, they developed several relevance assessment tasks to be issued as Human Intelligence Tasks (HITs) on Amazon Mechanical Turk[4], a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce. Workers (with at least 90% acceptance rate) were presented with the base query and the narrative of a topic with a link to the archived Webpage to be assessed, and classified the documents as “Highly relevant” (2), “Somewhat relevant” (1) and “Not relevant”

[2]http://commoncrawl.org/
[3]https://www.elastic.co/pt/products/elasticsearch
[4]https://www.mturk.com/

(0). Understandability assessments were also collected, using an integer value between 0 and 100 (lower values mean harder to understand).

These assessments are stored in a tab-separated file, containing the query id, the iteration number, the document id (that is the same as the filename of the Web pages) and the integer value representing the assessment.

5.1.2 Portuguese test collection

The Portuguese test collection was compiled within the scope of this dissertation, the result of a diverse number of research tasks with given queries from three editions of an Information Retrieval course. Each student was assigned with 2 queries representing different information needs, with the purpose of introducing each of these queries into 4 different search engines (Google[5], Bing[6], Yahoo![7] and HONSearch[8]) and to evaluate the first 30 results for each of these engines considering relevance and understandability.

5.1.2.1 Queries

The queries used in this test collection are the ones from CLEF eHealth Lab 2016 [KGS+16] and are available in Appendix A.2.

To create these queries, real health information needs expressed by the general public on posts from public health forums were considered. These posts were extracted from the *AskDocs*[9] section of Reddit. In this section, users can post a description of a medical case or ask a medical question seeking medical information. Other users can also comment on those posts.

The organization chose posts that were descriptive, clear and understandable. Posts containing information regarding the author, such as age, gender, medical history and current medical condition were preferred. The comments were also taken into account in the selection. The selection was manually done by a student, and a total of 50 posts were used for query creation.

The selected post forums were presented to 6 query creators: 3 medical experts and 3 lay users, that created a total of 300 queries.

Since all these queries are in the English language, they were translated to Portuguese with the help of a medical doctor.

5.1.2.2 Dataset

The previous queries in Portuguese were divided by three editions of the course "Recuperação de Informação", 2016/2017, 2017/2018 and 2018/2019. Each student of this course had two queries assigned, using 204 of the 300 queries. These queries were then submitted on the four search engines enumerated before and the first 30 URLs from the results list were stored.

[5] https://www.google.com/
[6] https://www.bing.com/
[7] https://www.yahoo.com/
[8] https://www.healthonnet.org/HONsearch/Patients/index.html
[9] https://www.reddit.com/r/AskDocs/

At the time of writing this dissertation, these URLs were used to create the dataset. Each URL was downloaded using the wget command for Linux. From a total of 16505 URLs, only 14146 were available to download, because some of these pages had already been accessed a long time ago and were no longer available.

To index this collection, the FSCrawler for Elasticsearch[10] tool was used. This crawler helps to index binary documents, such as .html files, like the ones downloaded before. The configuration file used for this tool is available at Appendix B.2 and the Elasticsearch version used was the v7.1.1.

5.1.2.3 Assessments

While saving the URLs of the top 30 results for each query, students were also asked to familiarize themselves with the health problem adjacent to the query and to evaluate the returned document for relevance and understandability.

In the relevance assessment the students should consider that the documents must contain the necessary information to satisfy the information need. According to that, the relevance values should be considered as "Highly relevant" (2), "Somewhat relevant" (1) and "Non relevant" (0). In the understandability assessment, the values should be 0 when the document uses too much medico-scientific terminology and therefore the document cannot be understood; 1 if the document is partially understood; and 2 if the document is completely understood.

Since it was not possible to download certain Web pages in the previous process, given their unavailability, and since these were associated with some of these assessments, some had to be discarded. By discarding these judgments, it was verified if all the queries had at least one associated assessment, which was not verified in some cases. Thus, these were also discarded, leaving a total of 200 queries.

These assessments are also stored in a tab-separated file, containing the query id, the iteration number, the document id (that is the same as the filename of the Web pages) and the integer value representing the assessment.

5.2 MetaMap's Evaluation

Since all the created methods to generate suggestions described in Section 4.1.2 rely heavily on the identification of medical concepts in the queries introduced by users, it is very important to see how the chosen tool (MetaMap) performs this task with queries introduced in English and Portuguese.

To do so, the 200 queries from the Portuguese test collection were used, since they come from a query initially written in English and have associated their translation in Portuguese. The MetaMap was used to obtain the number of medical concepts present in each of the queries, both

[10] https://fscrawler.readthedocs.io/en/latest/

in English and Portuguese. Using these values, different measures were calculated in order to evaluate the performance, and are presented in Table 5.1.

Table 5.1: MetaMap's performance using English and Portuguese queries.

Language	**Minimum number of detected concepts per query**	**Maximum number of detected concepts per query**	**Average number of detected concepts per query**	**Queries without mappings**
English	1	652	14.545	0%
Portuguese	0	24	0.83	79.5%

It is possible to conclude that MetaMap is able to detect at least one medical concept to a certain English query, which does not happen when the chosen language is Portuguese, and gives much more alternatives (looking to the maximum number of concepts detected) when the query is written in English. On average, there are 14.545 detected concepts in the English queries, while in the Portuguese ones the average does not reach at least one concept.

Concluding, it is adequate to use MetaMap to detect concepts using English queries, and it should not be used with Portuguese queries, since almost 80% of these do not have any map to medical concepts.

Given the bad performance of MetaMap in the Portuguese language, it was decided that, in the Portuguese evaluation, the English queries associated with them should be used throughout the process of generating suggestions, translating the query at the end of the whole process.

5.3 Methods' Evaluation

Each query created in the previously defined methods was used to retrieve the 10 best documents from one of the test collections, depending on the language of the query. To do so, we used Elasticsearch, in order to match the query content with the content of the indexed Web pages of the collections. Thus, for a certain query, 10 documents were returned, ordered by a ranking score representing the similarity of the query content with the Webpage content.

These results were saved in a tab-separated file, containing the query id, the iteration number, the document id, the ranking position and the score. This file follows the same format of the files containing the relevance and the understandability assessments, which makes it possible to compare the retrieved documents with the assessments for those same documents, considering relevance and understandability.

The baseline for this evaluation consists in the queries that are currently suggested by the Health Suggestions extension from Google Chrome. For a given query, it is saved the lay and medico-scientific suggestions, both in English and Portuguese, and the 10 best documents associated with each one are retrieved.

5.3.1 Evaluation measures

The evaluation measures used to compare the different methods are: Precision at k (P@k), since this is one of the classic measures adopted in document retrieval tasks; and Normalized discounted cumulative gain at k (nDCG@k), Binary preference (BPref), Rank-biased Precision (RBP) and others based on this one, called understandability-based RBP (uRBP) and uRBP graded (uRBPgr), since these were the measures adopted to evaluate the different submissions to the CLEF eHealth 2018 tasks.

Precision at k (P@k) is the proportion of documents that are relevant to the information need considering the top-k returned documents [Man09]. It can be expressed as:

$$P@k = \frac{\#\ of\ recommended\ items\ @k\ that\ are\ relevant}{\#\ of\ recommended\ items\ @k}$$

Normalized discounted cumulative gain (nDCG) is designed for situations of non-binary notions of relevance, since it can classify a document as partially relevant, which does not happen with precision. It imposes a discount function on the rank position of the documents, while the precision uses a uniform value for all positions. Therefore, more 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@k [JK02].

The binary preference (BPref) measure computes a preference relation of whether judged relevant documents are retrieved ahead of judged irrelevant documents. Thus, it is based on the relative ranks of judged documents only.

Rank-biased Precision (RBP) is a measure that considers that a user will start at the top of the retrieved document list and will proceed to the next document with a probability p, or finish the search with probability 1 - p. If the user chooses a document that is considered relevant, the RBP score increases, since the gain factor is 1 if the document is relevant and 0 otherwise. Therefore, RBP is computed as the sum of the probability of examining each relevant document:

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

where $r_i \in [0,1]$ is the relevance judgment of the ith 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 [PZ].

The understandability-based RBP (uRBP) is based on the previous measure, in which the gain factor considers both the relevance and understandability assessments. The assessment values must be binary. The uRBP graded (uRBPgr) allows graded assessment values (i.e. non-binary) for understandability of documents, associating different gains to different levels of understandability [Zuc16].

For each method, two evaluations are made: one considering word-sense disambiguation and one without it. For each one, relevance and understandability are assessed separately using the corresponding assessment files, where P@10, NDCG@10, BPref and RBP are calculated. In order to aggregate these two types of assessments, uRBP and uRBPgr are also calculated. The results for the English and Portuguese test collections are presented below.

P@10, NDCG@10 and BPref were calculated using the trec_eval[11] tool, an evaluation software used in the Text Retrieval Conference. In this tool, it is necessary to specify the assessment file to be considered and the results file. RBP, uRBP and uRBPgr were calculated using the ubire tool[12], developed by the Information Engineering Lab. This tool allows to specify both relevance and understandability assessment files to be considered at the same time, and the results file.

For the measures that consider only binary assessments (P@10, BPref, RBP and uRBP), the evaluation tools consider the non-zero values from the graded assessments as 1, i.e., relevant or understandable, and the zero values as 0, i.e., non-relevant or non-understandable. Since NDCG@10 and uRBPgr support non-binary assessments, the tools use the values from the assessments without any adjustments.

Since the behavior of the metrics is similar for each method, that is, if one of the measures increases, the others usually increase as well, the separated evaluation is done considering the NDCG@10, while the combined evaluation considers uRBPgr, since they are the metrics that support judgments with non-binary values, better illustrating the variations of relevance/understandability.

5.3.2 Results with the English test collection

Using both lay and medico-scientific suggestions generated for the 50 English queries and the retrieved documents for each one, it was used the relevance and understandability assessment files to evaluate those suggestions according to relevance and understandability. The measures described before were used and the results are represented in the next sections. For all tables, it is selected the best value for each column, both for lay and medico-scientific terminology.

5.3.2.1 Relevance

Using the relevance assessment file, that contains judgments between 0-2 about several documents considering the set of queries, and the results file containing the documents retrieved for the suggested queries as well as the corresponding score, P@10, NDCG@10, BPref and RBP were calculated. The results are represented in Table 5.2.

By analyzing Table 5.2 it is possible to conclude that the method that has better performance considering relevance is the MM + CHV-preferred for the lay terminology and the MM + UMLS-preferred for the medico-scientific one. For this specific case, using word-sense disambiguation or not, does not affect the performance of the method.

[11] https://github.com/usnistgov/trec_eval
[12] https://github.com/ielab/ubire

Table 5.2: Relevance evaluation of the methods using the English test collection.

		Without WSD				With WSD			
Terminology	**Method**	**P@10**	**BPref**	**NDCG@10**	**RBP**	**P@10**	**BPref**	**NDCG@10**	**RBP**
Lay	Health Suggestions (Baseline)	0.3500	0.0142	0.3141	0.3364	0.3500	0.0142	0.3141	0.3364
	MM + CHV-preferred	**0.6280**	**0.0278**	**0.5464**	**0.5856**	**0.6140**	**0.0266**	**0.5406**	**0.5770**
	MM + All preferred/synonym atoms	0.4460	0.0192	0.3922	0.4107	0.3920	0.0172	0.3428	0.3603
	MM + All Atoms	0.4020	0.0175	0.3613	0.3741	0.4300	0.0184	0.3735	0.4069
	MM + All Atoms + Child Relations	0.1720	0.0076	0.1592	0.1639	0.2620	0.0112	0.2460	0.2520
	MM + All Atoms + Parent Relations	0.2740	0.0127	0.2235	0.2481	0.3420	0.0153	0.2971	0.3162
	MM + All Atoms + Same Relations	0.4340	0.0178	0.3688	0.3940	0.4460	0.0188	0.3789	0.4120
	MM + Broader Concepts	0.3460	0.0159	0.3062	0.3223	0.3980	0.0173	0.3486	0.3719
	MM + Narrower Concepts	0.2320	0.0103	0.1851	0.2007	0.3280	0.0137	0.2666	0.2856
Medico-scientific	Health Suggestions (Baseline)	0.3200	0.0130	0.2878	0.3094	0.3200	0.0130	0.2878	0.3094
	MM + UMLS-preferred	**0.5180**	**0.0227**	**0.4537**	**0.4752**	**0.5380**	**0.0233**	**0.4648**	**0.4949**
	MM + All preferred/synonym atoms	0.3820	0.0165	0.3428	0.3743	0.3060	0.0134	0.2771	0.3026
	MM + All Atoms	0.4160	0.0179	0.3569	0.3827	0.3960	0.0168	0.3406	0.3748
	MM + All Atoms + Child Relations	0.1460	0.0071	0.1281	0.1361	0.1960	0.0085	0.1859	0.1915
	MM + All Atoms + Parent Relations	0.2840	0.0126	0.2315	0.2441	0.3340	0.0144	0.2882	0.3014
	MM + All Atoms + Same Relations	0.4120	0.0189	0.3586	0.3810	0.4380	0.0194	0.3851	0.4145
	MM + Broader Concepts	0.2900	0.0133	0.2500	0.2626	0.3520	0.0151	0.2983	0.3180
	MM + Narrower Concepts	0.2440	0.0106	0.1971	0.2127	0.3360	0.0140	0.2738	0.2970
Both	MM + Preferred Atoms	0.4680	0.0209	0.4021	0.4197	0.4180	0.0189	0.3658	0.3967

Since the methods MM + All preferred/synonym atoms, MM + All atoms, MM + All Atoms + Same Relations and MM + Preferred Atoms also outperformed the baseline, it is possible to say that they are useful to retrieve relevant documents, although not as relevant as when using the highlighted method.

The methods that suggest too specific terms (MM + All Atoms + Child Relations and MM + Narrower Concepts) are the ones that retrieve less relevant documents, with a large difference from baseline. However, this difference seems to be reduced by using word-sense disambiguation.

These conclusions can be applied to both terminologies, although the medico-scientific one has slightly worse performance.

5.3.2.2 Understandability

Considering the understandability assessments, that vary between 0-10 to express users' understanding of a certain document, and the ranked results for the suggested queries, the same measures were calculated and are presented in Table 5.3.

In a quick analysis of Table 5.3 it is possible to see that the performance values of NDCG@10 taking into account the understandability are slightly lower compared to the ones that consider only relevance.

MM + CHV/UMLS-preferred is, again, the best method to generate suggestions in both terminologies, with very similar results using word-sense disambiguation or not.

MM + All Atoms + Child Relations, MM + All Atoms + Parent Relations, MM + Broader Concepts and MM + Narrower Concepts are, again, the methods that did not outperform the baseline, which means that they are also not useful in generating suggestions that increase the level of understandability of the documents to the users.

The understandability values obtained with the medico-scientific suggestions are lower, compared with the ones using lay suggestions, as expected, except the methods that use the parent and the same level relations.

Table 5.3: Understandability evaluation of the methods using the English test collection.

		Without WSD				With WSD			
Terminology	**Method**	**P@10**	**BPref**	**NDCG@10**	**RBP**	**P@10**	**BPref**	**NDCG@10**	**RBP**
Lay	Health Suggestions (Baseline)	0.4520	0.0087	0.1894	0.4360	0.4520	0.0087	0.1894	0.4360
	MM + CHV-preferred	**0.7620**	**0.0147**	**0.3171**	**0.6855**	**0.7720**	**0.0149**	**0.3149**	**0.6935**
	MM + All preferred/synonym atoms	0.5840	0.0113	0.2538	0.5290	0.5340	0.0104	0.2355	0.4835
	MM + All Atoms	0.5340	0.0104	0.2390	0.4831	0.5720	0.0111	0.2534	0.5223
	MM + All Atoms + Child Relations	0.2220	0.0043	0.0953	0.2066	0.3180	0.0061	0.1373	0.3016
	MM + All Atoms + Parent Relations	0.3660	0.0071	0.1445	0.3264	0.4620	0.0089	0.1862	0.4149
	MM + All Atoms + Same Relations	0.5680	0.0111	0.2310	0.5048	0.5940	0.0116	0.2508	0.5365
	MM + Broader Concepts	0.4320	0.0084	0.1824	0.3958	0.5080	0.0099	0.2185	0.4686
	MM + Narrower Concepts	0.3380	0.0064	0.1423	0.3076	0.4400	0.0084	0.1789	0.3987
Medico-scientific	Health Suggestions (Baseline)	0.3980	0.0077	0.1732	0.3830	0.3980	0.0077	0.1732	0.3830
	MM + UMLS-preferred	**0.6180**	**0.0119**	**0.2773**	**0.5607**	**0.6740**	**0.0130**	**0.2951**	**0.6016**
	MM + All preferred/synonym atoms	0.4840	0.0093	0.2152	0.4590	0.4160	0.0080	0.1885	0.3952
	MM + All Atoms	0.5260	0.0120	0.2344	0.4811	0.5140	0.0100	0.2312	0.4790
	MM + All Atoms + Child Relations	0.1900	0.0037	0.0695	0.1673	0.2580	0.0050	0.0971	0.2352
	MM + All Atoms + Parent Relations	0.3940	0.0076	0.1592	0.3387	0.4580	0.0089	0.1889	0.4044
	MM + All Atoms + Same Relations	0.5040	0.0097	0.2424	0.4596	0.5420	0.0105	0.2613	0.4972
	MM + Broader Concepts	0.3820	0.0074	0.1529	0.3491	0.4720	0.0092	0.1881	0.4247
	MM + Narrower Concepts	0.3380	0.0065	0.1387	0.3096	0.4400	0.0085	0.1769	0.4015
Both	MM + Preferred Atoms	0.5980	0.0115	0.2712	0.5342	0.5700	0.0110	0.2661	0.5268

5.3.2.3 Relevance and Understandability combined

The uRBP and uRBPgr were also computed, both for methods that use word-sense disambiguation and for those that do not. These results are available in Table 5.4.

Table 5.4: Relevance and understandability evaluation combined of the methods using the English test collection.

		Without WSD		With WSD	
Terminology	**Method**	**uRBP**	**uRBPgr**	**uRBP**	**uRBPgr**
Lay	Health Suggestions (Baseline)	0.2869	0.1257	0.2869	0.1257
	MM + CHV-preferred	**0.4961**	**0.2372**	**0.4846**	**0.2298**
	MM + All preferred/synonym atoms	0.3618	0.1750	0.3221	0.1558
	MM + All Atoms	0.3189	0.1634	0.3530	0.1763
	MM + All Atoms + Child Relations	0.1460	0.0665	0.2341	0.1001
	MM + All Atoms + Parent Relations	0.1972	0.0948	0.2685	0.1226
	MM + All Atoms + Same Relations	0.3477	0.1645	0.3693	0.1731
	MM + Broader Concepts	0.2852	0.1307	0.3321	0.1525
	MM + Narrower Concepts	0.1617	0.0775	0.2397	0.1121
Medico-scientific	Health Suggestions (Baseline)	0.2610	0.1167	0.2610	0.1167
	MM + UMLS-preferred	**0.4155**	**0.2073**	**0.4280**	**0.2122**
	MM + All preferred/synonym atoms	0.3164	0.1610	0.2510	0.1279
	MM + All Atoms	0.3381	0.1690	0.3269	0.1634
	MM + All Atoms + Child Relations	0.1192	0.0531	0.1772	0.0716
	MM + All Atoms + Parent Relations	0.2031	0.0999	0.2655	0.1231
	MM + All Atoms + Same Relations	0.3480	0.1753	0.3719	0.1873
	MM + Broader Concepts	0.2273	0.1042	0.2798	0.1243
	MM + Narrower Concepts	0.1774	0.0828	0.2547	0.1188
Both	MM + Preferred Atoms	0.3662	0.1905	0.3469	0.1754

Since these values are the result of the combination of the assessment of relevance and understandability, they are a good measure to evaluate the overall performance of the methods. After analyzing those same values, it is possible to say that they corroborate the conclusions drawn before.

As expected, the method that returns the most relevant and understandable documents is the MM + CHV-preferred for the lay suggestions and the MM + UMLS-preferred for the medico-scientific suggestions, outperforming the baseline.

Other methods, such as MM + All preferred/synonym atoms, MM + All atoms, MM + All Atoms + Same Relations and MM + Preferred Atoms also retrieve documents that are more relevant or understandable than the ones retrieved using the baseline, but not as good as the highlighted method.

It is possible to see that using word-sense disambiguation helps to retrieve more relevant and understandable documents when the methods use lots of different atoms and their relations. When methods use preferred atoms, the word-sense disambiguation is not very helpful.

By suggesting too specific suggestions, like the ones from MM + All Atoms + Child Relations or MM + Narrower Concepts, the relevance and understandability is relatively lower comparing with suggestions with broader terms (MM + All Atoms + Parent Relations or MM + Broader Concepts) or terms related with the concepts of the suggestions (MM + All Atoms + Same Relations). In these methods, that explore the relations between concepts, only the latter revealed improvements compared to the baseline.

5.3.3 Results with the Portuguese test collection

To complement the evaluation with the English test collection, the lay and medico-scientific suggestions associated with the 200 queries of the Portuguese test collection were also used to verify if there were large variations regarding the performance of the methods to suggest new queries. For this purpose, both relevance and understandability assessments for this collection, as well as the ranking results file were used to evaluate the methods, resulting in the tables present in the next sections. For all tables, it is selected the best value for each column, both for lay and medico-scientific terminology.

5.3.3.1 Relevance

In Table 5.5 are presented the values of P@10, NDCG@10, BPref and RBP considering the relevance assessments and the ranking results file for the suggested queries. The assessment values range between 0-2.

With the Portuguese test collection, the MM + CHV/UMLS-preferred method is, once again, the best to generate suggestions in both terminologies considering relevance.

For all methods, using the word-sense disambiguation did not seem to improve the performance considerably.

Considering the lay terminology, only the MM + All Atoms + Child Relations method was the one that did not have better results than the baseline. In the medico-scientific terminology, all the methods outperformed the baseline.

Comparing the lay and medico-scientific methods, it is possible to say that they all have very similar performance in retrieving documents, considering relevance.

Table 5.5: Relevance evaluation of the methods using the Portuguese test collection.

		Without WSD				With WSD			
Terminology	**Method**	**P@10**	**BPref**	**NDCG@10**	**RBP**	**P@10**	**BPref**	**NDCG@10**	**RBP**
Lay	Health Suggestions (Baseline)	0.0780	0.0217	0.0673	0.0747	0.0780	0.0217	0.0673	0.0747
	MM + CHV-preferred	**0.1865**	**0.0576**	**0.1619**	**0.1793**	**0.1840**	**0.0575**	**0.1608**	**0.1777**
	MM + All preferred/synonym atoms	0.1410	0.0453	0.1205	0.1324	0.1455	0.0493	0.1282	0.1391
	MM + All Atoms	0.1355	0.0437	0.1171	0.1240	0.1455	0.0462	0.1271	0.1343
	MM + All Atoms + Child Relations	0.0655	0.0241	0.0566	0.0627	0.0730	0.0273	0.0647	0.0702
	MM + All Atoms + Parent Relations	0.1275	0.0392	0.1042	0.1178	0.1230	0.0405	0.0990	0.1097
	MM + All Atoms + Same Relations	0.1445	0.0438	0.1221	0.1373	0.1490	0.0461	0.1193	0.1313
	MM + Broader Concepts	0.1395	0.0437	0.1145	0.1316	0.1470	0.0454	0.1195	0.1417
	MM + Narrower Concepts	0.1085	0.0320	0.0932	0.1015	0.1295	0.0387	0.1149	0.1254
Medico-scientific	Health Suggestions (Baseline)	0.0760	0.0205	0.0644	0.0715	0.0760	0.0205	0.0644	0.0715
	MM + UMLS-preferred	**0.1865**	**0.0532**	**0.1377**	**0.1510**	**0.1745**	**0.0561**	**0.1493**	**0.1640**
	MM + All preferred/synonym atoms	0.1355	0.0434	0.1213	0.1248	0.1360	0.0425	0.1146	0.1274
	MM + All Atoms	0.1165	0.0378	0.0973	0.1028	0.1270	0.0374	0.1041	0.1128
	MM + All Atoms + Child Relations	0.0835	0.0245	0.0694	0.0797	0.1000	0.0305	0.0859	0.0946
	MM + All Atoms + Parent Relations	0.1290	0.0358	0.1080	0.1204	0.1325	0.0402	0.1088	0.1178
	MM + All Atoms + Same Relations	0.1235	0.0420	0.1096	0.1169	0.1235	0.0390	0.1022	0.1071
	MM + Broader Concepts	0.1310	0.0423	0.1114	0.1267	0.1525	0.0472	0.1322	0.1491
	MM + Narrower Concepts	0.1100	0.0330	0.0938	0.1026	0.1310	0.0386	0.1167	0.1275
Both	MM + Preferred Atoms	0.1720	0.0530	0.1497	0.1667	0.1780	0.0547	0.1520	0.1721

5.3.3.2 Understandability

Considering the understandability assessments, the same measures were calculated and are presented in Table 5.6. For this specific case, the values of the understandability assessments vary between 0-2.

Table 5.6: Understandability evaluation of the methods using the Portuguese test collection.

		Without WSD				With WSD			
Terminology	**Method**	**P@10**	**BPref**	**NDCG@10**	**RBP**	**P@10**	**BPref**	**NDCG@10**	**RBP**
Lay	Health Suggestions (Baseline)	0.1215	0.0161	0.1047	0.1177	0.1215	0.0161	0.1047	0.1177
	MM + CHV-preferred	**0.3505**	**0.0491**	**0.3110**	**0.3335**	**0.3305**	**0.0464**	**0.2932**	**0.3130**
	MM + All preferred/synonym atoms	0.2425	0.0340	0.2149	0.2311	0.2535	0.0360	0.2235	0.2440
	MM + All Atoms	0.2335	0.0331	0.2005	0.2173	0.2535	0.0358	0.2236	0.2416
	MM + All Atoms + Child Relations	0.1205	0.0173	0.1018	0.1098	0.1390	0.0199	0.1226	0.1299
	MM + All Atoms + Parent Relations	0.2215	0.0308	0.1897	0.2111	0.2285	0.0318	0.1977	0.2142
	MM + All Atoms + Same Relations	0.2460	0.0337	0.2125	0.2384	0.2485	0.0349	0.2117	0.2289
	MM + Broader Concepts	0.2510	0.0350	0.2148	0.2359	0.2710	0.0378	0.2328	0.2549
	MM + Narrower Concepts	0.1735	0.0242	0.1484	0.1669	0.2120	0.0294	0.1902	0.2081
Medico-scientific	Health Suggestions (Baseline)	0.1150	0.0154	0.0968	0.1100	0.1150	0.0154	0.0968	0.1100
	MM + UMLS-preferred	**0.2960**	**0.415**	**0.2498**	**0.2688**	**0.3045**	**0.0426**	**0.2607**	**0.2823**
	MM + All preferred/synonym atoms	0.2300	0.0330	0.2036	0.2200	0.2330	0.0331	0.2049	0.2224
	MM + All Atoms	0.2040	0.0291	0.1727	0.1855	0.2325	0.0330	0.2023	0.2180
	MM + All Atoms + Child Relations	0.1390	0.0195	0.1164	0.1335	0.1655	0.0232	0.1410	0.1551
	MM + All Atoms + Parent Relations	0.2235	0.0313	0.1897	0.2113	0.2330	0.0330	0.1984	0.2173
	MM + All Atoms + Same Relations	0.2270	0.0321	0.1952	0.2167	0.2235	0.0320	0.1847	0.2011
	MM + Broader Concepts	0.2350	0.0326	0.2055	0.2200	0.2730	0.0376	0.2424	0.2596
	MM + Narrower Concepts	0.1785	0.0247	0.1502	0.1688	0.2145	0.0299	0.1910	0.2087
Both	MM + Preferred Atoms	0.2955	0.0408	0.2485	0.2774	0.2975	0.0410	0.2561	0.2842

The method that returns documents with better understandability is the same, MM + CHV/UMLS-preferred.

Once again, only the MM + All Atoms + Child Relations was the only method that did not outperformed the baseline in the lay terminology, while in the medico-scientific terminology all methods were better than the baseline.

The word-sense disambiguation seems to improve the results of the MM + Broader/Narrower Concepts, both in lay and medico-scientific terminology.

In general, the methods also have similar performance when comparing the suggestions from both terminologies. Although, comparing the understandability and relevance values, it is possible to say that the methods are capable of retrieving more understandable documents than relevant ones, which did not happen in the previous test collection.

5.3.3.3 Relevance and Understandability combined

By combining the relevance and understandability assessments in the uRBP and uRBPgr measures in Table 5.7 it is possible to get an overall performance of the methods. After analyzing those same values, it is possible to say that they corroborate the conclusions drawn before.

Table 5.7: Relevance and understandability evaluation combined of the methods using the Portuguese test collection.

		Without WSD		With WSD	
Terminology	**Method**	**uRBP**	**uRBPgr**	**uRBP**	**uRBPgr**
Lay	Health Suggestions (Baseline)	0.0404	0.0567	0.0404	0.0567
	MM + CHV-preferred	**0.1237**	**0.1496**	**0.1258**	**0.1500**
	MM + All preferred/synonym atoms	0.0878	0.1070	0.0879	0.1102
	MM + All Atoms	0.0813	0.1006	0.0888	0.1091
	MM + All Atoms + Child Relations	0.0400	0.0507	0.0466	0.0578
	MM + All Atoms + Parent Relations	0.0705	0.0918	0.0675	0.0870
	MM + All Atoms + Same Relations	0.0919	0.1126	0.0910	0.1096
	MM + Broader Concepts	0.0839	0.1056	0.0928	0.1147
	MM + Narrower Concepts	0.0590	0.0772	0.0801	0.0999
Medico-scientific	Health Suggestions (Baseline)	0.0385	0.0537	0.0385	0.0537
	MM + UMLS-preferred	**0.0969**	**0.1214**	**0.108**	**0.1334**
	MM + All preferred/synonym atoms	0.0821	0.1022	0.0823	0.1020
	MM + All Atoms	0.0628	0.0805	0.0716	0.0894
	MM + All Atoms + Child Relations	0.0496	0.0632	0.0573	0.0744
	MM + All Atoms + Parent Relations	0.0715	0.0933	0.0715	0.0922
	MM + All Atoms + Same Relations	0.0721	0.0932	0.0673	0.0860
	MM + Broader Concepts	0.0848	0.1031	0.1017	0.1228
	MM + Narrower Concepts	0.0597	0.0781	0.0803	0.1012
Both	MM + Preferred Atoms	0.1011	0.1295	0.1084	0.1369

MM + CHV/UMLS-preferred is the preferred method to generate suggestions capable of retrieving more relevant and understandable documents.

For the Portuguese language, almost every method outperformed the baseline, with the exception of the MM + All Atoms + Child Relations for the lay terminology.

Comparing the methods that use word-sense disambiguation with the ones that did not, it is possible to conclude that there is a very little variation in the performance, which did not happen in the previous test collection. Since these ones use more queries, it is possible that the performance tends to normalize.

As in the previous evaluation, the best methods are the ones that use the preferred terms. However, using same level concepts proved to be useful in the lay terminology, while using broader terms in medico-scientific terminology also improved the results, since these methods are also in the top-ranked ones.

5.3.4 Results discussion

For both English and Portuguese test collections, the best methods to generate suggestions are the ones that consider the preferred atoms of the different vocabularies from the UMLS, mainly the CHV.

It has been observed that using child relations does not help to return more relevant or understandable documents, because the generated query may be so specific that it does not satisfy the information need of a certain user. On the other hand, using broader terms (parent and broader relations) proved to be more useful, since other designations for the same concept are being explored.

In the case of the English test collection, the use of word-sense disambiguation does not improve the performance of the methods that use terms considered preferred in the UMLS vocabularies, but is useful when exploring relations. As in the Portuguese test collection, the values do not vary as much, the overall conclusion is that when it is intended the use of the preferred atoms of some vocabulary associated with a concept, it is preferable not to use word-sense disambiguation. On the contrary, when it is intended the exploration of the various relations, it is preferable to use it, because context can be important for relations.

For all methods, the total number of processed atoms (PA), the selected atoms (SA) and the time associated with their identification and retrieval (TIR) were calculated. When necessary, the time taken to classify a concept (TCC) as lay or medico-scientific has been calculated. The total time (TT) corresponds to the sum of these times for all queries, and for a better idea of the time needed to formulate only one suggestion, this value was divided by the total number of queries, represented by TQ. All these times are in seconds. The values for the English test collection are in Table 5.8 and in Table 5.9 for the Portuguese one.

Table 5.8: Time used by the methods to generate suggestions using the 50 English queries.

	Without WSD						**With WSD**					
Method	**PA**	**SA**	**TIR (s)**	**TCC (s)**	**TT (s)**	**TQ (s)**	**PA**	**SA**	**TIR (s)**	**TCC (s)**	**TT (s)**	**TQ (s)**
MM + CHV/UMLS-preferred	115	115	21	-	**21**	**0.42**	115	115	17	-	**17**	**0.34**
MM + Preferred Atoms	115	115	67	-	67	1.34	115	115	54	-	54	1.08
MM + All preferred/synonym atoms	2707	2707	167	1360	1527	30.54	2206	2206	141	1145	1286	25.72
MM + All Atoms	4512	4512	183	2318	2501	50.02	3676	3676	165	1915	2080	41.6
MM + All Atoms + Child Relations	66531	878	5137	438	5575	111.5	49178	666	4643	322	4965	99.3
MM + All Atoms + Parent Relations	66531	2150	7736	1036	8772	175.44	49178	1666	5913	823	6736	134.72
MM + All Atoms + Same Relations	66531	1229	5743	624	6367	127.34	49178	976	4527	509	5036	100.72
MM + Broader Concepts	3258	1557	92	779	871	17.42	2343	1111	88	452	540	10.8
MM + Narrower Concepts	3258	239	87	112	199	3.98	2343	202	88	95	183	3.66

Table 5.9: Time used by the methods to generate suggestions using the 200 Portuguese queries.

	Without WSD						**With WSD**					
Method	**PA**	**SA**	**TIR (s)**	**TCC (s)**	**TT (s)**	**TQ (s)**	**PA**	**SA**	**TIR (s)**	**TCC (s)**	**TT (s)**	**TQ (s)**
MM + CHV/UMLS-preferred	635	635	44	-	**44**	**0.22**	635	635	51	-	**51**	**0.26**
MM + Preferred Atoms	635	635	304	-	304	1.52	635	635	290	-	290	1.45
MM + All preferred/synonym atoms	13931	13931	743	6976	7719	38.59	11251	11251	721	5789	6510	32.55
MM + All Atoms	23015	23015	902	12068	12970	64.85	18356	18356	860	8536	9396	46.98
MM + All Atoms + Child Relations	410991	4586	37875	2320	40195	200.98	306591	3631	39587	1871	41458	207.29
MM + All Atoms + Parent Relations	410991	27665	50045	14258	64303	321.52	306591	21960	39377	10724	50101	250.51
MM + All Atoms + Same Relations	410991	4711	49172	2316	51488	257.44	306951	3776	39181	1941	41122	205.61
MM + Broader Concepts	19870	6660	479	3178	3657	18.29	18324	7145	482	3726	4208	21.04
MM + Narrower Concepts	19870	1355	469	695	1164	5.82	18324	1120	480	603	1083	5.42

In both cases, it is possible to see that the methods that analyze the relationships for each atom take a longer time compared to the others.

The use of the relations from concepts should be preferred, since it takes less time to process them and the performance is more or less equal.

With the Portuguese test collection, the time elapsed for the methods that use the relations of the atoms is slightly higher compared with the English one, but in the other methods the time is very similar. The main reason for this is due to the fact that the queries of the Portuguese collection are more extensive than the English ones, which means that, when identifying more atoms, the number of relations increases proportionally.

Overall, the use of word-sense disambiguation helps to reduce the time elapsed in the methods, since fewer atoms are retrieved, and it is necessary to perform fewer calculations to assign a terminology to an atom.

The MM + CHV/UMLS-preferred is the method that takes less time to retrieve the suggestions, since it only needs to identify the concept and retrieve the corresponding CHV/UMLS-preferred expression.

5.4 Conclusions

Overall, it is clear that the best method to suggest queries in both lay and medico-scientific terminology, in English and Portuguese, is the MM + CHV/UMLS-preferred, since it was the one that outperformed the baseline, as well as every other method, considering relevance and understandability. It is also the method that retrieves suggestions more quickly.

Since the word-sense disambiguation reduces the time that is necessary to generate new suggestions, and slightly improve or does not affect the overall performance, it is concluded that, normally, it should be used. However, for the best method, sometimes the performance is reduced, so it should not be used.

Chapter 6

Implementation of the best strategy in the Health Suggestions extension

After evaluating all the implemented methods for the generation of suggestions in both terminologies, we decided to update the Health Suggestions extension to use this best method. In this chapter, we give a brief description of the extension before our update, as well as the required steps followed to implement the strategy and the differences between the two versions.

6.1 Extension status before the update

As briefly described in Section 2.3, the Health Suggestions consists in an extension for Google Chrome that, for a given query introduced by the user in Google, Bing or Yahoo!, retrieves suggestions in lay and medico-scientific terminologies, both in English and Portuguese. These suggestions are provided in a panel, in the bottom right corner of the browser as shown in Figure 6.1. If one of these suggestions is clicked, a new search in the same search engine is performed. By clicking on the search engines' icons, a new search is also performed with the initial query. The panel is only visible if the extension is able to provide any suggestion for the given query.

The extension also contains a settings page as shown in Figure 6.2, that can be accessed by clicking on the blue heart icon in the panel. There, the user can enable/disable the extension; allow the logging of his interactions; define if the database is local or remote and define the queries' language or if the extension should automatically detect it.

The extension is composed by two distinct modules: suggestion engine and logging engine.

The suggestion engine is based on the English version of the Consumer Health Vocabulary and its corresponding translation to Portuguese. Using those, two inverted indices were created, associating each stemmed term to an inverse string frequency and a postings lists, that is, the list of CHV strings where it appears. The vocabulary of terms is created with the tokens from the CHV's

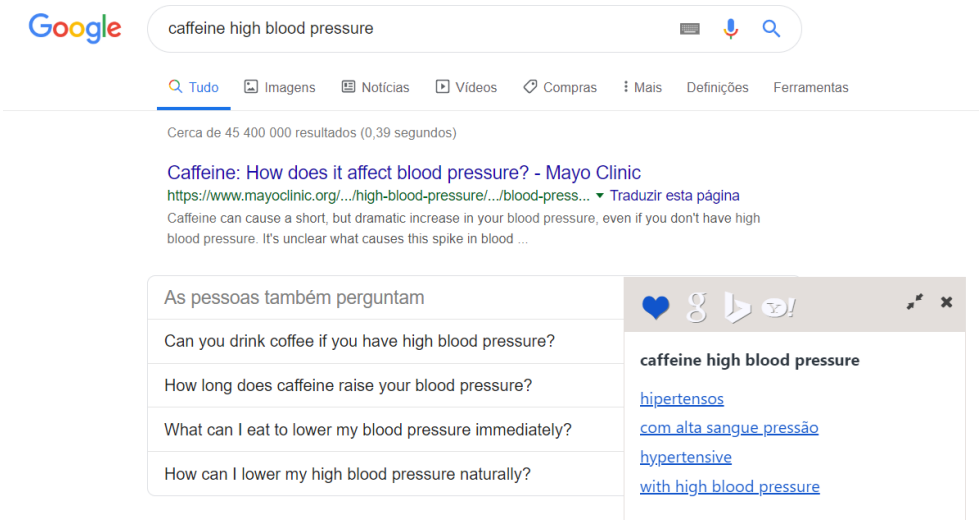


Figure 6.1: Suggestions' panel before the update.

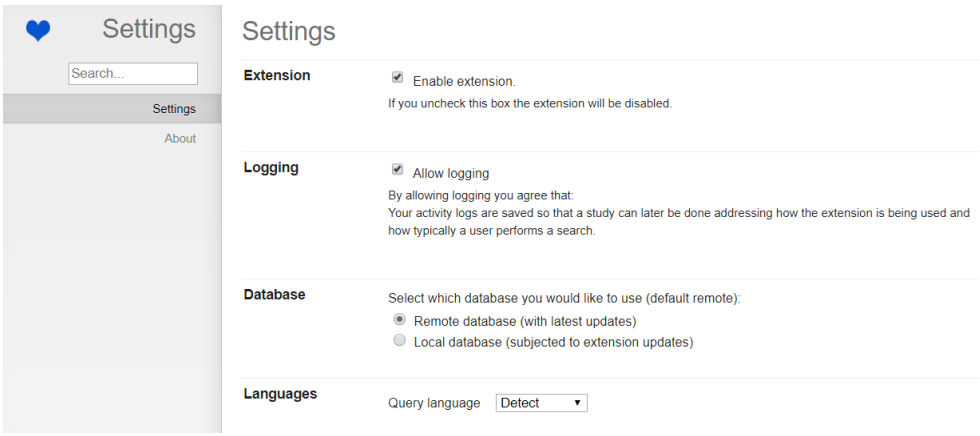


Figure 6.2: Options page before the update.

strings after removing the stop words. The terms were stemmed, reduced to lower case and the diacritics were removed [LF16].

Given a query introduced by the user, the engine detects the language, removes the diacritics and stop words and stems it. After that, it is computed the sum of the inverse string frequency to all the terms included in the query for each CHV string. The CUI associated with the string with maximum score is used to retrieve the CHV/UMLS-preferred terms, that are used as lay and medico-scientific suggestions. This process was previously described in Figure 2.4.

The logging engine is responsible for tracking users' actions in order to understand their search process, saving information such as: query, search engine, search results, the suggestions provided, clicks on the suggestions, time spent on web pages, among others.

6.2 Extension status using the best strategy

After the evaluation of the implemented strategies in the previous chapter, we concluded that the best method to generate suggestions in lay and medico-scientific terminologies uses the MetaMap tool to identify the different biomedical concepts included in the query, retrieving the CHV/UMLS-preferred expressions for each of the identified concepts.

Since the method that is currently implemented in the extension only uses the concept associated with the highest score string, it discards other possible ones included in the query. With the use of MetaMap, it is possible to solve this problem, since the tool is able to identify the different concepts included in the query.

Thus, the first step followed to implement the best method was the installation of MetaMap on the server. Since this tool is able to do lexical/syntactic analysis, as previously described in Section 2.5.1.1, the process of removing the diacritics, stop words and stemming the query was removed from the extension and is done by MetaMap. Since we concluded in Section 5.2 that MetaMap only works for English text, any query introduced by the user is translated into English using the Google Translator API, ensuring that the biomedical concepts included in the query are always identified, since they are always in English. Thus, the identification of the query's language by the extension was also removed, since this process is sensitive to failures and, sometimes, MetaMap might not identify the concepts.

The SQL database containing the CHV was updated, since a more recent version had already been released. This database contains the CHV/UMLS-preferred expressions to all the concepts of CHV, both in English and Portuguese.

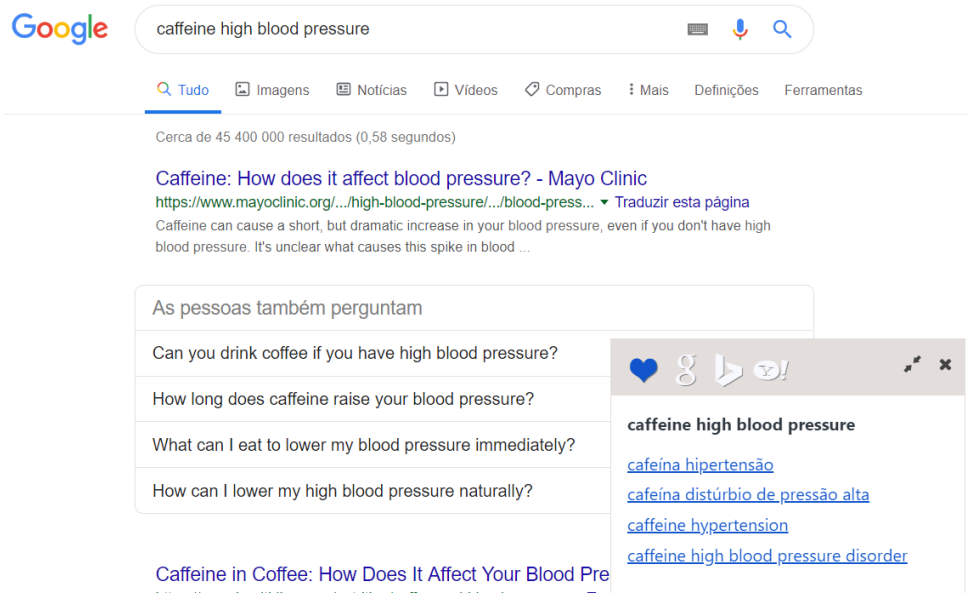


Figure 6.3: Suggestions' panel before the update.

For each identified concept included in the query, the corresponding CHV/UMLS-preferred

expressions are retrieved, both in English and Portuguese. The concatenation of these expressions forms the lay and medico-scientific suggestions, respectively, as can be seen in Figure 6.3. Comparing the suggestions from this figure with the ones from Figure 6.1 it is possible to see that the new suggestions are more precise since they are related with more than one concept.

Since the use of MetaMap always requires the use of the server, the option of using a local database was removed. Since it only detects concepts from English text and the way we solved it was always translating the query into English, the option of defining the query's language or the automatic detection was also removed. These changes in the settings page can be seen in Figure 6.4.

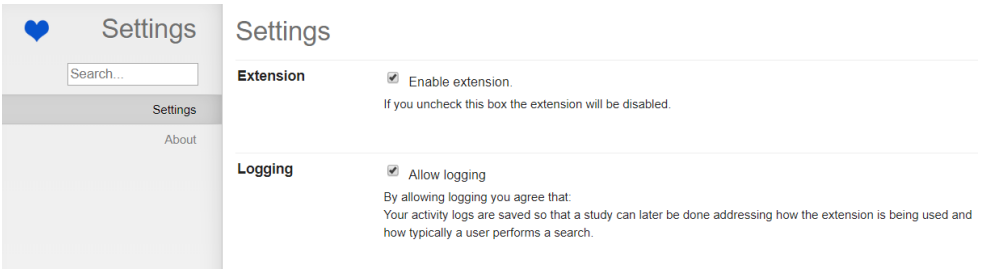


Figure 6.4: Options page before the update.

The new version of the extension is available to download on the Health Suggestions web page[1], as well as the instructions for installing this extension in Google Chrome. To access this web page and use the extension it is necessary to be connected to the FEUP network, which can be done through a VPN connection.

[1] http://healthsuggestions.fe.up.pt/

Chapter 7

Conclusions and Future Work

In this chapter, the main conclusions of the conducted work are detailed, as well as if the developed solution met the initial goals. Some possible improvements are also presented, that can be implemented in future work.

7.1 Conclusions

In this dissertation, the need to improve the patients' health management was shown, since some users face difficulties to meet their information needs with the increasing number of accessible information online. A health technology, the Health Suggestions extension for Google Chrome, serves as the basis of this work. This extension suggests queries in lay and medico-scientific terminology in English and Portuguese, since it is proved that these suggestions help the lay users in retrieving more relevant and understandable documents to satisfy their needs.

In order to improve this extension, query formulation techniques were explored, principally by analyzing works in the health information retrieval domain. Since the queries submitted by the users normally have different concepts, tools and techniques of concept recognition were also studied and a tool was chosen to be used for this purpose. For the basis of query formulation, different resources were analyzed, namely vocabularies and the existing relationships between the concepts.

By using the different names associated to those concepts, a problem arose: the identification of the terminology to which a concept belonged. In order to solve this problem, an approach to automatically classify the concepts in the lay or medico-scientific terminology was developed. This approach used the different Consumer Health Vocabulary expressions to form different subsets to serve as a vocabulary, computing the similarity of an expression with this vocabulary. The best method to do this uses all the expressions of the CHV and considers the difficulty of these expressions, by benefiting the most familiar terms and harm the less familiar. It defines a threshold

from which the expressions are considered lay and if the similarity is lower than that value the expression is considered medico-scientific.

With this problem solved, it was possible to define different methods to generate the suggestions. They all used MetaMap to recognize the terms present in the initial query. Some of these methods used the preferred terms of different vocabularies of the UMLS to generate the suggestions, and they were the ones that had better performance. Others explored broader terms, specific terms and similar terms, but did not retrieved results as good as the previous ones. The best method to suggest lay queries is the one that uses the CHV-preferred expressions of the CHV (the most familiar ones) to substitute the identified concepts by these expressions, and to suggest the medico-scientific ones it uses the UMLS-preferred ones (the less familiar).

The previously developed extension only identified one concept per query, no matter how extensive the query was. For that concept, it retrieved the CHV-preferred and the UMLS-preferred expressions too, as the lay and medico-scientific suggestions, respectively. The results of the evaluation showed that this is still the best method, but that was very important to firstly correctly identify the multiple concepts that could be present in the initial query, that is what the best method does. This method was then implemented in the extension.

In conclusion, the vast majority of the developed methods proved to be better than the baseline, helping the user to retrieve more relevant and understandable documents to satisfy his needs.

7.2 Future work

The method that assigns a terminology to a concept sometimes returns the value 0 as the value of similarity with used vocabulary. This makes it impossible to determine to which terminology the concept belongs to, but as a workaround, we considered this concept as lay. Therefore, this is a problem that could be solved as future work. The use of other strategies such as word embedding could also improve the performance of this method as a way to explore alternatives other than the cosine similarity.

Taking into account the developed methods to generate the suggestions, other relations could be explored, to see if they retrieve better results than the chosen ones. The fact that every time it is needed to use UMLS, it is necessary to make an HTTP request, makes the process quite slow. One of the possible strategies would be to download the entire Metathesaurus, accessing it locally, making the process more efficient.

References

[ALA10] Alan R Aronson, François-michel Lang, and Alan R Aronson. An overview of MetaMap : historical perspective and recent advances An overview of MetaMap : historical perspective and recent advances. pages 229–236, 2010.

[Aro01] A R Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings. AMIA Symposium*, pages 17–21, 2001.

[Art04] Review Articles. Indian Journal of Medical Sciences HEALTH INFORMATION ON THE INTERNET : PATIENT EMPOWERMENT OR PATIENT DECEIT ? 58(8):2000–2002, 2004.

[BMM11] Sumit Bhatia, Debapriyo Majumdar, and Prasenjit Mitra. Query suggestions in the absence of query logs. pages 795–804, 2011.

[Bod04] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32 Databas:D267–70, 2004.

[BSAS] Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. Automatic Query Expansion Using SMART : TREC 3.

[CBB+13] Stefano Ceri, Alessandro Bozzon, Marco Brambilla, Emanuele Della Valle, Piero Fraternali, and Silvia Quarteroni. Web information retrieval. *Web Information Retrieval*, pages 1–284, 2013.

[CCCdR18] Wanyu Chen, Fei Cai, Honghui Chen, and Maarten de Rijke. Personalized query suggestion diversification in information retrieval. *Frontiers of Computer Science*, 2018.

[CMO13] David Campos, Sérgio Matos, and José Luís Oliveira. Gimli: Open source and high-performance biomedical name recognition. *BMC Bioinformatics*, 14(May 2014), 2013.

[CNR08] Guihong Cao, Jian-yun Nie, and Stephen Robertson. Selecting Good Expansion Terms for Pseudo-Relevance Feedback. pages 243–250, 2008.

[CRB12] Claudio Carpineto, Giovanni Romano, and Fondazione Ugo Bordoni. A Survey of Automatic Query Expansion in Information Retrieval. 44(1), 2012.

[EN16] Liana Ermakova and Elena Nikitina. Proximity Relevance Model for Query Expansion. pages 1–6, 2016.

[FD12] Susannah Fox and Maeve Duggan. Health online 2013. *Pew Research Internet Project Report*, 01 2012.

REFERENCES

[GMX+09] James Geller, C Paul Morrey, Junchuan Xu, Michael Halper, Gai Elhanan, Yehoshua Perl, and George Hripcsak. Comparing inconsistent relationship configurations indicating UMLS errors. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2009(November 2014):193–7, 2009.

[GNN+10] Wei Gao, Cheng Niu, Jian-Yun Nie, Ming Zhou, Kam-Fai Wong, and Hsiao-Wuen Hon. Exploiting query logs for cross-lingual query suggestions. *ACM Transactions on Information Systems*, 28(2):1–33, 2010.

[HSBD11] Orla Higgins, Jane Sixsmith, Mm M Barry, and C Domegan. A literature review on health information- seeking behaviour on the web: a health consumer and health professional perspective. *European Centre for Disease Prevention and Control*, 1(October):1–9, 2011.

[JAJ17] Wura Jacobs, Ann O. Amuta, and Kwon Chan Jeon. Health information seeking in the digital age: An analysis of health information seeking behavior among US adults. *Cogent Social Sciences*, 3(1):1–11, 2017.

[JK02] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[JZH09] Shen Jiang, Sandra Zilles, and Robert Holte. Query suggestion by query search: A new approach to user support in web search. *Proceedings - 2009 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2009*, 1:679–684, 2009.

[JZP+18] Jimmy, Guido Zuccon, João R M Palotti, Lorraine Goeuriot, and Liadh Kelly. Overview of the CLEF 2018 Consumer Health Search Task. In *CLEF*, 2018.

[KGB09] Diane Kelly, Karl Gyllstrom, and Earl W Bailey. A Comparison of Query and Term Suggestion Features for Interactive Searching. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 371–378, New York, NY, USA, 2009. ACM.

[KGS+16] Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurélie Névéol, João Palotti, and Guido Zuccon. Overview of the CLEF eHealth Evaluation Lab 2016. In Norbert Fuhr, Paulo Quaresma, Teresa Gonçalves, Birger Larsen, Krisztian Balog, Craig Macdonald, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 255–266, Cham, 2016. Springer International Publishing.

[KTC+07] Alla Keselman, Tony Tse, Jon Crowell, Allen Browne, Long Ngo, and Qing Zeng. Assessing consumer health vocabulary familiarity: an exploratory study. *Journal of medical Internet research*, 9(1):e5, 2007.

[KZ04] Reiner Kraft and Jason Zien. Mining Anchor Text for Query Refinement. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 666–674, New York, NY, USA, 2004. ACM.

[LC07] Zhenyu Liu and Wesley W. Chu. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *Information Retrieval*, 10(2):173–202, 2007.

[LF16] Carla Teixeira Lopes and Tiago Almeida Fernandes. Health suggestions: A chrome extension to help laypersons search for health information. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9822 LNCS:241–246, 2016.

[LGT+08] Luo, Gang Gang, Tang, Chunqiang Chunqiang, Yang, Hao Hao, Wei, and Xing Xing. MedSearch: a specialized search engine for medical information retrieval. *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 143–152, 2008.

[LND08] K.W.-T. Leung, W. Ng, and Dik Lun Lee. Personalized Concept-Based Clustering of Search Engine Queries. *IEEE Transactions on Knowledge and Data Engineering*, 20(11):1505–1518, 2008.

[LR16] Carla Teixeira Lopes and Cristina Ribeiro. Effects of Language and Terminology on the Usage of Health Query Suggestions. pages 83–95, 2016.

[LR18] Carla Teixeira Lopes and Cristina Ribeiro. Effects of Language and Terminology of Query Suggestions on the Precision of Health Searches. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric SanJuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 101–111, Cham, 2018. Springer International Publishing.

[Man09] Christopher D. Manning. Intro to Information Retrieval. *Information Retrieval*, (c):1–18, 2009.

[Men14] Lingling Meng. A Survey on Query Suggestion. *International Journal of Hybrid Information Technology*, 7(6):43–56, 2014.

[MH13] Diana Lynn MacLean and Jeffrey Heer. Identifying medical terms in patient-authored text: A crowdsourcing-based approach. *Journal of the American Medical Informatics Association*, 20(6):1120–1127, 2013.

[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.

[OMQL15] J Ooi, Xiuqin Ma, Hongwu Qin, and S C Liew. A survey of query expansion, query suggestion and query refinement techniques. In *2015 4th International Conference on Software Engineering and Computer Systems (ICSECS)*, pages 112–117, 2015.

[oMUM] National Library of Medicine (US) and Bethesda (MD). *UMLS Reference Manual*.

[Ona19] A Onan. Semantic query suggestion based on optimized random forests. *Advances in Intelligent Systems and Computing*, 764:91–102, 2019.

[ooPN95] Commission of the European Communities. Directorate-General for Industry (DGIII), Heymans Institute of Pharmacology, and Mercator Network. *Multilingual glossary of technical and popular medical terms in nine European languages*. Heymans Institute of Pharmacology, 1995.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,

M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[PW91] Helen J Peat and Peter Willett. The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems. 42(5):378–383, 1991.

[PZ] Laurence A. F. Park and Yuye Zhang. On the distribution of user persistence for rank-biased precision.

[RF18] Victoria Rideout and Susannah Fox. Digital Health Practices, Social Media Use, and Mental Well-Being Among Teens and Young Adults in the U.S. pages 1–95, 2018.

[Shu10] Nakatani Shuyo. Language detection library for java, 2010.

[SIM14] Isabelle Stanton, Samuel Ieong, and Nina Mishra. Circumlocution in diagnostic medical queries. 2014.

[SL16] Ricardo Silva and Carla Teixeira Lopes. The effectiveness of query expansion when searching for health related content: Infolab at clef ehealth 2016. In *CLEF*, 2016.

[STZ05] Xuehua Shen, Bin Tan, and ChengXiang Zhai. Context-sensitive Information Retrieval Using Implicit Feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 43–50, New York, NY, USA, 2005. ACM.

[SYYT+16] Luca Soldaini, Andrew Yates, Elad Yom-Tov, Ophir Frieder, and Nazli Goharian. Enhancing web search in the medical domain via query clarification. *Information Retrieval*, 19(1-2):149–173, 2016.

[TC03] Koichi Takeuchi and Nigel Collier. Bio-medical Entity Extraction Using Support Vector Machines. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13*, BioMed '03, pages 57–64, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[TCS10] Michael Tanenblatt, Anni Coden, and Igor Sominsky. The ConceptMapper Approach to Named Entity Recognition. *Lrec*, pages 546–551, 2010.

[TML+16] Eugene Tseytlin, Kevin Mitchell, Elizabeth Legowski, Julia Corrigan, Girish Chavan, and Rebecca S. Jacobson. NOBLE - Flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinformatics*, 17(1):1–15, 2016.

[TPR17] Carla Teixeira Lopes, Dagmara Paiva, and Cristina Ribeiro. Effects of language and terminology of query suggestions on medical accuracy considering different user characteristics. *Journal of the Association for Information Science and Technology*, 68(9):2063–2075, 2017.

[Vec03] Olga Vechtomova. Query Expansion with Long-Span Collocates. pages 251–273, 2003.

[VMHZ14] V G Vinod Vydiswaran, Qiaozhu Mei, David A Hanauer, and Kai Zheng. Mining consumer health vocabulary from community-generated text. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2014:1150–9, 2014.

REFERENCES

[VWSG97] Bienvenido Vélez, Ron Weiss, Mark A Sheldon, and David K Gifford. Fast and Effective Query Refinement. *SIGIR Forum*, 31(SI):6–15, 1997.

[ZCM+03] Qinghua Zou, Wesley W Chu, Craig Morioka, Gregory H Leazer, and Hooshang Kangarloo. IndexFinder: a method of extracting key concepts from clinical texts for indexing. *Annual Symposium of the American Medical Informatics Association*, (February):763–7, 2003.

[ZCP+06] Qing T. Zeng, Jonathan Crowell, Robert M. Plovnick, Eunjung Kim, Long Ngo, and Emily Dibble. Assisting consumer health information retrieval with query recommendations. *Journal of the American Medical Informatics Association*, 13(1):80–90, 2006.

[Zie03] Rita D. Zielstorff. Controlled vocabularies for consumer health. *Journal of Biomedical Informatics*, 36(4-5):326–333, 2003.

[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. *International Journal of Medical Informatics*, 73(1):45–55, 2004.

[ZL11] Michael Zarro and Xia Lin. Using Social Tags and Controlled Vocabularies As Filters for Searching and Browsing: A Health Science Experiment. *The Fifth Workshop on Human-Computer Interaction and Information Retrieval*, (January):1–4, 2011.

[ZLXX16] Nana Zhu, Xueting Li, Lei Xiong, and Han Xue. Query Rewriting for Archived Information Retrieval. In *Proceedings of the International Conference on Internet Multimedia Computing and Service*, ICIMCS'16, pages 323–326, New York, NY, USA, 2016. ACM.

[ZT06] Qing T Zeng and Tony Tse. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association : JAMIA*, 13(1):24–29, 2006.

[ZTC+05] Qing T Zeng, Tony Tse, Jon Crowell, Guy Divita, Laura Roth, and Allen C Browne. Identifying consumer-friendly display (CFD) names for health concepts. *Proceeding of the AMIA Annual Symposium*, pages 859–63, 2005.

[ZTD+07] Qing T. Zeng, Tony Tse, Guy Divita, Alla Keselman, Jon Crowell, Allen C. Browne, Sergey Goryachev, and Long Ngo. Term identification methods for consumer health vocabulary development. *Journal of Medical Internet Research*, 9(1), 2007.

[Zuc16] Guido Zuccon. Understandability biased evaluation for information retrieval. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Advances in Information Retrieval*, pages 280–292, Cham, 2016. Springer International Publishing.

[ZZZ+14] Hai-tao Zheng, Jie Zhao, Yi-chi Zhang, Yong Jiang, and Shu-tao Xia. An Ontology-Based Approach to Query. pages 437–444, 2014.

REFERENCES

Appendix A

Queries used to evaluate the methods

A.1 Queries from CLEF eHealth 2018

151001 anemia diet therapy
152001 emotional and mental disorders
153001 american diabetes association
154001 high blood pressure
155001 infectious disease prevention
156001 food allergy test
157001 orlistat drug profile
158001 radiation health effect
159001 smoking and heart disease
160001 smoking cessation products
161001 omega 3 fatty acids
162001 common medication errors
163001 Anxiety coping skills
164001 health benefits of spirulina
165001 do allergies cause migraines
166001 ketamine and its potential for abuse
167001 head and neck cancer
168001 hiv vaccine phase
169001 normal lab values
170001 rheumatoid arthritis prognosis
171001 breast reduction and lift
172001 involuntary trembling or quivering
173001 drug food interaction
174001 breakfast better student
175001 Voltaren Emugel 1%
176001 pelvic inflammatory disease
177001 breakfast school children
178001 drug addiction organization
179001 resevrse insulin ressitance
180001 sleep apnea syndrome
181001 marijuana's effects on mental health
182001 Delayed onset muscle soreness
183001 what percent of celery is made of water
184001 antiandrogen therapy for prostate cancer
185001 chemicals causing prostate cancer
186001 stem cell hearing loss
187001 How does smoking cause hypertension?
188001 caffeine high blood pressure
189001 long term effects of gonhrrea
190001 privatization of health care in canada
191001 liver function test uses
192001 diabetes mellitus and body fat mass
193001 health effects of ionized water therapy
194001 ankle avulsion fracture healing time
195001 affective treatments for chronic lyme disease
196001 parkinson's disease + titanium link
197001 Four month infant running nose
198001 high protein diet and its human health effects
199001 why is there a minimum drinking age and

what are the consequences of underage drinking?

200001 feeling of fullness with hiccups with a feeling of a lump in the back of the throat

A.2 Queries from a past user study

1 muitas marcas vermelhas nas pernas depois de viajar dos estados unidos
2 inchaço com manchas de sangue no nariz
3 pés vermelhos secos e descamados em crianças
4 caroços com comichão pele
5 chiado e tosse durante o sono + crianças
6 criança com chiadeira a respirar
7 sintomas rosácea
8 córnea turva e problema de visão
9 olhos vermelhos e com comichão
11 manchas brancas na boca
12 bebé tem tosse seca e problemas para engolir saliva
13 pupila turva afetando a visão
14 criança dificuldade em respirar
15 ataque de asma
16 pequenas nódoas vermelhas nas pernas
17 irritação com descamação
18 marcha e equilíbrio fracos e com tremores
19 unha do polegar lesionada
20 dificuldade de movimento com tremor involuntário na mão
21 irritações cutâneas com comichão comuns
22 excreção amarela e espessa no olho
23 olhos vermelhos e injetados de sangue
24 gosma amarela a sair de um olho comichão
25 irritação vermelha rosto de bebé
28 pele de bebé fungos
29 mancha vermelha na pele com bolha e pus seco
30 sons estranhos ao respirar
31 criança com respiração com chiadeira
32 pele vermelha e irritada no rosto do bebé
33 infecção branca na faringe
34 problema de cáries
35 irritação forte com lentes de contato
36 olhos estão a tremer
37 descamação da pele
38 pés com coceira e escamação em crianças
39 eczema cutâneo com bolhas e pus branco
40 bebé mancha vermelha no rosto
41 íris do olho grande
42 hemorragia irregular sub-cutânea
43 saliências com comichão pele
44 unha a ficar escura
45 sensação seca com irritação
48 não consigo parar de mexer os olhos problema de saúde
49 bebé a respirar sempre de boca fechada
50 mancha vermelha rosto do bebé
51 olhos vermelhos vasos sanguíneos
52 manchas brancas na gengiva
53 pernas inchadas
54 caroços vermelhos no nariz
56 moleza ao andar
57 respiração difícil e tosse com chiado em criança
58 39 graus e varicela
59 respiração pesada e com chiadeira
60 bebé pinta branca na íris
61 machucados nas unhas da mão
63 crostas na pele
64 movimento ocular rápido involuntário esquerda-direita
65 manchas castanhas estranhas na pele
66 tratamento de tosse em bebés
67 dente preto
68 hérnia inguinal reparação laparoscópica rede benefícios riscos
69 opções remoção marisca

70 dores de cabeça aliviadas pela doação de sangue
71 tosse incontrolável após infeção posição deitada
72 elevada ingestão de água sede constante causas
73 letargia insónia suplementos vitamínicos riscos
74 menstruação irregular tossir sangue
75 alternativas de medicação metabolismo lento da tiróide
76 espinha branca paciente diálise
77 fibromialgia severa difícil de gerir opção de tratamento
78 tempestade tiroideia sinais sintomas quando atuar
79 dador de órgãos factos versus mitos
80 dor abdominal depois de comer inchado flatulente sangue nas fezes
81 entorse tornozelo sem dor pisado ligeiro tempo de cura
82 paralisia de Bell sinais sintomas causas
83 sangramento anal riscos causas
84 riscos de combinação de Tylenol com Benylin extra forte gripe
85 síndrome compartimental aguda tratamento opções
86 uso de esteróides anabolizantes em pacientes com distrofia muscular
87 hepatite B rastreio análise sangue informação infeção
88 dormência nos dedos dos pés causas e tempo de recuperação
89 dor abdominal no lado esquerdo aliviada pelo vómito
90 efeitos secundários do Effexor
91 Informação sobre doença de Alzheimer
92 tiroidectomia compensação perda de peso
93 dormência nos dedos das mãos mulher fadiga
94 pneumonia riscos preocupações tratamentos opções
95 aumentar inteligência esteróides dieta vitaminas
96 desejos de sal aumento sede
97 dor na parte inferior das costas sangue no sémen dor testicular
98 náusea vómitos SCI explicação sangue no cocó
99 álcool calor dor de cabeça náusea imediatos novo
100 hipertensão feminina cefaleias com cegueira visão dupla causas
101 hipertensão severa gestão correta
102 letargia infantil PCR elevada nódulos linfáticos inchados dor depois de comer
103 jovem mulher com dor no peito constante raio-x pulmões limpo ombro braço direito
104 análises de sangue ao fígado explicação
105 sangue no sémen trauma na próstata
106 pingo no nariz opções de tratamento
107 sangramento do mamilo pós-menopausa
108 uso prolongado de IBP riscos e efeitos secundários
109 aumento da contagem de glóbulos brancos na urina sem sintomas de infeção
110 síndrome de Guillain-Barré informação para pacientes evolução da doença
111 testículo inchado costas doridas náusea vómitos febre
112 infeção da garganta garganta dorida olhos irritados opções de tratamento
113 medicação para as dores dor crónica retrição de opióides explicação
114 alteração na visão ponto desfocado riscos sinais sintomas
115 ereção dolorosa a piorar depois de cateter de Foley
116 ferro elevado dor de cabeça
117 tosse posição deitada
118 efeitos de beber muita água
119 suplemento natural energia acordar
120 tossir sangue escuro
121 depressão tiróide medicação
122 nódulo branco duro ao lado do olho
123 fibromialgia dor medicação
124 tiróide palpitações ansiedade emergência
125 os médicos não te tratam se fores um dador de órgãos

126 viagem diarreia tratamento para as dores
127 entorse tornozelo tempo de recuperação depois de exostectomia
128 como se pode apanhar herpes
129 bacteremia corte anus
130 Benylin e Tylenol juntos
131 síndrome compartimental tratamento sem cirurgia
132 distrofia muscular esteróides anabolizantes
133 hepatite B erro análise sangue
134 equilíbrio efeito secundário Efexor
135 como é que a doença de Alzheimer causa a morte
136 perda de peso hemitiroidectomia
137 fadiga perda de equilíbrio dormência tonturas palpitações causas
138 causas de pneumonia recorrente
139 melhor concentração após mononucleose
140 depressão e cólicas abdominais
141 sangue no esperma
142 infiltração linfocítica do pénis artefacto esmagamento nuclear marcado
143 pericardite alívio dor consumidor de drogas IV
144 dor de estômago, cólicas, enjoos, vómitos, diarreia clonazepam
145 álcool temperatura dor de cabeça enjoo novo início
146 hipertensão mulheres jovens causas e tratamento
147 reduzir pressão arterial muito alta rapidamente
148 criança TSH alta ferro baixo infeções recorrentes
149 dor no peito intensa mulher jovem causas
150 causas de elevação ligeira na ALT
151 sangue no esperma andar de bicicleta
152 tratamento congestão nasal crónica
153 sangue a sair do mamilo
154 Omeprazol efeitos secundários a longo prazo
155 glóbulos brancos na urina sem infecção
156 Síndrome de Guillain-Barré
157 testículo inchado dorido, dor, vómitos
158 dor de garganta constante causas e tratamento
159 probabilidade de serem prescritos opióides
160 dor de cabeça e alterações visuais
161 ereção dolorosa após algália
162 hérnia inguinal taxa de sucesso
163 remoção de mariscas
164 doação de sangue redução dores de cabeça
165 posição deitada tosse
166 alternativa 5 Hour Energy
167 tosse menstruação
168 causas metabolismo lento
169 irregularidades da pele
170 dador de órgãos riscos
171 dor topo estômago paragem movimento intestinal
172 tempo de cura entorse tornozelo
173 transmissão de herpes
174 sangramento anal número dois
175 médicos perto de mim
176 seguro síndrome compartimental aguda
177 distrofia muscular esteróides
178 análise sangue hepatite B reativo
179 dedo pé partido
180 dor abdominal infantil vómito
181 efeito secundário Efexor
182 causa de morte no Alzheimer
183 perda de peso remoção tiróide
184 problemas cardiovasculares
185 pneumonia regular
186 trauma testicular costas
187 pele peniana benigna infiltração linfocítica ligeira com artefacto de esmagamento nuclear marcado
188 analgésico não viciante
189 gastroenterologistas de topo
190 álcool dor de cabeça enjoo
191 hipertensão
192 queda súbita pressão sanguínea alta
193 gânglios linfáticos inchados baixos níveis de energia dor de estômago

Queries used to evaluate the methods

194 tosse dor no peito
195 resultados análise ALT
196 sangue no esperma de bicicleta
197 nariz congestionado
198 fluido de sangue seco mamilo
199 sintomas de Omeprazol
200 ITU
201 síndrome de guillain-barré
202 inchaço escrotal costas doridas dor abdominal enjoos lento
203 cabeça garganta olhos dor
204 prescrição de opiáceos médico
205 olho zona baça lágrimas
206 algália pisado
207 sintomas de intoxicação por água

Queries used to evaluate the methods

Appendix B

Configuration files

B.1 Elasticsearch indexing configuration file

```
{
    "settings": {
      "number_of_shards": 1,
      "number_of_replicas": 0,
      "analysis": {
        "analyzer": {
            "my_english": {
                "tokenizer": "standard",
                "filter": ["lowercase", "terrier_stopwords", "porter_stem"]
            }
        },
        "filter": {
          "terrier_stopwords": {
              "type": "stop",
              "stopwords_path": "stopwords/terrier-stop.txt"
          }
        }
      },
      "similarity": {
        "bm25_content": {
            "type": "BM25",
            "b": 0.75,
            "k1": 1.2
        }
      }
    },
    "mappings": {
      docType: {
        "_source": {
```

```
            "enabled": True
            },
        "properties": {
            "content": {
                "type": "text",
                "similarity": "bm25_content",
                "analyzer": "my_english"
            },
            "is_html": {
                "type": "boolean"
            }
         }
        }
    }
}
```

B.2 FSCrawler's configuration file

```
---
name: "ri"
fs:
  url: "path\\to\\data"
  filename_as_id: true
  update_rate: "15m"
  excludes:
  - "*/~*"
  json_support: false
  add_filesize: true
  remove_deleted: true
  add_as_inner_object: false
  store_source: false
  index_content: true
  attributes_support: false
  raw_metadata: false
  xml_support: false
  index_folders: true
  lang_detect: false
  continue_on_error: false
  ocr:
    language: "eng"
    enabled: true
    pdf_strategy: "ocr_and_text"
elasticsearch:
  nodes:
  - url: "http://127.0.0.1:9200"
```

```
bulk_size: 100
flush_interval: "5s"
byte_size: "10mb"
```