

Understandability and Expertise in Consumer Health Search

Retrieving topically relevant and understandable health information on the Web.

PhD THESIS

submitted in partial fulfillment of the requirements for the degree of

Doctor of Technical Sciences

within the

Vienna PhD School of Informatics

by

MSc. João Palotti

Registration Number 1128475

to the Faculty of Informatics

at the TU Wien

Advisor: Dr. Allan Hanbury

Second advisor: Dr. Guido Zuccon

External reviewers:

Norbert Fuhr. University of Duisburg-Essen, Germany.

Udo Kruschwitz. University of Essex, United Kingdom.

Vienna, 29th March, 2019

João Palotti

Allan Hanbury

Declaration of Authorship

MSc. João Palotti
Address

I hereby declare that I have written this Doctoral Thesis independently, that I have completely specified the utilized sources and resources and that I have definitely marked all parts of the work - including tables, maps and figures - which belong to other works or to the internet, literally or extracted, by referencing the source as borrowed.

Vienna, 29th March, 2019

João Palotti

*Simplicity is a great virtue but it
requires hard work to achieve it
and education to appreciate it.
And to make matters worse:
complexity sells better.*

Edsger W. Dijkstra (1984)

Acknowledgments

I finally reached the end of this long journey in which Information Retrieval was just a small part of all that I lived and learned.

I would like first to thank the people that helped me the most with the intellectual work of this thesis. Allan Hanbury welcomed me to his team when I, barely speaking English, arrived in Austria. Even though he did not need to, he hired me in the Khresmoi project which immensely helped the development of this thesis. Allan always had his door open and plenty of patience to hear and advice on my research ideas and never-ending side-projects.

In the Khresmoi project, I luckily met Lorraine Goeuriot and Liadh Kelly who integrated me into the CLEF eHealth laboratory and with whom I have been collaborating ever since. Because of CLEF, I met Guido Zuccon, whose input and collaboration in this work is inexpressible. The whole month that Guido kindly hosted me in his place in Brisbane resulted in most of the work in this thesis (I am also thankful to Magda and Darra for gently allowing me to stay for such long period).

Vienna was my home sweet home for 5 years. It felt like home, despite the long dark nights of winter, because of the many great friends I made there. Serwah, Navid, Aldo, Mihai, Florina, Linda, Alex, Paolo, Albin, Stephan, David, Bernardo, and Evelyn made sure that loneliness was something I rarely felt in Austria.

My parents Claudia and Ademar and my brother Pedro have always been role models for me and always encouraged me in my work and life decisions. I am really grateful for their support.

I am also eternally thankful to Giselle, who has been with me for more than a decade, sharing a bunch of discoveries, adventures, fears, and joyfulness.

Last, but never least, I would like to thank the reviewers of this thesis, Norbert Fuhr and Udo Kruschwitz, and the many anonymous reviewers of my work whose highly valuable feedback helped in shaping this thesis.

Abstract

Search engines are concerned with retrieving relevant information to support a user's information seeking task. In the health domain, access to understandable information is crucial as it has the potential to impact on people's health decisions. In this thesis, we study two aspects that should be taken into account by modern health search engines: the user health expertise in the health domain and the document understandability.

This thesis begins by considering the role of user expertise in the health domain. We investigate user search behavior through logfiles of several domain-specific health search engines. While most of the recent studies on health search behavior have been based on the search logs of commercial general purpose search engines, we performed here the important task of reproducing these studies on search logs of health search engines, finding out to what extent these results can be supported or not. Our query-log analysis can be used to understand health searchers better and even to predict the user expertise based on user behavior and their interactions with the search engine.

Our investigation of document understandability in the health domain arises from the increasing concern that health documents on the Web are not suitable for health consumers. For that, we study the impact that preprocessing pipelines have on readability formulas, which are commonly used to estimate the understandability of documents. We also examined domain-specific methods to estimate the understandability of documents and how machine learning approaches can be employed to predict document understandability.

In particular, for the health domain, documents should be considered more relevant if, apart from being topically relevant, they are also understandable by the searcher. For that, we need evaluation frameworks that consider other relevance dimensions beyond topicality. In this work, we propose a framework that delays the combination of scores for the different relevance dimensions, which facilitates the work of information retrieval practitioners by increasing the interpretability of the results. With such a framework, we evaluated various strategies to integrate understandability estimation into search engines, finding that learning-to-rank is the most effective approach.

This work contributes to improving search engines tailored to consumer health search because it thoroughly investigates promises and pitfalls of understandability estimations and their integration into retrieval methods. As shown by our experiments, these methods would undoubtedly improve current health-focused search engines.

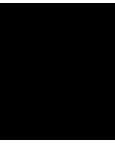
Contents

Abstract	ix
Contents	xi
I Preliminary	1
1 Introduction	3
1.1 Background and Motivation	3
1.2 Thesis Goal	5
1.3 Road Map and Contributions	6
1.4 Research Questions	8
1.5 Published Research	9
2 Related Work	13
2.1 Topics from Information Retrieval	13
2.2 User Search Behavior in the Health Domain	18
2.3 Understandability Estimation of Web Documents	21
2.4 Summary	24
II User Search Behavior in the Health Domain	25
3 Investigating Health Search through Query Logs	27
3.1 Query Logs and Preprocessing Steps	28
3.2 Enriching the Query Logs with MetaMap	32
3.3 Individual Query Analysis	37
3.4 Analysing Sessions	42
3.5 Summary	45
4 Exploiting Query Logs: Estimating User Medical Expertise	49
4.1 Data Collection	50
4.2 Classification Features	50
4.3 Classification Results	52
	xi

4.4	Summary	55
III Understandability Estimation of Web Documents		57
5	The effects of preprocessing HTML on the estimation of understandability	59
5.1	Traditional Readability Formulas	60
5.2	Preprocessing of Web Documents	61
5.3	The Influence of Preprocessing Methods on Retrieval Experiments	63
5.4	Summary	68
6	Analysing Documents: understanding understandability through correlation analysis	69
6.1	Experimental Methodology	70
6.2	Evaluation of Readability Formulas	77
6.3	Evaluation of Preprocessing Pipelines and Heuristics	81
6.4	The Best Understandability Estimators	85
6.5	Predicting Document Understandability	87
6.6	Summary	94
IV Understandability in Search Engines		95
7	Multidimensional Evaluation of Search Engines	97
7.1	Incorporating Understandability into Evaluation Metrics	98
7.2	A new Framework for Multi-Dimension IR Evaluation	100
7.3	Comparing frameworks Through System Simulations	101
7.4	Rank Correlations	103
7.5	Summary	104
8	Integrating Understandability into Search Engines	105
8.1	Methods to Integrate Understandability into Retrieval	105
8.2	Evaluation Measures	111
8.3	Evaluating Understandability Aware Retrieval	111
8.4	Summary	116
V Closure		117
9	Discussion and Conclusion	119
9.1	Discussion of Research Questions	119
9.2	Limitations and Future Work	124
A	CLEF eHealth Data	127

A.1 Introduction	127
A.2 Query Sets	128
A.3 Assessments	130
A.4 CLEF eHealth 2015	131
A.5 CLEF eHealth 2016	131
List of Figures	139
List of Tables	141
Bibliography	145

Part I
Preliminary



Introduction

1.1 Background and Motivation

Health is one of the most critical topics on the Web, both because of its popularity and its potential impact on people's life [90]. A 2013 Pew Research Center report found that 81% of U.S. adults use the Internet and, of those, 72% say they have looked online for health information in the previous year, and one in every three wanted to diagnose a medical condition [66].

Using the Web for health advice is a global tendency. In Europe, a survey from the European Commission estimates that 60% of the population have gone online to search for health-related information in 2014 [59]. In China, reports point to the fact that 28.9% of Chinese mobile Internet users have been using mobile healthcare apps, with 60.3% of them using apps for information search [126].

These reports show that search tasks vary from searching for very general information on health-related topics, such as diet or pregnancy, to searching for specific injuries and, in some cases, for rare diseases. It was also found that people often seek information on behalf of friends or relatives [48]; it is estimated that half of health information searches are on behalf of someone else [66]. Overall, these reports found that the search starts mostly in a commercial search engine, such as Baidu, Bing or Google.

Commercial search engines aim to retrieve relevant information to support a user's information seeking task. Commonly, signals about the topicality of a piece of information with respect to a query are used to estimate relevance, with other relevance dimensions, such as understandability, topical expertise, novelty, scope and trustworthiness [221] being relegated to a secondary position, or completely neglected. While this may be a minor problem for many information seeking tasks, there are some specific tasks in which dimensions other than topicality have an essential role. The seeking of health information on the Web by the general public is one such task.

In this thesis, we particularly study two of such dimensions: the user topical expertise and the document understandability in the health domain.

1.1.1 Topical Expertise

We define expertise as the amount of knowledge someone has on a topic. This knowledge can be acquired through study, training or experience in the subject. While expertise is intrinsically a continuous value, due to the nature of the data used in Part II of this thesis, we opt for defining two non-overlapping categories: the experts in the health domain and the non-experts in the health domain. We refer to them in this work, respectively, as *health experts* and *health consumers*.

Health consumers or just *consumers* are the general public, i.e., people without an in-depth and formal knowledge about health domain. *Patients* and *laypeople* are two terms often used in the literature to refer to this group. In this work, we avoid the term patient as the search engine user does not necessarily have a condition (as mentioned, searching on behalf of other people is a common practice). Note that some chronic condition patients might develop a broad knowledge on their diseases, often becoming experts in their one specific disease, but for the sake of simplicity we formally still consider these patients as health consumers.

On the other side of the spectrum, we opt for the term *health experts*, or simply, *experts*, to refer to people that have an in-depth and formal knowledge about the health domain, such as medical doctors, nurses, biologists or pharmacists. *Health professionals* or simply *professionals* are interchangeable terms often used in the literature to refer to health experts.

Studies have shown that the simple distinction between health consumers and health experts can significantly improve the quality of their interactions with search engines [207, 147, 207, 174, 39]. Schwarz et al. [174] show that the popularity of a webpage among experts is a crucial feature to help non-experts identify credible websites. Collins-Thompson et al. [39] discuss that re-ranking general search engine results to match the user's skills of readability can provide significant gains; however they also point out that estimating user profiles is a non-trivial task and needs to be further explored.

Part II of this thesis sheds light on this topic by investigating how both health consumers and health experts search online for health information. We profile these two types of users and study estimators that can be used to tell them apart.

1.1.2 Document Understandability

A key problem when searching the Web for health information is that retrieved documents might be too technical, unreliable and even misleading. While experts can identify the issues with low-quality results, the retrieval of unclear or incorrect health information poses potential risks to consumers, as they may dismiss severe symptoms, use inappropriate treatments or unfoundedly escalate their health concerns about common symptomatol-

ogy [17, 208]. In high-stakes search tasks such as this, access to poor information can lead to poor decisions which ultimately can have a significant impact on people's health and well-being [208, 205].

The use of general purpose commercial search engines for seeking health advice has been largely analyzed, questioned and criticised, despite the commendable efforts these services have put into providing increasingly better health information (e.g., the Google Health Cards [71]). While the access to a vast volume of online information could lead to better physician-patient relationships and better health outcomes [126, 215], an extensive number of studies has shown that the average user finds it difficult to understand the content of a significant portion of the results retrieved by current search engine technology, e.g., [81, 58, 62, 210, 161, 9, 133, 57]. In the context of consumer health information seeking, search engines should not only retrieve relevant information, but they should also promote information that is understandable, reliable and verified [208].

Parts III and IV of this thesis investigate the *understandability* of health information retrieved by search engines, and the improvement of search results to favor information understandable by the general public. Although very important in the health domain, we leave addressing the reliability and trustworthiness of the retrieved information for future work. Nevertheless, this can be achieved by extending the frameworks we investigate here.

1.1.3 Data Resources

In order to effectively support users in finding topical, high-quality, and accessible health information on the Web, new retrieval methodologies have to be developed and evaluated.

In this context, CLEF eHealth¹ is an evaluation lab organized within the Conference and Labs of the Evaluation Forum (CLEF) aiming to build resources and methods to support health consumers, their next-of-kin, clinical staff, and health scientists in understanding, accessing, and authoring eHealth information in a multilingual setting. The lab has been running yearly since 2013 [189, 107, 77, 106, 78] and historically has been built upon three main tasks in the health domain: information extraction, information management and information retrieval. In particular, the CLEF eHealth Evaluation Labs 2015 Task 2 [154] and 2016 Task 3 [226] were specifically designed to evaluate information retrieval systems aimed at health consumers to improve how the general public access medical information on the Web. The collections created in CLEF eHealth are extensively used in this thesis.

1.2 Thesis Goal

Information retrieval collections, such as CLEF eHealth 2015 and 2016, are developed to foster research in specific areas in which few or no solutions have been proposed. In

¹<https://sites.google.com/site/clefehealth/>

the case of consumer health search, only a few ad-hoc solutions exist. These solutions are typically supported by government initiatives or medical practitioner associations, e.g., HealthOnNet.org (HON) and HealthDirect.gov.au, among others. They aim to provide better health information to *health consumers*. For example, HON's mission statement is "to guide Internet users to reliable, understandable, accessible and trustworthy sources of medical and health information". But, do the solutions that these services currently employ actually provide this type of information to the health-seeking general public?

As an illustrative example, we analyzed the top 10 search results retrieved by HON on 01/10/2017 in answer to 300 health search queries generated by regular health consumers in health forums. These queries are part of the CLEF 2016 eHealth collection mentioned above. The understandability score of the retrieved documents was estimated with the most effective readability formula and preprocessing settings analyzed in the late chapters of this thesis. The understandability scores approximately correspond to the number of years in the school necessary to understand the text in the document. Thus low scores correspond to easy to understand Web documents. As the target audience of HON is health consumers without deep health expertise, one might expect that the content retrieved is accessible and easy-to-understand.

Figure 1.1 reports the cumulative distribution of understandability scores for these search results. Note that we did not assess the topical relevance of the documents here, only their estimated understandability score. We also report the understandability scores for the "optimal" search results (*oracle*), as found in CLEF 2016. In the oracle setting, we retrieved only results that were assessed as topically relevant ranked by their understandability scores. The other two systems reported are a typical baseline method (*BM25*) and our best retrieval method (*XGB*).

The results clearly indicate that, despite solutions like HON being explicitly aimed at supporting access to understandable health information, they often fail to do so (note how the cumulative distribution of HON and the BM25 baseline are similar).

The overall goal of this thesis is to investigate and establish methods and best practice for developing search engines that retrieve relevant and understandable health information from the Web for non-expert health consumers.

1.3 Road Map and Contributions

This thesis is divided into 5 parts. Part I contains this introduction and the related work of this thesis, shown in Chapter 2. Each of its subsections focuses on a particular forthcoming part of this thesis.

Next, we highlight here the three main general contributions of this thesis which we investigate, respectively, in Parts II, III and IV of this work.

The *first contribution* of this thesis is a study of user interactions with health/medical search engines. Part II analyzes health search behavior. Throughout a detailed query-log

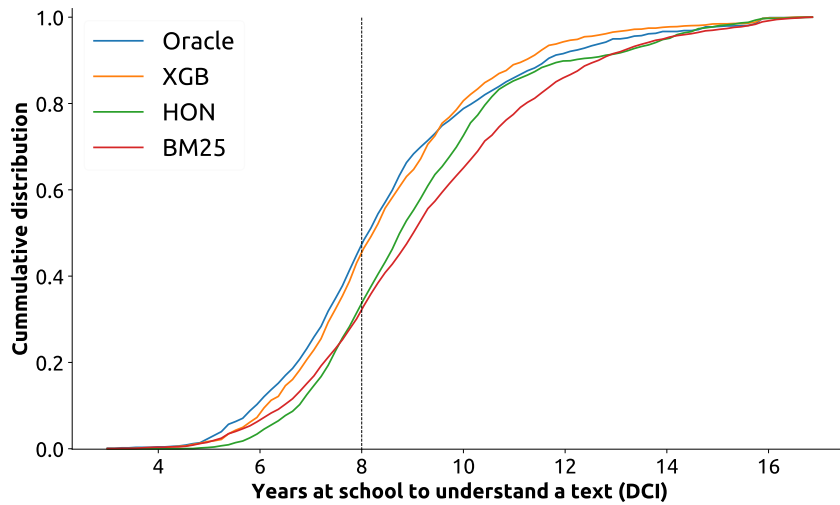


Figure 1.1: The cumulative distribution of the Dale-Chall Index (DCI) of search results. DCI measures the years of schooling required to understand a document. The average US resident reads at or below an 8th-grade level (dashed line) [43, 199, 50, 187], which is the level suggested by NIH for health information on the Web [196]. The distribution for HON is similar to that of the baseline used in this article (BM25). Our best method (XGB) re-ranks documents to provide more understandable results; its distribution is similar to that of an “Oracle” system.

based analysis, we investigate what distinguishes health consumers and health experts looking at what and how they search for medical advice on the Web. In Chapter 3, the query logs of four health search engines are investigated to understand the differences in search behavior of health experts and health consumers. These differences are then explored to build a health expertise classifier, which can be used to automatically classify search engine users as health experts or health consumers. Search engines could make use of such a classifier to foster search results’ content that, while topically relevant, has the highest level of understandability.

The *second contribution* of this thesis is an investigation into methods for estimating understandability of health documents. Part III changes the focus of this thesis from users to documents. Chapter 5 looks into how different HTML preprocessing settings impact the estimation of understandability when popular readability formulas are used. Readability formulas are widely used methods for estimating understandability of documents, but we identify hidden pitfalls in their usage. In Chapter 6, we develop other domain-specific methods which on top of being resilient to such pitfalls, are more effective. We make use of understandability assessments collected in CLEF eHealth to evaluate the different HTML preprocessing pipelines. In order to provide the most suitable document for

a health searcher, search engines need to be able to estimate how understandable a document is.

The *third contribution* of this thesis is the investigation and evaluation of methods to re-rank search engine results to boost understandable and relevant results. For that, Part IV focuses on building retrieval systems that, while aiming to find the most relevant documents for a given query, are specially trained to promote more understandable search results.

Topical relevance is central to the notion of relevance, but not the only factor (also called dimension) that determine the relevance of a document. Chapter 7 investigates evaluation frameworks that consider other relevance dimensions beyond the topical relevance. We present the current state-of-the-art framework for multidimensional search engine evaluation, UBIRE [224, 223], and its limitations. We propose a similar framework, named *H*, aiming to overcome the limitations found in UBIRE. With this new evaluation framework, in Chapter 8, we propose and evaluate different learning-to-rank approaches to retrieve topically relevant and understandable results.

Finally, Chapter 9 in Part V of this thesis presents the conclusion of this work, its limitations, and future directions.

1.4 Research Questions

This thesis is motivated by the need for a better health consumer experience when seeking out information on the Web, in particular retrieving information that can be understood by everybody. Based on that, the overall research question is: *How can we make search engines retrieve relevant and understandable information, especially for health consumers?*

We break this broad research question into many smaller ones which will be tackled in the following parts/chapters of this thesis:

- **Part II: Health Search Behavior**
 1. **Chapter 3:** *What and how do consumers and health professionals search in the health domain?*
 2. **Chapter 3:** *How suitable is an automatic health text annotator, such as UMLS MetaMap, to analyze and annotate short Web queries?*
 3. **Chapter 4:** *Can we automatically infer user health expertise through user search behavior?*
 4. **Chapter 4:** *What are the most useful features to infer user health expertise through search behavior automatically?*
- **Part III: Understandability Level of Web Documents**

5. **Chapter 5:** *What is the effect of preprocessing pipelines on readability formulas when estimating the understandability of Web documents?*
6. **Chapter 5:** *Among the readability formulas, what are the most and the least robust ones? Which readability formula should we use?*
7. **Chapter 6:** *What are the best understandability estimators among the various studied?*
8. **Chapter 6:** *How do preprocessing pipelines affect methods of understandability estimation?*

- **Part IV: Understandability Integrated into Search Engines**

9. **Chapter 7:** *How can we incorporate other relevance dimensions (e.g., understandability) into existing system evaluation metrics?*
10. **Chapter 7:** *What is the limitation of the state-of-the-art multidimensional evaluation framework and how can we overcome its limitation?*
11. **Chapter 8:** *How can understandability estimations be integrated into retrieval methods to enhance the quality of the retrieved health information?*

1.5 Published Research

The scientific work done during my Ph.D. is not limited to the work presented in this thesis. In this section, I compile a list of all my publications during the years of my Ph.D. studies and how they are linked (or not linked) to the text in this thesis:

- **Analysing Search Behavior - Part II - Chapters 3 and 4.**

1. Exploiting health related features to infer user expertise in the medical domain. **WSCD 2014.** *J Palotti, A Hanbury, H Müller* [147]
2. User intent behind medical queries: an evaluation of entity mapping approaches with MetaMap and Freebase. **III X 2014.** *J Palotti, V Stefanov, A Hanbury* [140]
3. How Users Search and What They Search for in the Medical Domain – Understanding Laypeople and Experts Through Query Logs. **Information Retrieval Journal 2016** *J Palotti, A Hanbury, H Müller, C Kahn* [148]

- **Analysing Document Content - Part III - Chapters 5 and 6.**

4. The Influence of Pre-processing on the Estimation of Readability of Web Documents. **CIKM 2015.** *J Palotti, G Zuccon, A Hanbury* [141]

5. Consumer Health Search on the Web: Study of Web Page Understandability and Its Integration in Ranking Algorithms, **JMIR 2019** *J Palotti, G Zuccon, A Hanbury* [157]
 - **Understandability in Search Engines - Part IV - Chapters 7 and 8.**
6. Ranking Health Web Pages with Relevance and Understandability. **SIGIR 2016** *J Palotti, G Zuccon, L Goeuriot, A Hanbury* [145]
7. Beyond Topical Relevance: Studying Understandability and Reliability in Consumer Health Search. **SIGIR 2016** *J Palotti* [142]
8. MM: A new Framework for Multidimensional Evaluation of Search Engines, **CIKM 2018** *J Palotti, G Zuccon, A Hanbury* [156]
 - **CLEF eHealth. The CLEF collections are extensively used in Chapters 5, 6, 7 and 8.**
9. ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred Health Information Retrieval. **CLEF 2014**. *L Goeuriot, L Kelly, W Li, J Palotti, P Pecina, G Zuccon, A Hanbury, G Jones, H Müller* [107]
10. Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. **CLEF 2014**. *L Kelly, L Goeuriot, H Suominen, T Schreck, G Leroy, D Mowery, S Velupillai, W Chapman, D Martínez, G Zuccon, J Palotti* [76]
11. Overview of the CLEF eHealth Evaluation Lab 2015. **CLEF 2015**. *L Goeuriot, L Kelly, H Suominen, L Hanlen, A Névéol, C Grouin, J Palotti, G Zuccon* [77]
12. CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving Information About Medical Symptoms. **CLEF2015**. *J Palotti, G Zuccon, L Goeuriot, L Kelly, A Hanbury, G Jones, M Lupu, P Pecina* [154]
13. The IR Task at CLEF eHealth evaluation labs 2016: user-centred health information retrieval. **CLEF 2016**. *G Zuccon, J Palotti, L Goeuriot, L Kelly, M Lupu, H Müller, J Budaher, A Deacon* [226]
14. Overview of the CLEF eHealth Evaluation Lab 2016. **CLEF 2016**. *L Kelly, L Goeuriot, H Suominen, A Névéol, J Palotti, G Zuccon* [106]
15. CLEF eHealth Evaluation Lab Overview. **CLEF 2017**. *L Goeuriot, L Kelly, H Suominen, A Névéol, A Robert, E Kanoulas, R Spijker, J Palotti, G Zuccon* [78]
16. CLEF 2017 Task Overview: The IR Task at the CLEF eHealth Evaluation Lab. **CLEF 2017**. *J Palotti, G Zuccon, Jimmy, P Pecina, M Lupu, L Goeuriot, L Kelly, A Hanbury* [155]

17. Building Evaluation Datasets for Consumer-Oriented Information Retrieval. **LREC 2016**. *L Goeuriot, L Kelly, G Zuccon, J Palotti* [79]
18. An analysis of evaluation campaigns in ad-hoc medical information retrieval: CLEF eHealth 2013 and 2014. **Information Retrieval Journal 2018** *L Goeuriot, G Jones, L Kelly, J Leveling, M Lupu, J Palotti, G Zuccon* [74]
19. Overview of the CLEF eHealth Evaluation Lab 2018. **CLEF 2018** *H Souminen, L Kelly, L Goeuriot, A Névéol, L Ramadier, A Roberts, E Kanoulas, R Sijker, L Azzopardi, D Li, Jimmy, J Palotti, G Zuccon* [188]
20. Overview of the CLEF 2018 Consumer Health Search Task. **CLEF 2018** *Jimmy, G Zuccon, J Palotti, L Goeuriot, L Kelly* [103]

- **Research papers related to health search**

21. *Khresmoi - Multilingual Semantic Search of Medical Text and Images*. **MEDINFO 2013**. *All Project Collaborators* [8]
22. TUW @ TREC Clinical Decision Support Track. **TREC 2014**. *J Palotti, N Rekabsaz, L Anderson, A Hanbury* [150]
23. TUW @ TREC Clinical Decision Support Track 2015. **TREC 2015**. *J Palotti, A Hanbury* [146]
24. Diagnose This If You Can – On the Effectiveness of Search Engines in Finding Medical Self-diagnosis Information. **ECIR 2015**. *G Zuccon, B Koopman, J Palotti* [225]
25. Exploring Understandability Features to Personalize Consumer Health Search. **CLEF 2017**. *J Palotti, N Rekabsaz* [149]
26. Interactive Exploration of Healthcare Requests. **CBMI 2016**. *A Bampoulidis, J Palotti, J Brassey, M Lupu, S Metallidis, A Hanbury* [12]
27. Does Online Evaluation Correspond to Offline Evaluation in Query Auto-Completion? **ECIR 2017**. *A Bampoulidis, J Palotti, J Brassey, M Lupu, A Hanbury* [13]
28. Query Variations and their Effect on Comparing Information Retrieval Systems. **CIKM 2016** *G Zuccon, J Palotti, A Hanbury* [227]
29. Assessors Agreement: A Case Study across Assessor Type, Payment Levels, Query variations and Relevance Dimensions. **CLEF 2016**. *J Palotti, G Zuccon, J Bernhardt, L Goeuriot, A Hanbury* [153]

- **Other research papers non-related to health search**

30. Insight to Hyponymy Lexical Relation Extraction in the Patent Genre Versus Other Text Genres. **IPaMin@KONVENS 2014**. *L Andersson, M Lupu, J Palotti, F Piroi, A Hanbury, A Rauber* [3]
31. TUW @ Retrieving Diverse Social Images Task 2014. **MediaEval 2014**. *J Palotti, N Rekabsaz, M Lupu, A Hanbury* [151]
32. TUW @ MediaEval 2015 Retrieving Diverse Social Images Task. **MediaEval 2015**. *S Sabetghadam, J Palotti, N Rekabsaz, M Lupu, A Hanbury* [169]
33. Learning to Rank for Personalized E-Commerce Search at CIKM Cup 2016. **CIKM Cup 2016 Workshop** *J Palotti* [143]
34. When is the time ripe for natural language processing for passage patent retrieval? **CIKM 2016**. *L Andersson, M Lupu, J Palotti, A Hanbury, A Rauber* [2]
35. Fixed-Cost Pooling Strategies based on IR Evaluation Measures. **ECIR 2017**. *A.Lipani, J Palotti, M Lupu, F Piroi, G Zuccon, A Hanbury* [123]
36. Fixed Budget Pooling Strategies based on Fusion Methods. **SAC 2017**. *A.Lipani, M Lupu, J Palotti, G Zuccon, A Hanbury* [122]
37. Leveraging Wikipedia’s Article Structure to Build Search Agents. **CLEF 2017**. *J Palotti* [144]
38. TrecTools: an Open-source Python Library for Information Retrieval Practitioners Involved in TREC-like Campaigns. **SIGIR 2019**. *J Palotti, H Schells, G Zuccon* [152]

Related Work

The goal of this thesis is studying the role of expertise and understandability in consumer health search. We start by expanding on this last word, *search*. For that, we review, in Section 2.1, the basic concepts of the Information Retrieval (IR) field and its components. We then focus on health search and the role of expertise on the health domain in Section 2.2. Finally, we review the literature on document understandability in Section 2.3. Note that this chapter condenses all the related work of this thesis.

2.1 Topics from Information Retrieval

As defined by Baeza and Ribeiro Neto, Information Retrieval deals with the representation, storage, organization of, and access to information items [11]. Simply put, the goal of this field is providing easy access to information in which a person is interested [11]. Croft, Metzler, and Strohman call the three core issues for the IR field: *information need*, *relevance* and *evaluation* [47].

Information need is the topic or piece of information that a person desires to learn more, and it is expressed by her with a *query* [128]. In other words, the information need is the underlying cause of the query that a person submits to an IR system/search engine/search system [47]. Unfortunately, it is not always easy to represent an information need with a query. There is an inherent gap between the information need and the query, and this mismatch often results in the retrieval of information that is not relevant for the searcher [206]. Information is relevant if it is one that the person perceives as containing value with respect to their personal information need [128]. In order to find the relevant information, researchers have proposed a number of retrieval models and systematically evaluate the performance of them.

In Section 2.1.1, we briefly overview the related work on IR evaluation, introducing the CLEF eHealth collections that were created and used in this thesis. Section 2.1.2

better defines the concept of relevance and, in particular, the concept of multidimensional relevance used in our work. One way to evaluate search systems is through the use of query logs generated by them. However, in this thesis, we employ user search logs to understand user search behavior better. We introduce the research on understanding user behavior with user query logs of general search engines in Section 2.1.3 and detail the related work for the health domain in Section 2.2.

2.1.1 Evaluating Search Systems

Evaluation has always been an important part of information retrieval research [114]. Starting in the late 1950s, the first large-scale evaluation of search performance, known as the Cranfield paradigm [33, 34], focused primarily on laboratory experiments designed for *batch evaluation*. In the 1990s, much more attention has been paid to the evaluation of real-life experiments with *user studies* and *online evaluation* [11]. Despite this tendency, laboratory experimentation is still dominant and the two main reasons for that are the repeatability and the scalability provided by the closed setting of a laboratory [11]. In this thesis, we do not perform any user study or online evaluation, instead, we focus on batch evaluation following the Cranfield paradigm.

In the Cranfield paradigm, a test collection is created with three components: (1) a document collection containing a very large set of documents; (2) a sample of typical information needs, represented by queries; and (3) a (very incomplete) set of assessments stating whether a document is relevant or not to fulfill an information need. Search systems can then be evaluated using pre-defined metrics, such as precision and recall, and research can be empirically conducted evaluating what works and what does not work. Every year, test collections with different purposes are created following this paradigm. The best-known test collections are those associated with evaluation forums such as the Text Retrieval Conference (TREC)¹, the Conference and Labs of the Evaluation Forum (CLEF)² and the NII Testbeds and Community for Information access Research (NTCIR)³. In particular, in this thesis, we make extensive use of CLEF eHealth 2015 and 2016 collections [154, 226].

The CLEF eHealth 2015 task provides a test collection to evaluate the effectiveness of search engines in answering self-diagnosing queries [154]. The collection⁴ is composed of a crawl of about one million documents, which have been made available to CLEF eHealth through the Khresmoi project⁵ [85, 8, 105]. The query set explored circumlocutory queries that users may pose when faced with signs and symptoms of a medical condition [186, 225]. The evaluation framework explicitly accounted for both the topical relevance of the search results and their understandability, interpreted as how easy it is for a health consumer to understand the content of a specific search result. This was done using understandability-

¹<http://trec.nist.gov>

²<http://www.clef-initiative.eu/>

³<http://http://research.nii.ac.jp/ntcir/index-en.html>

⁴Available at http://catalog.elra.info/product_info.php?products_id=1218

⁵<http://khresmoi.eu/>

biased evaluation measures, where the gain obtained from relevant information was weighted by how hard it is for a consumer to understand that information [223, 224]. Zuccon and Koopman [224], and later Zuccon [223], have proposed and investigated a family of measures based on the gain-discount framework, where the gain of a document is influenced by both its topical relevance and its understandability. They showed that, although generally correlated, topical-relevance evaluation alone provides differing system rankings compared to understandability-biased evaluation measures.

In turn, the CLEF eHealth 2016 Information Retrieval task [226] introduced a much larger Web corpus (ClueWeb12 B13⁶) which aimed to make the challenge more realistic, simulating well how users would search for health information online. Topics were extracted by mining the AskDocs health Web forums from Reddit⁷ to identify example information needs [226]. The evaluation framework of the CLEF eHealth 2015 collection was kept.

In this thesis, we use the CLEF eHealth 2015 and 2016 test collections along with the explicit understandability assessments distributed and the understandability-biased RBP measure (see [154, 223, 224]). In addition, we review the understandability-biased evaluation framework in detail and propose improvements in Chapter 7.

Initial attempts to use understandability estimation for improving search results in consumer health search were proposed [177, 201, 197] by participating teams in the CLEF eHealth information retrieval task or shortly after the collections were created. For example, Silva and Lopes experimented with ways to directly combine relevance score with understandability scores from Flesch-Kincaid, SMOG and Gunning Fog index [177], and Wang, Lu and Ren proposed a method based on query expansion using the consumer health search vocabulary [201]. The results of the best participating teams in both CLEF eHealth 2015 and 2016 are directly compared to our proposed methods in Chapter 8. Van Doorn et al. [197] have shown that learning a set of rankers that provide trade-offs across many relevance criteria, including readability/understandability, increases overall system effectiveness.

A general data analysis of these two CLEF eHealth collections is presented in Appendix A as they are extensively used in Parts III and IV of this thesis.

2.1.2 Multidimensional Document Relevance

A crucial part of system evaluation in the Cranfield paradigm is defining relevance. One of the first notions of relevance was straightforward: a document is relevant to a query if the topic of the information retrieved matches the topic of the request. This was called *topical relevance* by Eisenberg and Schamber [56] and measured whether a document was “on the topic”. Due to its simplicity and clear definition [20], this notion was adopted by most of the modern evaluation tasks, such as TREC, CLEF and NTCIR, which heavily rely

⁶<http://lemurproject.org/clueweb12/>

⁷<http://reddit.com/r/AskDocs/>

on evaluation metrics that exclusively measure the topical relevance of documents [198]. Recently, extensions were made based on notions of novelty and diversity (e.g., [32]), but these new metrics still mainly consider relevance with respect to document topicality.

However, document relevance cannot be attributed to just one factor such as topicality. Instead, it is multidimensional and situational [20]. Early research [172, 190, 86], reviewed by Borlund, divided relevance into two classes: (1) objective or system-based relevance and (2) subjective or human based relevance [20]. This first class of relevance exclusively refers to the document “aboutness” aforementioned and it is context-free, while the second one refers to the subjective factors in both users and documents and it is context dependent. A large corpus of research was dedicated to identifying the subjective aspects of relevance. Park, for example, identified individual’s subject knowledge, professional training, and educational background as a user-based influential factor, while scarcity, availability, timeliness, and scope were identified as document-based factors [159]. Schamber published a compiling and non-exhaustive list of 80 relevance criteria suggested in the literature [173]. Recently, Fuhr et al. [70] proposed to enrich documents with an automatically generated information label, similarly to the nutrient information label found on packaged foods. Such label would have scores for many dimensions discussed here such as factuality, readability/understandability, virality, emotion, opinion, controversy, authority/credibility/trustworthiness, technicality, and topicality. In the same way that a person decides if she wants to buy a dish for lunch after inspecting its nutrient information, she should be able to determine whether a document is relevant or not after examining its information label.

In the consumer health search domain, Hersh highlights two important factors that should be considered by modern search engines [90]: understandability and information trustworthiness. This thesis focus on the first one.

Initial work to take document understandability into account for better personalization has been evaluated and successfully implemented in general search engines [191, 39, 109, 40, 176, 125].

Tan, Gabrilovich and Pang, for example, used a month of Yahoo search engine logs to evaluate a system that personalizes content according to user familiarity in a domain [191]. They estimate text understandability using a classifier trained on 40,000 aligned pairs of articles from both Simple Wikipedia (easy-to-read documents) and standard English Wikipedia (hard-to-read documents). As the feature set, they used the output of different readability formulas as well as bag-of-words features (850 basic English words from a Wikipedia list⁸). Interestingly, “health and wellness” was the hardest domain among the 17 domains evaluated. Their evaluation framework was based on user clicks and their experiments reported that content ranking was significantly improved.

Similarly to Tan et al., Collins-Thompson et al. [39] and their follow up work, Kim et al. [109], modeled the understandability of Web documents retrieved by the Bing

⁸https://simple.wikipedia.org/wiki/Wikipedia:Basic_English_ordered_wordlist

Search Engine and the reading level of Bing users. They explicitly modeled 12 reading levels corresponding to the 12 US school grades, as was previously done by Collins-Thompson [40]. Their main finding is that reading level and topic metadata used together were more effective than either one used alone.

There are two fundamental differences between the aforementioned work and this thesis: (1) the focus on our work is the health domain rather than the general domain: this allowed us to explore much more domain related features; (2) instead of using commercial search engines to model the effects of integrating document understandability into search engines, we opt to foster research in developing freely available collections, which can be used by any other researcher in industry or academia. These efforts in creating such freely available collections led us to establish the Information Retrieval task of CLEF eHealth Evaluation Lab [75, 76, 154, 226, 155], from which the author of this thesis actively collaborated since 2014. In particular, the integration of document understandability was explicitly explored in the information retrieval task of CLEF eHealth 2015 [154] and 2016 [226], which are used in Chapters 6, 7 and 8.

2.1.3 User Query Logs

As soon as modern search engines appeared, the first studies on query logs started [99, 178, 98, 87]. For example, Jansen et al. [99] and Silverstein et al. [178] analysed the logs from Excite and Altavista, respectively, popular search engines at that time. Both studies point out some essential results such as the fact that the vast majority of users issue only one single query and rarely access any result page beyond the first one. These results helped to model the research in the information retrieval field by, for example, given emphasis to precision over recall, making query recommendation [10, 94], and understanding user sessions [104, 72].

The most recent general search engine to disclose query logs to researchers was America Online (AOL) in 2006 [160]. The AOL data were afterward used in various studies, such as Brenes et al. [23], which provides methods to group users and their intents, and Torres et al. [54], who analyze queries targeting children's content. In this work, we compare the analysis made in the literature for general search engines [99, 179] with medical domain search engines, and we adopt a method similar to [54] to divide the AOL logs into queries related or not to health.

It is important to mention that the AOL log had known privacy problems in the past [1, 72], resulting in some users being identified even though the logs were supposedly anonymized. Despite this problem, we opt to use this dataset in Chapters 3 and 4 for several reasons. One reason is that it can be freely downloaded, as well as the code used for all the experiments of this thesis, making the experiments reproducible. In the absence of a more recent large search engine query log, we consider that the AOL logs are still the best choice for researchers in academia. A complete reference of the previous 20 years of research on log analysis and its applications is well described by Silvestri [179].

There are a number of studies analyzing query logs in the medical domain. We highlight in details in the next section some important work for this research, including work based on general search engines [185, 207, 27, 209], as well as specialized ones [91, 96, 194, 131, 220].

2.2 User Search Behavior in the Health Domain

Part II of this thesis investigates user search behavior in the health domain based on the query logs of various search engines. Each search engine was designed explicitly for either being used by health consumers (e.g., laypeople, general public) or by health experts (e.g., nurses and physicians). We start by reviewing the literature on health search behavior when the search is done by health consumers with general search engines (Section 2.2.1) and by health experts with specialized search engines (Section 2.2.2). We also review the related work on the role of user expertise in search behavior (Section 2.2.3).

2.2.1 Health Search using General Search Engines

In one of the first studies on search query logs, Spink et al. [185] investigated medical queries issued in 2001 in both Excite and AlltheWeb.com, popular search engines at that time. They found that medical web search was decreasing since 1999, suggesting that users were potentially shifting from general-purpose search engines to specialized sites for health-related queries. They also showed that health-related queries were equivalent in length, complexity, and lack of reformulation to general web searching.

Toms and Latter [193] observed 48 consumers searching for four health-related topics using Google. They used transaction logs, video screen capture, retrospective verbal protocols and self-reported questionnaires to study user behavior. Their results indicated significant problems in query formulation (on average 4.2 keywords were used per query, but out of those, 3.2 were stopwords and thus not processed by the search engine) and in making efficient selections from result lists. While the searching behavior in a lab environment might substantially differ from natural searching behavior, the queries issued by the participants were successfully used in the TREC 10 Interactive Track⁹.

The European Center for Disease Prevention and Control surveyed the research on consumer health search published between 2006 and 2010 reporting that, among other findings, females are more likely than males to search for health information and online health consumers tend to be more educated, earn more and have high-speed internet access at home and work [139]. Interestingly, they also report that those with limited literacy skills have less knowledge of disease management and health-promoting behaviors, poorer health status and are less likely to use preventive services than those with average or above average literacy skills.

More recently, White and Horvitz studied how users start looking for a single symptom and end up searching for severe diseases, a phenomenon they named cyberchondria [208, 209].

⁹<http://trec.nist.gov/data/t10i/t10i.html>

They used the logs of the Windows Live Toolbar to obtain their data and list of keywords to annotate symptoms and diseases in queries. Instead of word matching, we used the US National Library of Medicine MetaMap to do the same in Chapter 4. Similarly to our work, they define user sessions as a series of queries followed by a period of user inactivity of more than 30 minutes and they made use of the Open Directory Project (ODP) hierarchy to identify medical sessions.

Another important recent work is Cartright et al. [27], in which a study on user behavior when searching for health information online is presented. They classified user queries into three classes (symptoms, causes, and remedy), and analyzed the change of search focus along a session. They showed that it is possible to build a classifier to predict what is the next focus of a user in a session. We decided to use the same classes in order to make our study comparable, however, we used the semantic annotator of MetaMap instead of hand-coded rules.

Not based on query logs, but on the ranking lists of major general search engines, Wang et al. compared the results of Google, Yahoo!, Bing, and Ask.com for one single query *breast cancer* [200]. Among their conclusions is the fact that results provided rich information and highly overlapped between the search engines. The overlap between any two search engines was about half or more. Thus, if the quality of results retrieved by a specific search engine for a health query is poor, it is likely that it will be poor for other search engines as well.

2.2.2 Health Search using Specialized Search Engines

Herskovic et al. analyzed an arbitrary day in PubMed¹⁰, the largest biomedical database in the world [91]. They found that the usage of PubMed differs from the usage of general Web search engines. For example, PubMed queries are longer than the ones issued on Excite and Altavista. Subsequently, Dogan et al. studied an entire month of PubMed log data [96]. Their main finding comparing PubMed and general search engines was that PubMed users are less likely to select results when the result sets increase in size, users are more likely to reformulate queries and are more persistent in seeking information. In Chapter 3, whenever it is possible, we compare our findings with the ones made for PubMed.

Meats et al. analyzed the 2004 and 2005 logs of the TRIP Database, together with a usability study with nine users [131]. Their work concluded that most users used a single term and only 12% of the search sessions made use of Boolean operators, under-utilizing the search engine features. Tsikrika et al. examined query logs from ARRS GoldMiner, an expert search engine for radiology images [194]. They studied the process of query modification during a user session, aiming to guide the creation of realistic search tasks for the ImageCLEFmed benchmark. Meats used 620,000 queries and Tsikrika only 25,000, while we use respectively nearly 3 and 9 times more queries from TRIP and GoldMiner in Chapters 3 and 4, allowing us to perform a more in-depth analysis.

¹⁰<https://www.ncbi.nlm.nih.gov/pubmed>

Zhang analyzed how 19 students solved 12 tasks using MedlinePlus¹¹ [220]. The tasks were created based on questions from the health section of Yahoo! Answers. Even though the log analysis made is very limited due to the artificial scenario created and the small number of users, Zhang could investigate browsing strategies used by users (amount of time searching and/or browsing MedlinePlus) and the users' experience with Medline-Plus (usability, usefulness of the content, interface design) through questionnaires and recording the users performing the tasks. Although our user analysis is limited to the query logs, a larger analysis is made with different search engines and the user behavior is captured in a very natural setting.

2.2.3 Influence of Expertise in Search

One of the first studies to report how expertise influences the process of search dates from the 1990s. In that work, Hsieh-Yee reported that experienced library science students could use more thesauri, synonymous terms, combinations of search terms and spend less time monitoring their searches than novices [93]. Later, Bhavnani studied search expertise in the medical and shopping domains [18]. He reported that experts in a topic could easily solve the task given even without using a search engine because they already knew which website was better adapted to fill their needs. Bhavnani also reported that experts started their search by using websites such as MedlinePlus, instead of a major search engine, while consumers started with Google.

White et al. [207] showed a log-based analysis of expertise in four different domains (medicine, finance, law, and computer science), developing an expertise classifier based on their analysis. Apart from showing that it is possible to predict user expertise based on their behavior, they showed that experts have a higher success rate only in their domain of expertise, with success in a session being defined as a clicked URL as the final event in a session. Therefore, an expert in finance would have a comparable or worse success rate in medicine than a non-expert. A significant limitation of their work is the approach used to separate experts from non-experts. They assume that search leading to PubMed was done by medical experts and search leading to ACM Digital library (ACM-DL)¹² was made by computer science experts. In the medical domain this is a weak premise for two reasons: (1) it is estimated that one-third of PubMed users are health consumers [117], (2) PubMed is more important for medical researchers than practitioners [113]. Tracing a parallel between medicine and computer science, a general practitioner would be like a software developer that does not necessarily need to consult the ACM-DL (the correspondent for PubMed) to perform his/her work. One could manually expand the list of expert sites to include, for example, StackOverflow¹³ or an API website for experts in computer science and treatment guidelines or drug information sites for medicine but it would be a laborious task and unstable over time.

¹¹MedlinePlus is a web-based consumer health information system developed by the American National Library of Medicine (NLM): <http://www.medlineplus.gov/>

¹²<http://dl.acm.org/>

¹³<http://stackoverflow.com/>

Hence, to cope with this challenge, we use the logs of different search engines made for distinct audiences.

Other few recent user studies were also conducted to infer user expertise in the medical domain. For example, Zhang et al. [219] and Cole et al. [36] are based on TREC Genomics data. The former employed a regression model to match user self-rated expertise and high-level user behavior features such as the mean time analyzing a document and the number of documents viewed. They found that the user’s domain knowledge could be indicated by the number of documents saved, the user’s average query length, and the average rank position of opened documents. Their model, however, needs to be further investigated because the data was limited, collected in a controlled experiment. Similarly, but using only eye movement patterns as features, the latter conducted a user study instead of log analysis and employed a linear model and random forests to infer the user expertise level. Their main contribution is demonstrating that models to infer a user’s level of domain knowledge without processing the content of queries or documents is possible, however they only performed one single experiment and in one single domain.

2.3 Understandability Estimation of Web Documents

Understandability refers to the ease of comprehension of the information presented to a user. Health information is understandable “when health consumers of diverse backgrounds and varying levels of health literacy can process and explain key messages” [175]. Often the terms understandability and readability are used interchangeably: we use readability to refer to formulas that estimate how easy it is to understand a text, usually based on its words and sentences. We use understandability to refer to the broader concept of ease of understanding: this is affected by text readability, but may also be influenced by how legible a text is and its layout, including, e.g., the use of images or diagrams to explain difficult concepts. In general, increasing readability tends to improve understanding/understandability [121].

There is a large body of literature that has examined the understandability of Web health content when the information seeker is a member of the general public. For example, Becker reported that the majority of health Web sites are not well designed for the elderly [16], while Stossel et al. found that health education material on the Web is not written at an accessible reading level [187]. A common finding of these studies is that, in general, health content available on Web pages is often hard to understand by the general public. Often, understandability research in the medical domain includes content that is retrieved in top-ranked positions by current commercial search engines [81, 58, 62, 210, 161, 9, 133].

Previous Linguistics and Information Retrieval (IR) research have attempted to devise computational methods for the automatic estimation of text readability and understandability, and for the inclusion of these within search methods or their evaluation. We divide the computational approaches to understandability estimation into the two following categories: (1) *Readability Formulas*: Section 2.3.1 presents the most common readability

formulas, which generally exploit word surface characteristics of the text; (2) *Machine Learning*: Section 2.3.2 presents machine learning approaches for understandability estimation, these include the use of medical specialized dictionaries or terminologies, often compiled with information about understandability difficulty.

2.3.1 Readability Formulas

While recent research has proposed sophisticated readability estimation methods [39, 82], often tailored to specific domains [214], traditional readability measures such as the Automated Readability Index [180] and the Gunning Fog Index [83] are extensively used for assessing information on the Web (see [210, 224], for example). These long-established readability measures consider the surface level of the text contained in Web pages, that is, the wording, syllables counts and sentence syntax. In this framework, the presence of long sentences, words containing many syllables and unpopular words, are all indicators of difficult text to read.

In this thesis, we consider a large number of readability formulas, which are listed in Table 2.1. For example, the Dale-Chall readability formula is based on a corpus of 3,000 words that can be understood by fourth-grade students [49], the Flesch-Kincaid and Flesch Reading Ease measures compute a readability score based on a weighted combination of the number of words and the number of syllables in a sentence [110]. The Gunning-Fog index combines the intuitions of these preceding approaches using sentence length and frequency of “complex” words [83].

Generally, traditional readability measures estimate the minimum required level of knowledge to comprehend a text, often measured using the U.S. grade level system. For example, a text with a score of 1 would be suitable for a 6-7-year-old child, while a score of 13 requires the knowledge of a freshman undergrad student. Among the metrics used here, Flesch Reading Ease (FRE) and Lasbarhetsindex (LIX) are the only ones which do not follow the U.S. grade level system. With the exception of Flesch Reading Ease, for all other measures, the higher the readability score, the harder it is to understand the text.

Note that several (if not all) experiments published in the medical area often consider one or more of these readability formulas as a proxy for understandability (e.g. [81, 62, 210, 161, 9, 133]). We analyze the pitfalls of using such readability formulas to infer understandability of Web documents in Chapter 5.

2.3.2 Machine Learning for Understandability Predictions

Earlier research explored the use of statistical natural language and language modeling [176, 125, 40, 89, 39] as well as linguistic factors, such as syntactic features or lexical cohesion [162]. Si and Callan [176], for example, devised a new formula which linearly combined two components: (1) a value for word frequency in a background corpora, using a Unigram Language Model; and (2) a value to cope with the sentence length distribution of a document, modeled with a normal distribution. They showed that the combination of these two components was more accurate than using either of them alone. The use of

Table 2.1: The most frequently used readability formulas. C is the number of characters, DCW is the number of words found in the Dale-Chall list of 3,000 common words, LW is the number of long words (words with 6 or more characters), PW is the number of polysyllables words (words with 3 or more syllables) S is the number of sentences, Sy is the number of syllables and W is the number of words in the text.

Automated Readability Index (ARI) [180]

$$ARI = 4.71 \times \frac{C}{W} + 0.5 \times \frac{W}{S} - 21.43$$

Coleman-Liau Index (CLI) [37]

$$CLI = 5.89 \times \frac{C}{W} - 30.0 \times \frac{S}{W} - 15.8$$

Dale-Chall Index (DCI) [49]

$$DCI = 15.79 \times \frac{DCW}{W} + 0.0496 \times \frac{W}{S}$$

Flesch-Kincaid Grade Level (FKGL) [110]

$$FKGL = 0.39 \times \frac{W}{S} + 11.8 \times \frac{Sy}{W} - 15.59$$

Flesch Reading Easy (FRE) [110]

$$FRE = 206.835 - 1.015 \times \frac{W}{S} - 85.6 \times \frac{Sy}{W}$$

Lasbarhetsindex (LIX) [19]

$$LIX = \frac{W}{S} + \frac{LW \times 100}{W}$$

Gunning Fog Index (GFI) [83]

$$GFI = 0.4 * \left(\frac{W}{S} + 100.0 \times \frac{PW}{W} \right)$$

Simple Measure of Gobbledygook (SMOG) [129]

$$SMOG = 1.0430 * \sqrt{PW \times \frac{30.0}{S}} + 3.1291$$

the Unigram Language Model, though, requires a (large) set of training documents, i.e., documents which were previously classified either manually or through a set of rules.

Liu et al. [125] and Collins-Thompson & Callan [40] followed the steps of Si and Callan [176], successfully building models based on manually labeled Web documents using the American K-12 school system. Because of the large vocabulary difference, models trained on general school text for estimating the understandability of general English documents are not the best alternative for estimating the understandability of specialized medical content. Nevertheless, the idea of using statistical models is further explored in the health domain in Chapter 6. In the medical domain, Zeng et al. explored features such as word frequency in different medical corpora to estimate concept familiarity, which prompted the construction of the Consumer Health Vocabulary (CHV) [217, 218, 216].

The CHV is a prominent medical vocabulary dedicated to mapping consumer vocabulary to technical terms [218]. It attributes a score for each of its concepts with respect to their difficulty, with lower/higher scores for harder/easier concepts. Researchers have evaluated CHV in tasks such as document analysis [120] and medical expertise prediction [147]. The hierarchy of MeSH was previously used in the literature to identify hard concepts, assuming that a concept deep in the hierarchy is harder than a shallow one [213]. Other approaches combined vocabularies with word surface characteristics and syntactic features, like part of speech, into a unique readability measure [108]. Also in Chapter 6, we investigate approaches to estimate understandability that are based on both CHV and MeSH vocabularies.

2.4 Summary

We centralized in this chapter the related work of each part of this thesis.

In Section 2.1, we overviewed the core concepts of the information retrieval field used in this thesis. In particular, we introduced the query logs used to understand user search behavior in Part II of this thesis, and the 2015 and 2016 CLEF eHealth collections which are extensively used in Parts III and IV of our work.

Section 2.2 focused on the health domain and reported the related work on health search and user expertise, which is the main subject of Part II of this thesis.

Finally, Section 2.3 described the related work of Part IV by presenting methods to assess the understandability of health documents on the Web.

Part II

User Search Behavior in the Health Domain

Investigating Health Search through Query Logs

User interactions with a search engine are usually recorded to improve the quality (both in terms of effectiveness and efficiency) of information retrieval systems. These interactions, *the search query logs*, typically contains information about users, issued queries and clicked results, along with others. The knowledge acquired from the interactions between users and search engines is the essential part of personalization strategies used by most of the commercial search engines, as it is an unobtrusive method which captures the user behavior in a natural setting [101, 179].

In this part of this thesis, we make extensive use of search engine logs. We divide the users of medical search engines into health consumers and health experts, where health consumers, i.e., the general public, are considered to be searchers that do not have an in-depth knowledge about the health/medical topic being searched, while health experts do have an in-depth knowledge about the medical topic being searched.

Previous research has pointed to the fact that distinguishing consumers and experts can significantly improve their interactions with the search engine [207, 147]. We assume that it is possible to distinguish the level of expertise of the searcher based on the vocabulary used and the search style. Note that while it would be realistic to represent a continuum of expertise levels, we define only two classes (consumers and experts) in this study, allowing us to investigate the most relevant differences between the classes.

In our analysis, we use health-related queries from the America Online (AOL) query log [160], as well as the Health on the Net (HON) search engine log to represent the logs generated to a significant extent by health consumers. Health experts also use general search engines to seek health content; however, their queries are drowned in the consumer queries. White et al. [207], for example, hypothesize that search leading to PubMed was

done by experts. However using this hypothesis, only 0.004% of the whole AOL log was issued by health experts.

Besides the fact that PubMed is more frequently used in a research environment rather than in a clinical environment [113], it is also frequently visited by health consumers [117]. Therefore, instead of the PubMed queries, we use the logs from the evidence-based search engine TRIP Database and the radiology image search engine American Roentgen Ray Society (ARRS) GoldMiner to represent queries entered by physicians usually when facing a practical problem.

One limitation of using such specialized search engines, though, is that they only index medical content, rather than general health information. Thus, we shall restrict our analysis only to *queries in the medical domain*, instead of the full health domain. Important topics, such as diet, physical activity, and well-being should be excluded from our further analysis. However, it is important to note that the search for medical information, e.g., which includes self-diagnosis, advice on drug use or treatment options, is likely the most harmful if misunderstood by health consumers.

After introducing the datasets and preprocessing steps in Section 3.1, we present and evaluate MetaMap, the tool used to enrich the information contained in the query logs, in Section 3.2. In Section 3.3, we use the mappings to analyze individual queries, following a very similar approach carried out by Herskovic et al. [91], being able to compare our results for individual queries. Later, in Section 3.4, the focus is on the session level. Summary and discussion are presented in Section 3.5.

3.1 Query Logs and Preprocessing Steps

This section describes the datasets used to analyze user behavior on the Web (Section 3.1.1), their relationship (Section 3.1.2) and the required preprocessing that were applied (Section 3.1.3).

3.1.1 Query Logs

Four query logs from search engines taking free text queries were divided into five datasets in our analysis: two focused on health consumer queries, two made up of queries from health experts and one consisted of queries not related to health or medical information.

The query logs that are assumed to consist almost completely of queries submitted by health consumers were obtained from medical-related searches in America Online's search service [160]¹ and from the Health on the Net Foundation website (HON²).

The AOL logs were obtained from March to May 2006. We divided them into two non-overlapping sets: **AOL-Medical** and **AOL-NotMedical**. For this purpose, the click-through information available in the AOL data was used. A common approach

¹Obtained from <http://www.gregsadetsky.com/aol-data/>

²<http://www.hon.ch/HONsearch/Patients/index.html>

Table 3.1: ODP categories used to filter the AOL-Medical. These categories are the most relevant ones related to Medicine in ODP hierarchy (see <http://www.dmoz.org/Health/Medicine/>)

ODP Category	URL Examples
\Top\Health\Medicine	http://www.nlm.nih.gov http://www.webmd.gov
\Top\Health\Alternative	http://www.acupuncturetoday.com http://www.homeopathyhome.com
\Top\Health\Dentistry	http://www.dental--health.com http://www.animated-teeth.com
\Top\Health\Conditions_and_Diseases	http://www.cancer.gov http://www.cancer.org
\Top\Health\Organisations\Medicine	http://www.ama-assn.org http://www.aafp.org
\Top\Health\Resources	http://health.nih.gov http://www.eyeglassretailerreviews.com

to infer the topic of a URL is checking if it is listed in the Open Directory Project (ODP)³ [27, 39, 54, 207, 209]. For the clicked URLs that are not present in ODP, some researchers use supervised learning to automatically classify them [39, 207, 209]. However, it is very important to note that this approach cannot be used here, as 47% of the AOL log entries lack the clicked URL information.

Alternative approaches can be designed. One is to keep only queries in which the clicked URL is found in ODP, excluding all the rest. Although valid, this approach results in removing 73% of all queries, as only 27% of the queries had a clicked URL found in ODP. This has a substantial impact on the behavior analysis, such as a vast reduction in the number of queries per session. Another possibility is doing as in Cartright’s work [27], in which a list of symptoms was used to filter sessions on health information. However, this approach creates a strong bias when analyzing what users are searching for, as it certainly results in a dataset in which everyone searches for symptoms.

Our solution is based on user sessions – this approach is not as restricted as to analyzing single queries and does not suffer from the bias of filtering by keywords. First, we divide the query log into user sessions, continuous queries from the same user followed by an inactivity period exceeding 30 minutes. After this, we attribute one of the following labels for each clicked URL, if any: (1) *Medical*, (2) *Not Medical*, or (3) *Not Found*. This depends on whether the URL is (1) found in any Medical category listed in Table 3.1; (2) found in any other category: News, Arts, Games, Health/Animals, Health/Beauty, etc.; or (3) not found in either of these. Last, we assign to the whole session the Medical label only if the proportion of URLs on Medical information is greater than a threshold t .

³<http://www.dmoz.org/>

3. INVESTIGATING HEALTH SEARCH THROUGH QUERY LOGS

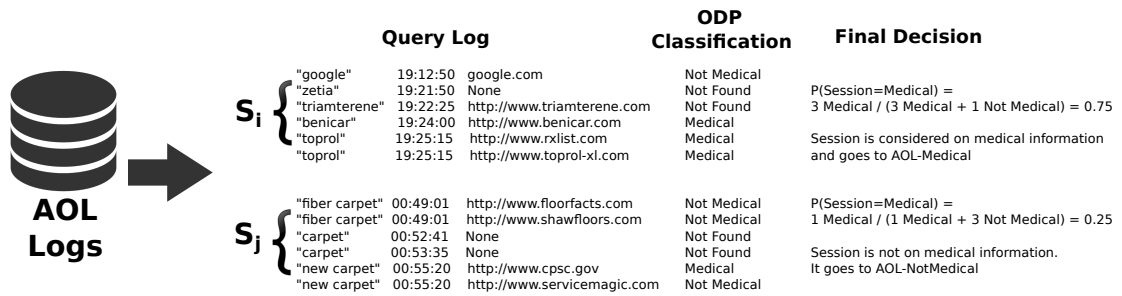


Figure 3.1: Two real user sessions extracted from AOL logs, S_i is classified as a search for medical content, while S_j is not.

Medical search sessions classified this way are attributed to the set **AOL-Medical**, while the rest goes to the **AOL-NotMedical** set. Figure 3.1 illustrates the session assignment procedure. For the experiments performed in this work, we use $t = 0.5$. This value is a fair trade-off between two extremes: considering an entire session as being on medical information because one single URL on medical information was clicked (see the second part of Figure 3.1), and considering an entire session as being on medical information only if all the known clicked links are on medical information (see the first part of Figure 3.1).

For the first part of Figure 3.1, it is important to note that the first query could belong to another session, as the user intent might be different from the rest of the session. The second and third queries, drug names that are clearly for medical content, were not used to calculate whether the session was on medical information or not, as their clicked URLs were not found in ODP. After the label estimation is done, all the queries of a session are assigned to the same class, therefore all six queries in S_i are assigned to AOL-Medical.

While only 27% of the queries have their URLs found in ODP, using the session approach described above allows us to have 50% of all sessions with at least one URL found in ODP. Altogether, 68% of all AOL queries were evaluated, as they belong to sessions that had at least one clicked URL in ODP. A more accurate way to define sessions is a field of research by itself [87, 104, 72] and it is not the goal of this work.

The **HON** dataset is composed of anonymous logs ranging from December 2011 to August 2013. This non-governmental organization is responsible for the HONcode, a certification of quality given to websites fulfilling a pre-defined list of criteria [21]. HON provides a search engine to facilitate the access to the certified sites. Although the majority of the queries are issued in English, the use of French or Spanish is frequent. Aiming to reduce noise, every query in the HON dataset was re-issued in a commercial search engine and the snippets of the top 10 results were used as input for an automatic language detection tool [127], which presented a precision of 94% in filtering English queries.

As health expert datasets, we use the query logs from the Turning Research Into Practice

(**TRIP**) database⁴ and ARRS **GoldMiner**⁵. The former is a search engine indexing more than 80,000 documents and covering 150 manually selected health resources such as MEDLINE and the Cochrane Library. It intends to allow easy access to online evidence-based material for physicians [131]. The query logs contain queries of 279,280 anonymous users from January 2011 to August 2012. GoldMiner consists of logs from an *image* search engine that provides access to more than 300,000 radiology images based on text queries of text associated with the images. Although the usage of an image search engine is slightly different from document search, previous work in the literature [194, 92] showed that the user search behavior is similar. We had access to more than 200,000 queries with last logged query being issued in January 2012. Due to a confidentiality agreement, we cannot reveal the start date of this collection. The GoldMiner search engine is interesting because its users are very specialized. Thus, it represents the particular case of catering to experts in a narrower domain inside medicine. As GoldMiner is so specialized, the number of health consumers using it is likely small. It is, therefore, a good example of the extreme specialization end of the expert continuum, allowing the effects of this specialization on the vocabulary and search behavior of the users to be found.

3.1.2 Sorting the Data by Expertise Level

We make the assumption that experts and health consumers are more likely to use different search engines to satisfy their information needs. Therefore we assume that almost all queries issued into a particular search engine are issued by only one of the two classes of users under consideration. This assumption is justified as we are using search logs from search engines clearly aimed at users of specific expertise. This assumption is also more inclusive than another assumption that has been used to separate medical experts from consumers: that only searches leading to PubMed were done by medical experts [207]. As discussed earlier, this assumption would only tend to detect health researchers, as health practitioners make less use of PubMed [113]. We do not take into account that many users are in between consumers and experts as levels can vary.

On one extreme, we have AOL health consumer users. There might be a few experts using AOL, but their queries are drowned in the consumer queries. Also focused on patients, HON is a search engine for consumers searching for reliable health information. The primary target audience is health consumers concerned about the reliability of the information they access. On the other extreme, mainly targeting health experts looking for health/medical evidence, the TRIP database can also be accessed by consumers but these few consumers might be already considered specialists in their diseases. Finally, the GoldMiner search engine is made by radiologists and for radiologists, consumers have practically no use for this kind of information, but a variety of health experts might access the system. We position each dataset on an expertise axis in Figure 3.2 to help in understanding how each dataset relates to each other.

⁴<http://www.tripdatabase.com/>

⁵<http://goldminer.rrs.org>

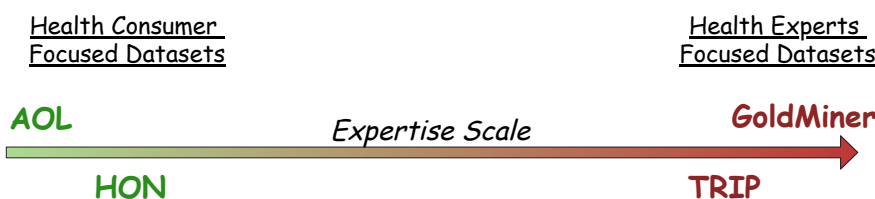


Figure 3.2: The datasets used here are placed on an expertise scale. The expertise level increases as a dataset is placed more to the right-hand side of the scale.

3.1.3 Preprocessing Log Files

The first challenge dealing with different sources of logs is normalizing them. Unfortunately, the clickthrough URL information is available only for the AOL and HON datasets, limiting a detailed click analysis. Thus, we focus on a query content analysis, using only the intersection of all possible fields: (1) timestamp, (2) anonymous user identification, and (3) keywords. Neither stop word removal nor stemming was used.

Sessions were defined as follows. They begin with a query and continue with the subsequent queries from the same user until a period of inactivity of over 30 minutes is found. This approach for sessions, as well as the 30-minutes threshold, is widely used in the literature [27, 209, 104]. We excluded extremely prolific users (over 100 queries in a single session) since they likely represent “bots” rather than individuals.

3.2 Enriching the Query Logs with MetaMap

The US National Library of Medicine MetaMap is intensively used in this work to enrich the information contained in the query logs, adding annotations regarding the concepts searched in the queries. MetaMap is widely employed to map biomedical text to the Unified Medical Language System (UMLS) Metathesaurus, a compendium of many controlled vocabularies in the biomedical sciences [4]. This type of mapping has been already used for many different tasks in the health domain, such as query expansion [7, 107], concept identification and indexing [5, 137], question answering [51], knowledge discovering [202], and more related to this work, to enrich query logs to understand user goals [91, 96]. To explain how mapping queries to UMLS can give us some insights about the user intent, we first have to explain what UMLS is and how MetaMap maps text to UMLS. We explain how the mapping works in the next section and we evaluate the mapping in Section 3.2.2.

3.2.1 MetaMap

The Unified Medical Language System (UMLS) Metathesaurus is a multi-purpose, multi-lingual vocabulary database, containing information about biomedical and health related concepts. A Metathesaurus can be defined as a very large, multi-purpose, and multi-lingual vocabulary resource that contains information about biomedical and health related

concepts, their various names, and the relationships among them [135]. UMLS is updated quarterly with new vocabularies and currently contains 169 different sources⁶. Altogether, UMLS comprises more than 1 million biomedical concepts. In its 2013 version, the UMLS Metathesaurus has more than one hundred different controlled vocabulary sources and a large number of internal links, such as alternative names and views of the same concept.

1. + Anatomy [A]
2. + Organisms [B]
3. - Diseases [C]
 - [Bacterial Infections and Mycoses \[C011\]](#) +
 - [Virus Diseases \[C021\]](#) +
 - [Parasitic Diseases \[C031\]](#) +
 - [Neoplasms \[C041\]](#) +
 - [Musculoskeletal Diseases \[C051\]](#) +
 - [Digestive System Diseases \[C061\]](#) +
 - [Stomatognathic Diseases \[C071\]](#) +
 - [Respiratory Tract Diseases \[C081\]](#) +
 - [Otorhinolaryngologic Diseases \[C091\]](#) +
 - [Nervous System Diseases \[C101\]](#) +
 - [Eye Diseases \[C111\]](#) +
 - [Male Urogenital Diseases \[C121\]](#) +
 - [Female Urogenital Diseases and Pregnancy Complications \[C131\]](#) +
 - [Cardiovascular Diseases \[C141\]](#) +
 - [Hemic and Lymphatic Diseases \[C151\]](#) +
 - [Congenital, Hereditary, and Neonatal Diseases and Abnormalities \[C161\]](#) +
 - [Skin and Connective Tissue Diseases \[C171\]](#) +
 - [Nutritional and Metabolic Diseases \[C181\]](#) +
 - [Endocrine System Diseases \[C191\]](#) +
 - [Immune System Diseases \[C201\]](#) +
 - [Disorders of Environmental Origin \[C211\]](#) +
 - [Animal Diseases \[C221\]](#) +
 - [Pathological Conditions, Signs and Symptoms \[C231\]](#) +
 - [Occupational Diseases \[C241\]](#) +
 - [Chemically-Induced Disorders \[C251\]](#) +
 - [Wounds and Injuries \[C261\]](#) +
4. + Chemicals and Drugs [D]
5. + Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. + Psychiatry and Psychology [F]
7. + Phenomena and Processes [G]
8. + Disciplines and Occupations [H]
9. + Anthropology, Education, Sociology and Social Phenomena [I]
10. + Technology, Industry, Agriculture [J]
11. + Humanities [K]
12. + Information Science [L]
13. + Named Groups [M]
14. + Health Care [N]
15. + Publication Characteristics [V]
16. + Geographicals [Z]

Figure 3.3: MeSH hierarchy with the Disease branch expanded

The white row of Table 3.2 is the original version of the classical UMLS example from [135]. It illustrates how different atoms can have the same meaning. Atoms are the basic building blocks from which the Metathesaurus is constructed, containing the concept names or strings from each of the source vocabularies. The atoms shown are part of two vocabularies PSY (Psychological Index Terms), and MSH (Medical Subject Headings, MeSH), mapping different strings and terms to the same concept, C0004238, which states that atrial fibrillation is a pathological function. The other row of this table shows another concept, C1963067, mapped from the vocabulary NCI (National Cancer Institute), which states that atrial fibrillation can be an adverse event associated with the use of a medical treatment or procedure, although we do not know which medical treatment or procedure.

⁶<http://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html>

3. INVESTIGATING HEALTH SEARCH THROUGH QUERY LOGS

Table 3.2: A concept is potentially linked to various AUI (atom), SUI (string), and LUI (term). We used MetaMap to map a user query, e.g. “Atrial Fibrillation” to the different existing concepts (C0004238, C1963067). Note that each concept is associated to one single semantic meaning.

Concept (CUI)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs)
C0004238 [Pathologic Function] Atrial Fibrillation (preferred) Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	L0004238 Atrial Fibrillation (preferred) Atrial Fibrillations	S0016668 Atrial Fibrillation (preferred)	A0027665 Atrial Fibrillation (from MSH)
		S0016669 (plural variant) Atrial Fibrillations	A0027667 Atrial Fibrillation (from PSY) 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)
C1963067 [Finding] Atrial fibrillation (Atrial Fibrillation Adverse Event)		Auricular Fibrillations (from NCI)

The task of MetaMap is to map biomedical text to its corresponding concept(s). MetaMap generates a candidate set for a piece of text, based on its internal parser and variant generation algorithm, which takes into account acronyms, synonyms, inflections and spelling variants of the text. Then, based on metrics such as centrality, variation, coverage and cohesiveness, MetaMap ranks each candidate [4]. Occasionally, more than one candidate may have the same score. We collect all the top candidate(s) and its (their) associated semantic type(s), shown in bold below the CUIs in Table 3.2. In the running example, a text containing only ‘atrial fibrillation’ is mapped to both C0004238 and C1963067 with the same top score, and the types ‘Pathologic Function’ and ‘Finding’ are assigned to the query. To the best of our knowledge, MetaMap is the state of the art for mapping biomedical text to UMLS concepts.

An interesting way to capture the user intent is mapping the queries to a well known domain corpus. In this work we use the Medical Subject Headings, MeSH, as it is a rich and well-structured hierarchy that has already been studied to examine user query logs [91], allowing us to compare the behavior of the users studied here with PubMed users. The whole MeSH hierarchy contains more than 25,000 subject headings in the 2013 version, the one used in this work, containing 16 top categories such as ‘Anatomy’ and ‘Diseases’. Figure 3.3 shows an example of the MeSH hierarchy with the first level of the disease branch expanded.

We use the approach of Herskovic et al. [91] in this chapter, mapping each query onto

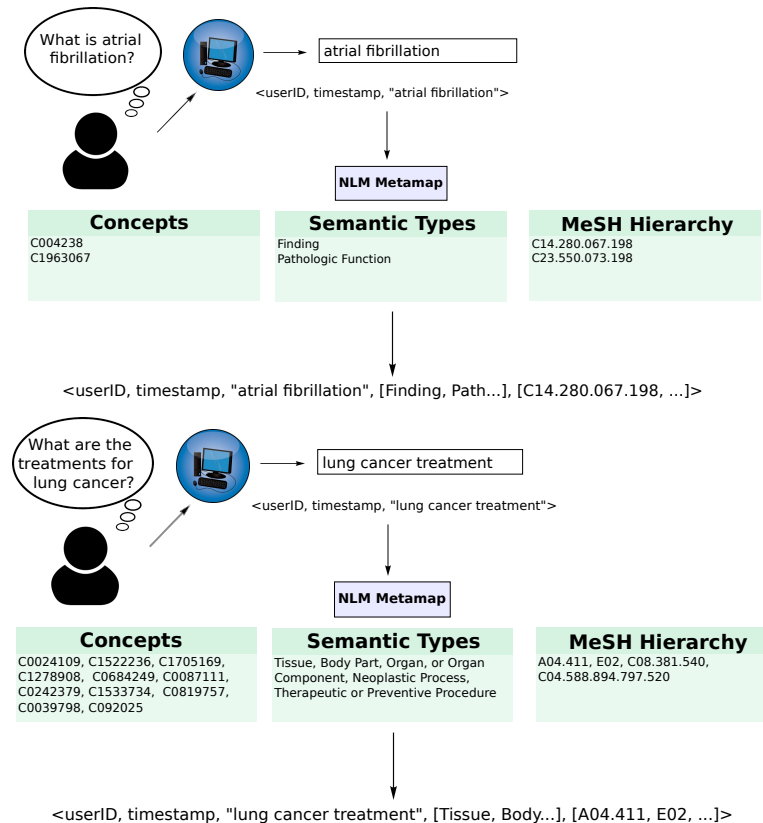


Figure 3.4: Two different user queries are enriched with information extracted with MetaMap. In the top part, the same example used in Table 3.2 is processed by MetaMap. In the bottom part, the query “lung cancer treatment” is more ambiguous and results in different mappings, such as *Lung (Entire lung) / Cancer Treatment (Cancer Therapeutic Procedure)* and *Lung Cancer (Malignant neoplasm of lung) / Treatment (Therapeutic procedure)*

one or more MeSH terms with MetaMap. As shown in Figure 3.4, one query can be mapped to multiple MeSH identifiers. For example, the query ‘atrial fibrillation’ is mapped to both MeSH ids *C14.280.067.198* and *C23.550.073.198*, both in the topmost *Disease* category (represented by the starting letter ‘C’ as shown in Figure 3.3). After the mapping to MeSH is done, we can easily have an overview regarding the subjects the users are more interested in. In this case, we would conclude that this user is interested in diseases, as her/his only query maps only to category ‘C’, more specific in cardiovascular diseases, C14, and pathological conditions, C23.

After preprocessing, each query is converted into the following format: `<timestamp, userID, query, semanticTypes, meshIDs>`, where the *timestamp*, *userID* and *query* are originally query log fields, while *meshIDs* and *semanticTypes* are the set of semantic types and MeSH identifiers generated by MetaMap. These two fields are examined in

detail in Sections 3.3.2 and 3.4.2. Figure 3.4 illustrates how the queries were enriched with the information provided by MetaMap and the final format.

Finally, it is important to mention that the queries were mapped to concepts in the UMLS 2013 AA USAbase Strict Data and no special behavior parameter was used. We manually examined the behavior for two important parameters: allowing acronyms/abbreviations (-a) and using the word sense disambiguation module (-y), and decided not to activate them. Our experiments show that activating the former parameter decreases the precision significantly for the sake of a small increase in recall, as MetaMap is already capable of matching some of the most frequently used abbreviations (HIV, HPV, AIDS, COPD). For the latter, we have an inverse scenario, where we had a small gain in precision but a larger loss in recall, as MetaMap always picks only one possibility when more than one concept is possible. It means that MetaMap would be forced to choose between concepts C0004238 and C1963067 of Table 3.2, even when both are equally likely. The last important reason for not using any other parameter is that we want to compare our results with [91], in which no special option was used either. For the experiments shown in Section 3.3.2 we used the parameter (-R) to restrict MetaMap to use only MeSH as vocabulary source.

3.2.2 Evaluation of the MetaMap Mappings

As recently reported by MetaMap’s authors [6], a direct evaluation of MetaMap against a manually constructed gold standard mapping to UMLS concepts has almost never been performed. Usually, indirect evaluations are made, where the effectiveness of a task is measured with and without MetaMap. For example, query expansion using the related concepts of a concept identified by MetaMap versus not using it. Here we are interested in the few articles that evaluate the effectiveness of MetaMap, especially the ones focused on mapping user queries.

In 2003, Pratt and Yetisgen-Yildiz [165] compared MetaMap mappings to UMLS with mappings made by six physicians and nurses. For the 151 concepts in their ground truth, MetaMap could match 81 concepts exactly, 60 partially and could not match only 10 concepts, of which 6 were not available in UMLS. In a scenario considering partial matches (e.g., mapping to *angiomatosis* instead of *leptomeningeal angiomatosis*), MetaMap had an F_1 of 76%. In another experiment in the same year, Denny et al. [52] built a bigger gold standard dataset of 4,281 concepts to evaluate MetaMap, reaching a precision of 78%, recall of 85% and F_1 of 81%.

More recently, Névél et al. [137] reported results on using MetaMap to detect disease concepts on both literature and query corpus. The results showed that MetaMap had a better effectiveness for long sentences (F_1 of 76%) than for short queries (F_1 of 70%), but they also pointed out that the average inter-annotator agreement of the 3 assessors for the query corpus was 73%, showing that MetaMap results are not far from humans performing the same task. Using 1,000 queries from partly the same datasets that are used here: AOL, HON and TRIP, we also showed in [140] an F_1 of 70% for query mappings.

Név  l et al. [136] created an annotated set of 10,000 queries that were mapped to 16 categories, in a similar way to what is done in Section 3.4.2, where the semantic types produced by MetaMap are used to define our own categories. We used N  v  l’s dataset to calibrate our mappings for our *Cause* and *Remedy* categories (see Section 3.4.2), as well as to make decisions regarding MetaMap’s parameters. We used the *Disorder* category of N  v  l as an equivalent of our *Cause* category, and we combined *Chemical and Drugs* (antibiotic, drug or any chemical substance), *Gene, Proteins and Molecular Sequences* (name of a molecular sequence) and *Medical Procedures* (activity involving diagnosis, or treatment) as the closed possible class of our *Remedy* class. We could reach an F_1 of 78% for the *Cause* category (Precision=75%, Recall=81%) and 72% for *Remedy* (Precision=70%, Recall=73%). These figures are in line with what is known for MetaMap when mapping medical abstracts to concepts (i.e., [137, 147]), encouraging us to use it for mapping short queries to concepts as well.

3.3 Individual Query Analysis

One goal of this section is to study how users search, based first on simple statistics to model their behavior. Also, we start exploring the content of their queries, but considering all the queries without dividing into user sessions.

3.3.1 How Users Search

We start by showing a few simple but important statistics about the logs. This section aims to understand the user behavior through general statistics, as well as to show how each log is composed. In Table 3.3 we depict several metrics that are used to characterize user interactions and compare their values to those in related studies. Torres et al. [54] use AOL logs to study queries performed by kids. White et al. [207] use a keyword-based method to filter domain-specific queries and divide them into those issued by consumers and those issued by experts. Their work also considers other types of queries, such as queries on computer science or financial information. We show only the data for the medical domain. Herskovic et al. [91] and Dogan et al. [96] analyze different periods of PubMed logs. For all datasets, “N/A” is used when the information is not available.

The query logs from the related work shown in Table 3.3 belong to the same period as the AOL logs. Query logs from HON, TRIP and GM are considerably newer than the others. Nevertheless, Table 3.3 shows that AOL-Medical and HON are very similar in many aspects, such as the average number of terms per query and the average time per session. The biggest difference between these two query logs was found for the average number of queries per session, however the difference is small (2.71 for AOL-M and 1.80 for HON) if compared to any other datasets (e.g., 5.20 for TRIP and 8.76 for AOL-Kids). In this aspect, HON query logs are more similar to enterprise and local Web site search [114]: for example, the logs from the University of Essex analyzed by Kruschwitz et al. [115, 116] contained on average 1.53 queries per session, and logs of queries submitted to the Utah government Web site showed on average 1.73 queries per session. In turn, the number of

queries per session for AOL-M (2.71) and AOL-NM (3.26) are similar to other general search engines at the time, such as AltaVista (2.02 [178]), Excite (2.8 [99]) and Dogpile (2.31 [100]).

The average terms/characters per query can be an indicator of the complexity and difficulty of the users to express their information needs. We note that AOL-Medical and HON queries are shorter than TRIP queries, and that TRIP logs are similar to PubMed logs in terms of query length. White’s work also found that expert queries are more complex than consumer queries. Note that the recent increase in the number of spoken queries submitted to search engines due to the popularity of smartphones has likely increased the average terms/characters per query [46, 84]. Experiments with more recent search logs are necessary to understand if expert queries continue to be longer than non-expert queries.

The average number of queries per session and time per session, although considerably smaller than what was found by White’s work, follow the same pattern, with TRIP data having longer sessions than HON and AOL-Medical. We could not find an explanation for such long sessions in the White et al. dataset. We show only the sessions made by experts and consumers in the medical domain from White’s work, but in their original paper they report that sessions are considerably smaller when the same set of people query in other domains: having a mean session length of fewer than 5 queries, and the mean time per session is never longer than 800 seconds.

We aggregate the log into two groups in Table 3.4: *consumers* and *experts*, making the comparison of our datasets with the literature possible. As done by White et al. [207], we use Cohen’s d to determine the effect size of each variable between each pair of groups. We randomly sampled 45,000 users from TRIP and merged them with the 45,090 users from GoldMiner, making all datasets have a comparable number of users. Cohen’s d is a useful metric for meta-analysis [35] that uses the means and standard deviations of each measurement to calculate how significant a difference is. Although there are controversies about what is a “small”, “medium” or “large” effect size, a recommended procedure is to define a Cohen’s d effect size of 0.2 or 0.3 as a “small” effect, around 0.5 as “medium” effect and greater than 0.8 as a “large” effect [35]. White et al. built a classifier to detect user expertise based on a superset of the features shown in Table 3.4. They argued that these are valuable features based on Cohen’s d value, as well as feature importance calculated by their regression classifier. Although considered to have a “small” effect, this was big enough to help separate experts from consumers. We reached very similar Cohen’s d values to White’s paper, hypothesizing that the behavior could be used to predict expertise in other logs as well. In particular, we found the same ranking that White et al. found, among the four features presented in Table 3.4.

3.3.2 What Users Search for

In order to understand what search engine users are looking for, we investigate popular terms and queries issued. Also, we use MetaMap to map queries to the MeSH hierarchy,

Table 3.3: General Statistics describing the query logs. We use * when the median value was used instead of the average, and N/A when the data was not available.

Dataset	This Work					Literature					
	Consumers		Experts		Non-Medical	AOL-Kids	AOL-NKids	Consumers	Experts	Pubmed	
	AOL-M	HON	TRIP	GM	AOL-NM	Torres et al. [54]	White et al. [207]	Herskovic et al.[91]	Dogan et al.[96]		
Logs Initial Date	Mar 2006	Dec 2011	Jan 2011	N/A	Mar 2006	Mar 2006	May 2007	Jan 2006	Mar 2008		
Logs Final Date	May 2006	Aug 2013	Aug 2012	Jan 2012	May 2006	May 2006	Jul 2007	Jan 2006	Mar 2008		
# Users	47,532	47,280	279,280	45,090	655,292	N/A	N/A	37,243	7,971	624,514	NA
# Queries	215,691	140,109	1,788,968	219,407	34,427,132	485,561	N/A	673,882	362,283	2,689,166	58,026,098
# Unique Queries	69,407	85,824	486,431	90,766	9,695,882	10,252	N/A	N/A	N/A	N/A	N/A
# Sessions	79,711	77,977	344,038	100,848	10,555,562	21,009	N/A	68,036	26,000	740,215	23,017,461
Avg Terms Per Query	2.61 (± 1.71)	2.72 (± 2.05)	3.40 (± 2.33)	2.28 (± 2.54)	2.46 (± 1.87)	3.23	2.5	2.92	3.30	3*	3.54
Avg Char Per Query	16.22 (± 9.11)	18.11 (± 11.48)	21.22 (± 9.69)	16.64 (± 10.20)	15.98 (± 9.67)	N/A	N/A	20.76	24.05	N/A	N/A
Avg Queries Per Session	2.71 (± 2.50)	1.80 (± 2.48)	5.20 (± 5.95)	2.18 (± 2.57)	3.26 (± 4.65)	8.76	2.8	9.90	13.93	N/A	4.05
Avg Time Per Session (s)	258 (± 531)	208 (± 592)	471 (± 758)	163 (± 520)	384 (± 809)	1238	N/A	1549.74	1776.45	N/A	N/A

Table 3.4: General Statistics – Stratified by expertise. C for consumers and E for experts

Dataset	Consumers	Experts	Cohen's d	
Total Number of Users	94,812	90,090	E - C	E - C from [207]
Total Number Of Queries	355,800	504,745		
Total Number Of Unique Queries	149,648	181,051		
Total Number Of Sessions	157,688	155,965		
Mean Terms Per Query	2.65 (± 1.85)	2.91 (± 2.09)	0.13	0.20
Mean Chars Per Query	16.97 (± 10.16)	19.18 (± 10.16)	0.22	0.30
Mean Queries Per Session	2.26 (± 2.53)	3.24 (± 4.29)	0.28	0.38
Mean Time Per Session (sec)	233 (± 562)	271 (± 629)	0.06	0.11

finding the high-level topics associated with the user queries.

Terms and Queries

We depict the most popular queries, terms (here excluding the stop words), and abbreviations used in all logs, as well as their frequency among the queries in Table 3.5. As expected, AOL-NotMedical contains navigational queries and several terms related to entertainment. Similarly, some of the most popular queries in AOL-Medical are navigational: in the top 10 queries, two are for the website *webmd.com* and one searching for the website *mayoclinic.com*. These same websites also appear in the HON search log. The analysis of AOL-Medical terms shows common medicine-related concepts, with people searching for information about different cancer types in more than 3% of the cases.

Most of the top queries in the TRIP log are related to diseases. In TRIP logs, we found ‘*area:*’ in 3% of the queries, ‘*title:*’ in 2.2%, ‘*to:*’ in 1.5% and ‘*from:*’ in 1.8%, in total these keywords were used in 6.7% of the queries, however, we do not show these terms in Table 3.5, as they do not reveal what the users search for, but how they search. These patterns were not found in the other datasets. The use of more advanced terms is also found in PubMed logs [91], we hypothesize that some users might just copy and paste their queries from PubMed into the TRIP search engine, resulting in queries such as ‘*palliative care (area:oncology)*’, indicating that the user wants material about palliative care specifically for the area of oncology. ‘*Title*’ is used in PubMed for performing a search only in the title of the indexed articles, while ‘*from:*’ and ‘*to:*’ specify periods of

3. INVESTIGATING HEALTH SEARCH THROUGH QUERY LOGS

Table 3.5: Top queries and terms and their relative frequency (%) among all queries

Rank.	Laypeople				Experts				AOL-NotMedical	
	AOL-Medical		HON		TRIP		GoldMiner		String	Freq
	String	Freq	String	Freq	String	Freq	String	Freq	String	Freq
QUERIES										
1	webmd	0.98	trustworthy health sites	4.24	skin	0.29	mega cisterna magna	0.44	google	0.95
2	web md	0.41	cancer	0.51	diabetes	0.22	baastrup disease	0.40	ebay	0.40
3	shingles	0.27	webmd	0.47	asthma	0.17	toxic	0.23	yahoo	0.37
4	mayo clinic	0.26	sleep apnea syndromes	0.27	hypertension	0.14	limbus vertebra	0.22	yahoo.com	0.28
5	lupus	0.25	lymphoma	0.22	stroke	0.13	cystitis cystica	0.20	mapquest	0.25
6	herpes	0.20	breast cancer	0.21	osteoporosis	0.11	thornwaldt cyst	0.14	google.com	0.23
7	diabetes	0.19	hypertension	0.18	low back pain	0.10	buford complex	0.13	myspace.com	0.22
8	fibromyalgia	0.18	mayoclinic.com	0.16	copd	0.10	splenic hemangioma	0.13	myspace	0.21
9	pregnancy	0.16	obesity	0.16	breast cancer	0.09	throckmorton sign	0.12	www.yahoo.com	0.12
10	hernia	0.16	drweil.com	0.14	pneumonia	0.09	double duct sign	0.12	www.google.com	0.12
TERMS										
1	cancer	3.40	health	6.39	treatment	3.03	cyst	3.17	free	1.24
2	hospital	3.00	sites	4.37	cancer	2.56	mri	1.89	google	1.04
3	pain	2.25	trustworthy	4.28	pain	2.13	disease	1.80	county	0.65
4	symptoms	2.14	cancer	2.74	care	2.10	ct	1.75	yahoo	0.62
5	disease	2.03	disease	1.53	children	1.98	fracture	1.68	pictures	0.60
6	blood	1.87	diabetes	1.17	therapy	1.81	tumor	1.65	lyrics	0.52
7	medical	1.62	treatment	0.96	diabetes	1.80	syndrome	1.47	school	0.51
8	webmd	1.21	syndrome	0.87	disease	1.78	liver	1.26	myspace	0.49
9	surgery	1.14	heart	0.83	pregnancy	1.70	pulmonary	1.22	ebay	0.46
10	syndrome	1.13	pain	0.80	acute	1.41	bone	1.16	sex	0.44
11	breast	1.11	care	0.77	syndrome	1.39	renal	1.13	florida	0.45
12	center	1.09	effects	0.75	management	1.14	sign	1.12	sale	0.41
13	health	1.04	medical	0.67	stroke	1.07	lung	1.11	city	0.40
14	heart	0.90	blood	0.65	surgery	1.06	brain	1.08	home	0.39
15	diabetes	0.86	pregnancy	0.61	prevention	1.05	cell	1.00	state	0.39

time in which a document was published.

The topmost query in the HON log and its top 3 terms are ‘*trustworthy health sites*’. It shows that many of the queries are from users that do not know which are the medical websites that they can trust, and also demonstrates a misunderstanding by the end users of the nature of the content indexed by the HON search engine (only HONcode-certified websites are indexed).

For the GoldMiner queries and terms, we clearly see the increase in the terminological specificity of the most popular keywords used.

Mapping to MeSH

MeSH is a hierarchical vocabulary used by US National Library of Medicine for indexing journal articles in the life sciences field. A query log analysis using MeSH was also carried out by Herskovic et al.[91] for the PubMed logs in order to understand what are the most popular topics searched by the users. We use the same weighting schema used in Herskovic’s work: if n categories are detected in one query, we give the weight of $1/n$ to these categories.

General statistics calculated for the mapping of user queries to MeSH terms are shown in Table 3.6. Here, we are testing MetaMap for the annotation of non-medical queries as

Table 3.6: General MeSH Statistics

Metric	Laypeople		Experts		
	AOL-M	HON	TRIP	GM	AOL-NM
% of queries containing MeSH terms	77.87	77.81	85.96	79.02	50.51
Mean MeSH Depth	3.99	3.83	3.86	4.01	4.37
Mean MeSH terms per query	2.14	2.19	2.78	2.07	1.12
Mean Disease terms per query	0.81	0.60	0.99	1.17	0.05

well, which to the best of our knowledge was never studied.

An interesting result is the fact that around 50% of AOL-NotMedical queries were successfully mapped to a MeSH concept. To investigate this, we collected a large random sample of mapped queries and analyzed them. We found that MetaMap is able to find many concepts not directly linked to medicine, such as geographic locations, animals and plants, food and objects. For example, ‘www’ (*L01.224.230.110.500*), used in 10% of all AOL queries, is recognized and annotated as *Manufactured Object*. Also, locations are usually very commonly found and help to explain the high mean MeSH depth found for this dataset, second row in column AOL-NM (California is mapped to both *Z01.107-.567.875.760.200* and *Z01.107.567.875.580.200*). It is important to have this in mind when building systems like in Yan et al. [213], in which the MeSH depth is used to model document scope and cohesion. When looking at false positive mappings, especially the ones mapping to diseases and symptoms, we detected that MetaMap’s errors fall into two main categories: (1) English common words: tattoo (tattoo disorder), Pokemon (ZBTB7A gene), and (2) abbreviations: park (Parkinson disease), dvd (Dissociated Vertical Deviation). For both types of errors, MetaMap, or a system using it, would have to use the context (words around the mapping) to detect that Pokemon is used as a cartoon or a game, and not as a gene name. Specifically for the second case, it would be desirable if MetaMap could allow the use of a pre-defined list of acronyms to increase its precision. In the current implementation, MetaMap has a parameter for user-defined acronyms (-UDA), but it is just used to expand more acronyms instead of overwriting its pre-defined ones. Also for AOL-NM, the third and fourth rows indicate the suitability of using mappings to MeSH for distinguishing between medical and non-medical queries. Queries from the medical logs have a larger number of MeSH terms and disease terms than AOL-NM. If the errors analyzed above could be amended using the query context or session, for example, then mapping to MeSH could be helpful to detect queries or sessions on medical information.

Going further, we present in Figure 3.5 the most popular categories for the first level of the MeSH hierarchy. We also show the results obtained by Herskovic et al. [91] for PubMed, in order to compare our findings. We show only the categories that have more than 5% of the queries containing MeSH terms mapped to it.

When Herskovic and colleagues did this experiment, they found that PubMed users were more interested in the category *Chemical and Drugs*. In general, the distributions over the categories for the AOL-Medical, HON and TRIP search logs are similar. However,

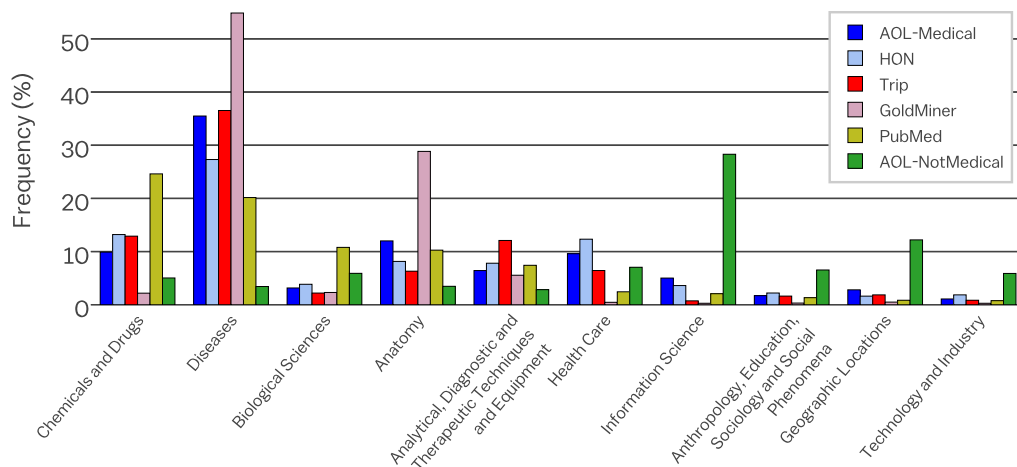


Figure 3.5: Popular categories according to MeSH mappings

differently from PubMed, we found that the users are generally most interested in *Diseases*, and then *Chemicals and Drugs*. The results for GoldMiner show another trend for the second most popular category, focused on anatomy rather than on drugs, likely because radiologists often have to append to their query the part of the body that they are interested in. In its actual version, GoldMiner has a filter for age, sex and modality (e.g., CT, X-ray), but it has no filter for body parts. This analysis suggests that it could be interesting to add a filter for body regions as well.

Last, the four classes to the right of Figure 3.5 partly explain the high percentage of AOL-NotMedical terms mapped to MeSH terms. The high percentage of mappings to these least medical categories of MeSH (e.g., *Information Science* and *Technology and Industry*), together with a low percentage of mappings to highly medical categories (e.g., *Diseases*), could be used as a discriminative feature to distinguish between medical and non-medical logs.

3.4 Analysing Sessions

The rest of this chapter considers user sessions instead of separate queries. Once more, we first study the user behavior, then the content of each session.

3.4.1 Session Characteristics

A series of queries, part of an information seeking activity, is defined as a session [72]. In order to better understand sessions, we define and study a set of search patterns. In this work, we consider that, after issuing the first query, a user may act in four different ways: (1) *Repetition*: repeat precisely the same query, (2) *Expansion (Specialization)*: repeat

Table 3.7: Aggregated percentages for query modifications along the sessions

Action	Laypeople		Experts		AOL-NM
	AOL-M	HON	TRIP	GM	
Expansion	6.66	13.83	14.85	5.96	3.71
Reduction	1.23	2.23	4.35	9.61	0.84
Reformulation	84.74	63.56	43.96	49.56	80.27
Exp. Red.	0.37	1.29	5.09	3.54	0.57
Exp. Ref.	5.43	13.90	15.27	8.28	9.66
Red. Ref.	1.01	2.21	5.63	12.01	2.09
Exp. Red. Ref.	0.56	2.98	10.85	11.04	2.86

the query adding one or more terms to increase precision, (3) *Reduction (Generalization)*: reduce the number of terms to increase recall, or (4) *Reformulation*: reformulate the query changing some or all the terms used. We ignore the first case because we cannot be sure if a user is indeed repeating the same query or just changing the result page, as some search engines record the same query as a result of a page change. Note that multiple search patterns have been employed in the literature [119, 88, 100, 94, 72] depending, among other factors, whether the semantics of the queries are taken into account or not. The search patterns employed in this work are a common subset of the different patterns definitions in the literature.

Table 3.7 depicts the changes made by users during the sessions. If during one single session a user adds a term to the previous query and then changes a few words, we count one action in the row Exp.Ref (for expansion and reformulation – the order is not important). In the end, we divide the number of actions of each row by the total actions in the query log. Hence, Table 3.7 shows that the most frequent user action is the reformulation alone, but it is more likely to happen in search engines targeting consumers, e.g., 84% of the sessions in the AOL-Health logs and 63% of HON had only reformulations. The last row of Table 3.7 shows that expert users might be more persistent than consumers, as more than 10% of the sessions in the expert search engines are composed of every type of action, while in consumers logs this number is less than a third of this. In the literature, White et al. [207] also hypothesize that expert users are more persistent than consumers.

3.4.2 What are the Sessions About?

In order to understand the sessions, we first attribute meaning to the users' individual queries, mapping them to search intents. We use the same search intents previously defined in Cartright, White and Horvitz [27]: symptom, cause and remedy, so that a direct comparison can be performed. A difference between their method and ours is that we map the queries in a session to search intents by using the UMLS semantic types of MetaMap, as done in [97, 136, 147, 140], instead of handmade rules.

In Figure 3.6, we show ten semantic types that are frequently found in the query logs (i.e., semantic types found in at least 5% in any query log). We inspect only these ten

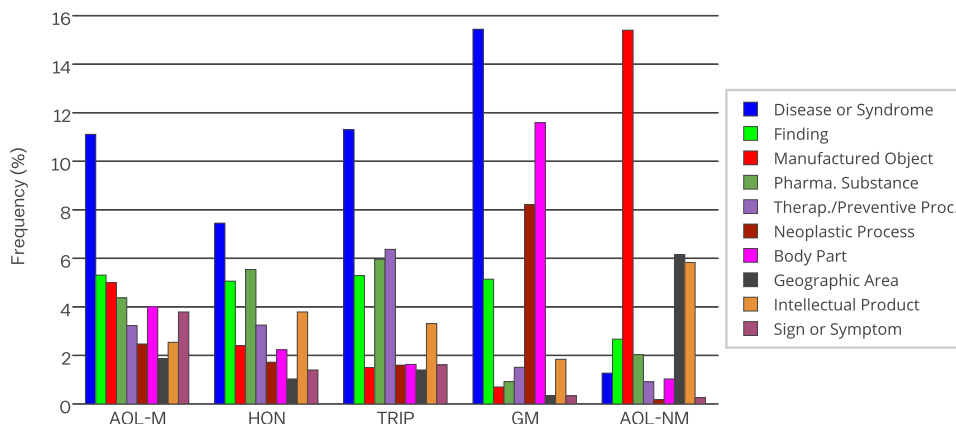


Figure 3.6: The top most frequently used semantic types (frequency in percentage). Many of the most used types are aggregated to study the user focus described in Table 3.8

semantic types for a matter of readability, as currently MetaMap recognizes 133 semantic types and it is not possible to visualize them all⁷. The single most common type in all the medical logs is ‘*Disease and Syndrome*’. As we expect, the top types in AOL-NotMedical are not related to the medical domain, and the second most common semantic type for GoldMiner is related to parts of the body, as one might expect for radiology queries.

After a meticulous analysis of the semantic meaning assigned for the queries (as previously described in experiments made in Section 3.2.2), we defined the following classification based on the three classes created in [27] (some examples of queries classified for each type are given for a better understanding):

- **Symptom:** Sign or Symptom (*cough; sore; headache; red eyes*), Findings (*stress; testicular cyst*)
- **Cause:** Anatomical Abnormality (*hiatal hernia*), Cell or Molecular Dysfunction (*macrocytos*), Congenital Abnormality (*scoliosis*), Disease or Syndrome (*diabetes; heart failure*), Experimental Model of Disease (*cancer model*), Injury or Poisoning (*achilles tendon rupture*), Mental or Behavioral Dysfunction (*bipolar disorder*), Neoplastic Process (*lung cancer; tumor*), Pathologic Function (*atypical hyperplasia*)
- **Remedy:** all 28 types belonging to the high-level group Chemicals & Drugs, which includes Clinical Drug (*cough syrup*), Antibiotic (*penicillin*), Pharmacologic Substance (*tylenol; mietamizol*), Amino Acid, Peptide, or Protein (*vectibix; degarilex*), Immunologic Factor (*vaccine; acc antibody*), Vitamin (*quercetin, vitamin B12*), Therapeutic or Preventive Procedure (*treatment; physiotherapy*), etc.

⁷A complete list of all semantic types can be found Online: <http://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml>

We analyze the most popular semantic types found in the queries and show them in Table 3.8, together with a direct comparison to Cartright et al. [27]. The largest difference between all four medical logs analyzed in this paper and the Cartright et al. results is in the symptom category. For the latter, 63.8% of the sessions are focused on symptoms, while between 5.5% and 9.1% are focused on symptoms in our analysis. The main reason for Cartright’s result is linked to the way in which they created their dataset: *keeping only sessions that had at least one query containing a term in a wordlist extracted from a **list of symptoms** from the Merck medical dictionary*. Their preprocessing step, therefore, explains the fact that most of the sessions were concentrated only on searching for symptoms. Conversely, our analysis reveals that the most common user focus is on causes rather than on symptoms. Also, the second most common focus is on the way to cure a disease. It is important to note that the Cartright et al. logs date from 2009, which means they are 3 years younger than AOL, but roughly 3 years older than HON, also suggesting that the considerable divergence found is due to the preprocessing steps and not to an evolution on how the users search.

Once more, GoldMiner presents a different behavior. We hypothesize that the low number of sessions on remedies is explained by the fact that radiologists are not interested in remedies when searching for images as they are rather in the diagnosis phase. It is interesting to note that searching for causes and remedies in the same session is a very frequent task for medical experts in the TRIP logs, with 16% of the sessions searching for both remedies and causes.

In Table 3.9, we show the behavior modifications along a session. One oscillation is characterized by a transition from one focus type to another and then back to the original type. Originally, this study was made to support the hypothetico-deductive searching process in which a user cyclically searches for a symptom, then a cause and then returns to symptom [27]. The symptom-cause pattern was also found in our experiments, but with a more balanced distribution in relation to the other patterns. Again, the large number of behaviors involving symptoms found in [27] is likely an artifact of how the dataset was constructed. We see that the cause-remedy pattern plays a very important role, especially in the TRIP log, in which this is the most common pattern. Finally, the least frequent pattern found in all four datasets is the symptom-remedy one. The study of the behavior modification was used in [27] to build a classifier to predict what is the next user action, allowing a search system to support medical searchers by pre-fetching results of possible interest or suggesting useful search strategies.

3.5 Summary

In this chapter, we focus on one of the most important topics in the health domain, medicine. We conducted a detailed study of medical information search behavior through query logs. We studied how users search for medical documents, as well as what they search for. Results were compared to those in published studies analyzing search logs in the medical domain. Almost all recent studies about the behavior of searchers looking

Table 3.8: User focus when searching for medical content in a single session.

Intent	Laypeople		Experts		AOL-NotMedical	Cartright et al.[27]
	AOL-Medical	HON	TRIP	GoldMiner		
None	34.0	40.4	16.8	21.2	82.9	3.9
Symptom	9.1	6.3	5.5	6.4	3.9	63.8
Cause	24.3	20.9	26.0	58.2	3.3	5.3
Remedy	14.7	16.2	17.4	3.3	7.5	1.1
Symptom and Cause	6.8	6.1	7.2	6.4	0.5	22.6
Symptom and Remedy	2.1	2.6	4.5	0.5	0.9	2.0
Cause and Remedy	7.1	5.0	15.9	3.0	0.8	0.4
All three	1.9	2.5	6.7	1.0	0.2	0.8

Table 3.9: Cycle Sequence along a single session

Pattern	Interaction	Laypeople		Experts		Cartright et al.[27]
		AOL-Medical	HON	TRIP	GoldMiner	
	Sessions with oscillations (%)	23.07	13.48	64.61	8.60	16.2
Symptom-Cause	Symptom→Cause→Symptom	19.2	15.6	13.2	22.7	51.4
	Cause→Symptom→Cause	19.9	18.8	14.5	35.3	38.4
Symptom-Remedy	Symptom→Remedy→Symptom	8.2	11.8	10.8	4.1	5.1
	Remedy→Symptom→Remedy	8.1	14.2	11.6	3.8	2.7
Cause-Remedy	Cause→Remedy→Cause	18.2	18.4	24.8	20.3	1.5
	Remedy→Cause→Remedy	26.4	21.2	25.1	13.8	0.9

for medical information have been based on the search logs of a large commercial general purpose search engine. This chapter performs the important task of reproducing these studies as far as possible on search logs from other search engines to find out to what extent these results can be supported or not. An important difference with this study compared to published studies is the use, in three of the four cases, of domain-specific medical search engines targeted at either experts or consumers, meaning that we have very strong priors about who is using the search engines and what they are searching for. This avoids assumptions that have to be made in order to extract medical queries or extract expert or consumer queries from the search log of a general purpose search engine.

This chapter covers the behavior of the users when searching for medical information. Analyses were done both at the level of individual queries and of sessions. It was found that the mean number of terms per query and mean number of chars per query were higher for experts than for consumers with a small effect (measured by Cohen's d value). We also found longer sessions both in terms of mean number of queries per session and mean amount of time spent per session, with a small effect. These small effects were sufficient to successfully train a classifier to predict expert and consumer classes in Chapter 4.

When analyzing the user behavior in terms of sessions, we conclude that experts are

more persistent than consumers, as more than 10% of the sessions in the expert search engines were composed of all possible query modification actions (expansion, reduction, reformulation). Alternatively, longer sessions could mean that experts are struggling to find relevant information. It supports the current efforts of the information retrieval community to help experts finding scientific material to improve their clinical decisions [167]. It would be interesting to study if the increase of expertise of consumers can change their behavior over time as suggested by Wildemuth [211], but this will likely require years of search engine logs.

The investigation of what the users search for led us to conclusions that are significantly different from results published in the literature. In both of our analyses, the one based on the MeSH hierarchy and the one based on semantic types, we observed that users are more concerned with diseases rather than symptoms. Understanding what users are searching for is an essential step towards providing more relevant search results.

We also identified patterns supporting the hypothetico-deductive searching processes, especially for the cause-remedy component, in which both consumers and experts cycle through searching for causes and remedies in sessions to discover potential treatments for a disease. Finally, we found that TRIP users, mainly users falling into our expert class, use the hypothetico-deductive method very often, in more than 60% of their sessions, versus less than 25% for AOL and HON. This supports the hypothesis that experts have much more complex information needs, which are not well addressed by the current search systems [167].

An interesting kind of search in the medical domain is the one for self-diagnosis purposes [65], which often arises before consulting medical expert (or to help the decision to consult). Previous research has shown that exposing people with no or scarce medical knowledge to complex medical language may lead to erroneous self-diagnosis and self-treatment and that access to medical information on the Web can lead to the escalation of concerns about common symptoms (e.g., cyberchondria) [208]. Also, current commercial search engines are far from being effective in answering such queries [225], presenting on average only 3 highly relevant results in the top 10 results. In the same manner that experts can assist health consumers in detecting credible content on the Web [174], a search system capable of inferring user expertise can learn about the decisions taken by experts to better support consumers. In the case of self-diagnosis, the symptom-cause cycle in the expert search logs can be explored to provide query suggestions for health consumers.

After consulting a medical expert, health consumers often query about a disease or about a treatment that was recommended to them [65]. When copying-and-pasting the complex terms into a search box, they are presented with documents that are potentially as complex as their queries [107, 225]. Inferring user medical knowledge can help matching health consumers with the suitable documents for them even for complex queries, significantly diminishing harmful situations and misunderstandings. Inferring user medical knowledge is the focus of the next chapter of this thesis.

Exploiting Query Logs: Estimating User Medical Expertise

We have seen in Chapter 3 that health experts and health consumers use different search strategies. Now, we shall take advantage of these differences to build an automatic classifier that can assist search systems to provide the most suitable content to the users based on the user domain knowledge.

In particular, we explore two out of the four query logs studied in Chapter 3: HON and TRIP. Those are the most representative query logs of the health consumers and experts and their search sessions, as GM is too limited to a very specific specialization (radiologists) and the AOLM requires preprocessing that is not needed by HON. Both HON and TRIP query logs are made for a distinct audience, although they share a very similar interface, taking free text queries in a single text box. Our assumption that health experts use TRIP and health consumers use HON does not require any complex filtering of users, as most users query medical content on these Websites.

One concern that arises when using two different sources of logs is that we could learn how to differentiate between the two search engines, instead of learning how to infer the correct user expertise. To overcome this drawback, we are focusing on *what* the users search for (e.g., analyzing the keywords used to find out the topics searched), rather than *how* (e.g., the number of words used or the number of queries per session). Therefore, we do not make use of user sessions and use behavior features only as baseline, rather we focus on creating and evaluating features for expertise prediction in the health domain based solely on the keywords used. As an outcome, we built a classification model capable of inferring user medical expertise that can be easily integrated into any search engine. The results show that a Random Forest classifier using the medical features proposed

can boost the classification accuracy by more than 12%, compared to the same classifier using only user behavior related features.

This chapter is divided as follows. Section 4.1 presents the datasets used in the classification task, while Section 4.2 lists all the features devised. Section 4.3 analyses the classification results. A summary and conclusion are presented in Section 4.4.

4.1 Data Collection

Our dataset consists of query of two types of users: users from the HON search engine, the health consumers; and users from the TRIP search engine, the health experts. The general statistics about the dataset were described in Table 3.3. In order to create a balanced dataset, we randomly selected 25,000 users (i.e., 25,000 query sessions) from both HON and TRIP datasets. Note that we opted for a balanced dataset as we are not aware of the actual distribution between health consumers and professionals in commercial search engines. While it has been observed that the naturally occurring distribution is not always the optimal distribution to train classification models [204, 30], experiments with different sampling strategies are beyond the scope of our work.

Only two primary pieces of information are extracted for each query: (1) the anonymous user identification, and (2) the keywords used. This information is the common intersection of the two query logs used, and potentially present in any other query log of a search engine or Q&A system. As done in Chapter 3, we enrich the user query using Metamap [4]. We augment the datasets with: (1) the concepts found in each query, (2) the sources of vocabularies, (3) the Medical Subject Headers (MeSH) identifiers, (4) the medical semantic types, and (5) the part-of-speech tagging. In the next section, we describe the 28 features generated in this work.

4.2 Classification Features

We divided the features devised into 5 groups, allowing us to later study the contribution of each group. Groups are described below, and a summary is presented in Table 4.1.

4.2.1 Behavior Features

Behavior features focus on *how* users search instead of *what* they search. We use the average number of words per query and average number of characters per query, because they have shown to be good predictors in the related work [207, 219]. Our goal with these two features is to be able to compare the performance of classifiers using them with classifier using other groups of features.

4.2.2 Semantic Features

As in Section 3.4.2, we explored the MetaMap mappings to acquire the semantic behind a query. Apart from the same three semantic classes: (1) **symptom**, (2) **cause** (disease)

and (3) **remedy**, we use the label **other** to refer to queries with no symptom, cause or remedy. A query can be mapped to multiple semantic classes *symptom*, *cause* and *remedy*, or belong to the *other* semantic class. Features are normalized by the number of queries issued by a users, e.g., if only one out of the five queries a user issue is mapped to the symptoms, then feature $\% \text{ Queries with Symptoms} = 0.2$.

4.2.3 Unified Medical Language System (UMLS) Features

Similarly to what we did in Chapter 3, we use MetaMap to map user queries to UMLS concepts. MetaMap also provides an easy way to access all the sources and different concepts to which a term might belong. Using this information, we model features based on the number of UMLS sources and the number of concepts that can be identified in a query. Additionally, we also map each UMLS concept to one or more concepts in the Medical Subject Headings (MeSH) hierarchy. MeSH was already used assuming that difficult concepts are lower in the hierarchy [213]. For example, the query ‘*wilson’s disease*’ is mapped to two concepts (ATP7B gene, and Hepatolenticular Degeneration), it is presented in 20 out of 169 sources used by Metamap and related to 11 concepts in the MeSH hierarchy. We normalized these counts by the number of queries issued by a user.

4.2.4 Consumer Health Vocabulary Features

The vocabulary gap between consumers and experts is a substantial barrier to health information access for consumers. The Consumer Health Vocabulary (CHV) was created to cope with this issue [218]. The CHV dataset (version 20110204) links part of the UMLS concepts, e.g., ‘*myocardial infarction*’, to everyday expressions, e.g., ‘*heart attack*’, which are called laypeople-preferred term. Moreover, for many terms in the UMLS Metathesaurus, a difficulty score is available, related to the frequency or the context in which the term is used. For any word without a difficult score, we used the mean difficult score of the complete CHV dataset, 0.29 (the data ranges from 0.0 – very difficult – to 1.0 – very easy). For example, *myocardial infarction* score is 0.54, while *heart attack* is 0.80.

For each query in our datasets, we compute five values: (1) the number of terms found in the CHV dataset; if the query contained an (2) expert term (i.e., a UMLS term that is not in the laypeople-preferred version), a (3) consumer term (i.e., a UMLS term that is in the laypeople-preferred version) or a (4) misspelled term (CHV dataset also contains a list of frequently misspelled terms); as well as (5) the average difficult score of all terms identified. We normalized these counts by the number of queries issued by a user. For example, the query ‘*heart attack*’ contains only one concept which is successfully mapped to CHV with MetaMap, a consumer term is identified and its difficulty score is 0.80.

Table 4.1: Groups and features used in this work. Two basic behavior features are used in the baseline models.

Behavior Features (Baseline Only)	
Avg. Words per Query	Avg. characters per query
Semantic Features	
% Queries with Symptoms	% Queries with Causes
% Queries with Remedies	% Queries with Other Types
Unified Medical Language System (UMLS) Features	
% Queries Using Sources	Avg. Sources Per Query
% Queries Using Concepts	Avg. Concepts Per Query
% Queries Using MeSH	Avg. MeSH Per Query
Avg. MeSH Depth Per Query	
Consumer Health Vocabulary (CHV) Features	
Avg. CHV Terms Per Query	% Queries with Consumer Terms
% Query with Expert Terms	% Queries with Misspelled Terms
Avg. Combo Score Per Query	
Part-of-Speech Tagging (POS) Features	
% of Nouns	% of Verbs
% of Auxiliary Verbs	% of Adjectives
% of Conjunctions	% of Adverbs
% of Determiners	% of Prepositions
% of Pronouns	% of Shapes
% of Punctuations	% of Modal Verbs

4.2.5 Part-of-Speech Tagging Features

We employed the part-of-speech tagger module MedPost/SKR¹ of Metamap to annotate each word in a query with one of the following lexicon tags: noun, verb, auxiliary verb, adjective, conjunction, adverb, determiner, preposition, modal verb, pronoun, punctuations and numbers. We count the percentage of queries with any one of the possible tags.

4.3 Classification Results

The classification problem presented here seeks to infer the user expertise based on the user queries by calculating the features shown in Table 4.1 for each user. As described in Section 4.1, there are 25,000 regular users from HON and 25,000 medical users from TRIP in the dataset, resulting in a baseline accuracy of 50.00% for a classifier that assigns all the users to one of the two classes.

Support Vector Machines with Linear Kernel, Logistic Regression and Random Forest from the Python package scikit-learn² were used with their respective best hyperparameters

¹<https://metamap.nlm.nih.gov/MedPostSKRTagger.shtml>

²<http://scikit-learn.org>

Table 4.2: Results of the experiment on classifying users according to their expertise. Different combinations of classifiers and feature sets are explored. Underline is used to show values that are statistically different from the classifier with Behavior Features using a student’s t-test with Bonferroni correction, $p < .00069$.

Feature Set	Classifier	Pos. Class	Acc	Prec	Recall	F1
Baseline I: None	Positive Class	Layp.	50.00	50.00	100.0	50.00
		Exp.				
Baseline II: Behavior Features (Avg. Words per Query & Avg. Chars. per Query)	Logistic Regression	Layp.	63.33	62.53	66.52	64.46
		Exp.		64.25	60.15	62.12
	SVM (Linear Kernel)	Layp.	59.74	60.35	73.22	63.98
		Exp.		61.71	51.24	54.18
	Random Forest	Layp.	68.01	67.10	70.70	68.84
		Exp.		69.05	64.51	66.69
Part Of Speech	Logistic Regression	Layp.	<u>61.62</u>	<u>58.30</u>	<u>81.66</u>	<u>68.02</u>
		Exp.		<u>69.42</u>	<u>41.59</u>	<u>52.00</u>
	SVM (Linear Kernel)	Layp.	61.56	58.32	81.07	67.83
		Exp.		68.98	42.07	52.25
	Random Forest	Layp.	65.33	<u>61.54</u>	81.75	70.22
		Exp.		73.51	47.84	57.96
Semantic	Logistic Regression	Layp.	<u>64.63</u>	<u>65.28</u>	62.51	63.86
		Exp.		64.03	<u>66.74</u>	<u>65.35</u>
	SVM (Linear Kernel)	Layp.	64.61	65.34	62.24	63.75
		Exp.		63.95	66.98	65.43
	Random Forest	Layp.	65.80	67.51	<u>60.92</u>	64.04
		Exp.		64.43	70.35	67.25
UMLS	Logistic Regression	Layp.	<u>65.67</u>	<u>68.84</u>	<u>57.27</u>	<u>62.51</u>
		Exp.		63.42	74.08	<u>68.33</u>
	Linear SVM	Layp.	64.17	68.12	60.00	62.33
		Exp.		64.79	66.20	62.17
	Random Forest	Layp.	70.19	71.26	67.68	69.42
		Exp.		69.91	<u>67.43</u>	<u>68.64</u>
CHV	Logistic Regression	Layp.	<u>67.09</u>	<u>68.35</u>	<u>63.67</u>	<u>65.92</u>
		Exp.		66.00	<u>70.52</u>	<u>68.18</u>
	SVM (Linear Kernel)	Layp.	66.89	68.46	62.67	65.43
		Exp.		65.58	71.12	68.23
	Random Forest	Layp.	70.99	71.62	69.57	70.57
		Exp.		71.30	<u>67.36</u>	<u>69.27</u>
Part of Speech + Semantic + UMLS + CHV	Logistic Regression	Layp.	<u>71.70</u>	<u>72.75</u>	70.24	<u>71.27</u>
		Exp.		<u>71.09</u>	<u>73.16</u>	<u>72.10</u>
	SVM (Linear Kernel)	Layp.	<u>71.58</u>	<u>72.49</u>	70.32	<u>71.15</u>
		Exp.		<u>70.76</u>	74.45	72.37
	Random Forest	Layp.	76.45	79.48	71.33	75.18
		Exp.		76.88	<u>74.43</u>	75.63
All Features Above	Logistic Regression	Layp.	<u>72.71</u>	<u>73.42</u>	<u>71.20</u>	<u>72.28</u>
		Exp.		<u>72.05</u>	<u>74.24</u>	<u>73.12</u>
	SVM (Linear Kernel)	Layp.	<u>69.33</u>	74.06	65.40	67.21
		Exp.		71.59	71.37	68.93
	Random Forest	Layp.	76.93	80.24	71.46	75.59
		Exp.		77.58	75.11	76.32

selected using grid search. The performance of each classifier was measured by precision, recall, F_1 ³ and accuracy scores, as these are well known and widely used metrics.

We performed a ten-fold cross-validation experiment across ten runs. Comparisons to the baseline with behavior features were made using a two tailed student's t-test with Bonferroni correction [69] (72 hypotheses tests = 3 machine learning models \times 6 groups \times 4 evaluation metrics – and an initial $\alpha = .05$ resulted in a new critical p-value of $\alpha/72 = .00069$ – i.e., results are significantly different from the baseline if $p < .00069$). We analyze the results of our model and compare them to two baselines: (1) a classifier that assigns all examples to the positive class, (2) classifiers using two basic user behavior related features - average words per query and average characters per query.

Table 4.2 summarizes the results. Models using the behavior features obtained a substantial improvement over the positive class baseline, with the Random Forest classifier reaching an accuracy score of 68.01, an absolute increase of 18 percentile points over the positive class baseline. The Random Forest model using only the part of speech features or only the semantic feature was not able to improve on the behavior features baseline. The CHV and UMLS features revealed to be strong signals to infer medical expertise. The use of UMLS features for the Random Forest classifier improved the accuracy by 2 percentile points when compared to the same classifier with behavior features; while CHV increased the accuracy by 2.9 percentile points. Finally, the use of all features yield a Random Forest model with an accuracy score of 76.45, 8.4 percentile points higher than the same classifier with behavior features. When adding the behavior features, the accuracy score of the Random Forest model reaches 76.93, a gain of 8.9 percentile points compared to the Random Forest using only the behavior features. In general, the Random Forest classifier obtained the best results within experiments with the same feature set, showing that non-linear models should be preferred to this user classification task.

The Random Forest classifier also allows us to compute the Gini importance score for each feature. This value is higher when the feature is more important, indicating how often a particular feature was selected for a split in a forest, and how large its overall discriminative value was for the classification problem under study. Figure 4.1 shows all features according to the Gini importance score when all the features are used. The reason why the Random Forest using only the two behavior features did relatively well is that these two features are among the top ten strongest features, with Number of Chars per query being the strongest feature. Unfortunately, the high variance of the Gini importance method does not allow us to say that one single feature is significantly better than the others.

A direct comparison with other works in the literature such as White et al. [207], Zhang et al. [219] and Cole et al. [36] would not be fair, because these other works use a different range of features and datasets. Particularly, many of the features are related to the result page: ranking of clicked results, domains of results, saved documents, among others. In

³ F_1 is defined as: $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

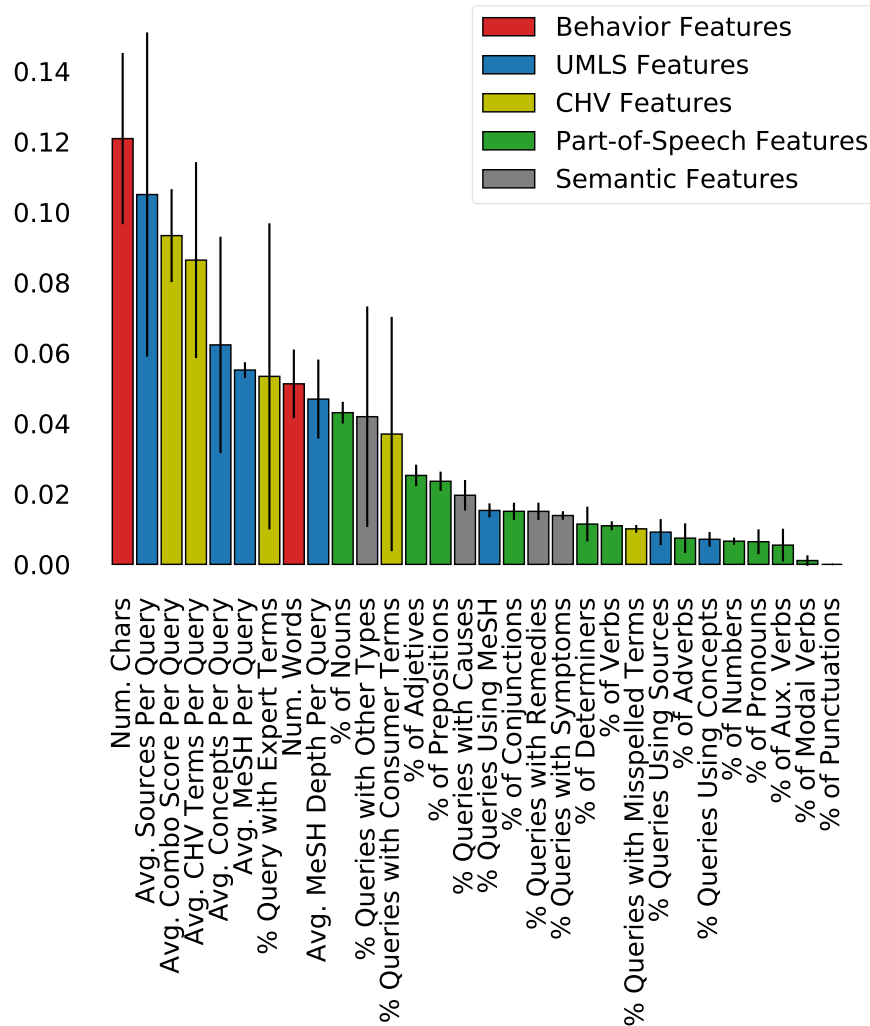


Figure 4.1: Feature importance according to the Gini importance score generated by the Random Forest classifier. The higher, the more important the feature. The error bars represent the standard deviation from the mean value for each feature.

contrast to the metrics used in this work, these are pieces of information more difficult to obtain and not always available in search logs.

4.4 Summary

In this chapter, we made use of two distinct query log dataset to investigate the feasibility of estimating user expertise from user queries. Our approach, when compared to similar work in the literature [207, 209], has the advantage of not requiring any complex filtering of users.

We have developed and evaluated features to be used in the medical domain to classify users according to their expertise. We concentrated on pieces of information easily obtained by any search engine: the keywords for each query. Many of the features devised here have never been used in the literature before.

We evaluated different machine learning models with various sets of features. We found that the best results using only the set of behavior features, simple yet efficient, were superior to using feature sets based on part-of-speech features and semantic analysis. Feature sets extracted from domain vocabularies, such as the UMLS and, in particular using the CHV, reached a better performance than using other feature sets. Altogether the best results were achieved when all features were used, significantly outperforming the two baselines, reaching an accuracy of 76.93%.

Search engines, such as Google, HON or PubMed, could make use of the classifier here proposed to personalize the results for their users. Once a user is classified as a health consumer, the result set could then be re-ranked to promote content that, while topically relevant, has the highest level of understandability. Methods to re-rank results in such way, as well as methods to evaluate the effectiveness of such re-ranking, are presented in Part IV of this thesis.

Part III

Understandability Estimation of Web Documents

The effects of preprocessing HTML on the estimation of understandability

In Part II of this thesis, we studied the users of different medical retrieval systems. We learned that health consumers and health experts have different behavior both on what they search and how they search. We devised and analyzed features based on these differences and showed how it is possible to distinguish between health consumers and health experts, which can be useful to better support search engine users seeking health or medical information online.

Now we turn our attention away from the users and focus on the results, i.e., documents, that are retrieved by search engines. In particular for the health domain, in this part of this thesis, we study methods to infer how difficult to read a document can be.

Researchers in the health domain are actively interested in measuring how capable search engines are in providing consumer-friendly results at the top of their rankings (among the many research papers in this area see [81, 62, 210, 161, 9, 133]). For example, Meillier and Patel [133] show that, in order to understand the top results of Google for *gastroparesis*, users need the knowledge of a 12th-grade student¹. Patel et al. [161] similarly report that results for *thyroid surgery* require at least 10 years in school to be understood, while the average American inhabitant reads at or below an 8th grade level (i.e., 8 years in school) [43, 199, 50, 187], which is the level suggested by NIH for health information on the Web [196].

¹Most of the readability/understandability studies use the K-12 American grade level as reference. K-12 is a shortening to designate the period from the kindergarten (K) for 4-to 6-year-olds to the twelfth grade (12) for 17-to 19-years-olds.

In order to estimate how difficult a Web document is to understand, researchers heavily rely on text readability formulas that are mostly based on surface level characteristics of text, such as the length of words and sentences.

In this chapter, we demonstrate that different tools for extracting text from Web documents lead to very different estimations of understandability. This has an important implication for search engines because search result personalization strategies that consider users reading ability may fail if incorrect text readability estimations are computed.

This chapter is divided as follows. Section 5.1 briefly revisits the traditional readability formulas and their limitations. Section 5.2 details the different preprocessing methods and tools used in our analysis. In Section 5.3, we experiment with a realistic collection to calculate the impact of preprocessing methods on estimation and analysis of readability formulas. Finally, we summarize our findings in Section 5.4.

5.1 Traditional Readability Formulas

Numerous studies have proposed and analyzed methods to accurately measure the level of knowledge required to read a text [55, 82, 38, 39, 192]. While recent research has proposed sophisticated readability estimation methods [39, 82], often tailored to specific domains [214], traditional readability formulas such as the Automated Readability Index and the Gunning Fog Index are extensively used for assessing information on the Web (see for example [210, 224]). As described in Chapter 2, these long-established readability formulas consider the surface level of the text contained in Web documents, that is, the wording and the syntax of sentences. In this framework, the presence of long sentences, words containing many syllables and unpopular words, are all indicators of difficult text to read.

We consider in this chapter a subset of the readability formulas listed in Table 2.1. We explicitly exclude the Dale-Chall Index and the Flesch Reading Ease (FRE) formulas, as their understandability estimation scales widely differ from that of the others, and thus make their comparison more difficult to interpret. An easier way to compare different metrics was also one motivation to update the FRE formula to the FKGL [110]. Note that more recent and sophisticated readability estimators are also not considered in this analysis for two main reasons: (1) they often introduce additional complexity to the estimation process and thus introduce more degrees of variation, which are difficult to control for and compare across; (2) several (if not all) experiments published in the health domain often consider only traditional formulas (see [81, 58, 62, 210, 161, 9, 133]). Nevertheless, further in Chapter 6, more recent understandability estimators are considered.

Two main factors have been identified as affecting the user perception of text difficulty and that thus characterize the readability formulas of Table 2.1:

Word Length: short words are commonly used and understood, while long words are usually rare, often containing many syllables, harder to read, write and remember.

This factor is measured by expressions such as $\frac{Sy}{W}$, $\frac{C}{W}$ and $\frac{PW}{W}$.

Sentence Length: short sentences are usually simple. Long sentences are usually complex, demanding more cognitive processing (memory) and attention. This factor is usually measured by the average words per sentence, i.e., $\frac{W}{S}$.

Each measure differs from the others in the way these factors are combined, usually via a coefficient that has been tuned through comprehensive readability experiments [55].

Web documents contain text strings that do not belong to the information content of documents, but are instead used to format, structure and layout (e.g., tags) and to embed functionalities (e.g., scripts). The presence of these strings affects both word length and sentence length. While the use of HTML parsers allow to remove all strings not associated with the actual informative content of the documents (and thus reducing the errors in estimating word lengths), the estimation of sentence lengths is heavily affected by how the text is extracted from Web documents, as we show with a concrete example in Section 5.2. This is because Web documents are rich in tables, menus, lists, figures, captions, titles and subtitles: these are often part of the information content of the documents, but do not follow the expected structure of a sentence as assumed by the traditional readability formulas. For example, often titles, menus and lists do not end with a punctuation mark that delimits the end of the sentence. In this chapter, we determine the effects that different ways of preprocessing Web documents to extract the text associated with their information content have on the estimation of readability scores.

It is interesting to note that, already in the 1960s, the precise identification of sentence boundaries was a topic of concern for evaluating the readability of text. For example, Smith and Senter [180], authors of the Automated Readability Index (ARI), recommend typists to add to the end of each sentence an equals sign, aligned with a full stop, to explicitly demarcate sentence boundaries.

5.2 Preprocessing of Web Documents

Because traditional readability formulas are based on surface level characteristics of text, the accurate parsing of Web documents is fundamental to ensure that readability is accurately estimated and taken into account for search result personalization. For example, text contained in different HTML fields, tables, lists, etc., should be adequately processed to determine the wording and the syntax of sentences, including sentence length. This preprocessing step is often omitted or simplified (see for example [224]) and the influence of parsing errors on the readability estimation of Web documents is unknown. On the other hand, the cleansing of Web documents' text has been recognized as an important issue in linguistics and language technology research [14].

Here we consider three different approaches to remove the HTML tags and the boilerplate text, to maintain only the text associated with the information content of the Web

<pre> 1 <body class="mediawiki_page-Readability_skin-vector_action-view"> 2 <div id="siteNotice"><!-- CentralNotice --></div> 3 <h1 id="firstHeading" class="firstHeading" lang="en">Readability</h1> 4 <div id="siteSub">From Wikipedia, the free encyclopedia</div> 5 <div id="jump-to-nav" class="mw-jump"> Jump to: 6 navigation, search </div> 7 <div id="mw-content-text" lang="en" dir="ltr" class="mw-content-ltr"> 8 <p>Readability is the ease with which a text can be understood. 9 </div> </pre>	<pre> CentralNotice. Readability. From Wikipedia, the free encyclopedia. Jump to: navigation, search. Readability is the ease with which a text can be understood. </pre>
--	---

Figure 5.1: Simplified Wikipedia entry for Readability (top) and the output of *Naïve* (bottom). In the bottom part, we show the result of the preprocessing approach termed *DoNotForcePeriod* (left), which does not modify the text extracted by the HTML parser, and that of the alternative preprocessing approach termed *ForcePeriod* (right), which adds a period as sentence boundary at the end of every line. The *DoNotForcePeriod* approach concatenates all the text till it reaches a sentence boundary, producing longer sentences than the *ForcePeriod* approach.

documents. The first approach to performing this text cleansing process is to use standard HTML parsing tools such as JSoup² for Java or BeautifulSoup³ for Python. We used BeautifulSoup version 4.3.2, and we term this approach as **Naïve**, as this is the simplest preprocessing method and only naïvely strips the HTML tags from a Web document.

We also consider two open source tools developed specifically for removing the boilerplate from HTML documents: **Boilerpipe** [111] and **JusText** [164]. In common, both of these two tools divide an HTML document into blocks depending on specific HTML tags (*Div*, *Form*, *H1*, so on) and estimate block by block if the text of the block represents boilerplate, thus should be removed, or represents a legit text, and thus should be kept. Boilerpipe [111] is based on decision trees and its basic algorithm uses two features only, the word count and the link density. These features are extracted from the current block (the block being classified) as well as from the previous and the next block. JusText [164] was created based on a set of heuristics such as the observation that short blocks which contain a link are almost always boilerplate. The key idea behind JusText is that long text blocks and some short text blocks can be classified with very high confidence. All the other blocks can then be classified by looking at the surrounding blocks. We used the Python version 1.2.0.0 of Boilerpipe⁴ and version 2.1.1 of JusText⁵.

Figure 5.1 shows the output of the *Naïve* approach applied to the first paragraph of a Web document from Wikipedia. The extracted text often presents an interesting characteristic: it lacks the punctuation marks to delimit the sentence boundaries. This has a clear

²<http://jsoup.org>

³<https://pypi.python.org/pypi/beautifulsoup4>

⁴<https://pypi.python.org/pypi/boilerpipe>

⁵<https://pypi.python.org/pypi/JusText>

effect on the readability measures that consider sentence length as an indication of text difficulty. To better understand the effect of this, we explore two possible approaches:

1. **ForcePeriod:** a sentence boundary (period / full stop) is added at the end of a line if no punctuation mark is found, possibly resulting in short sentences.
2. **DoNotForcePeriod:** no sentence boundary is added, possibly resulting in long sentences.

Note that the *Naïve-DoNotForcePeriod* approach is often used when processing Web documents to automatically estimate readability measures, see for example [224].

5.3 The Influence of Preprocessing Methods on Retrieval Experiments

Researchers in the health domain are interested in measuring how accessible for health consumers are the documents provided by search engines (e.g., [81, 62, 210, 161, 9, 133]). However, these researchers rarely mentioned how, i.e., which preprocessing steps were taken to evaluate the understandability of the retrieved documents. In this section, two different retrieval experiments are done: (1) we directly measure the impact of preprocessing methods into the interpretation of retrieval result sets; (2) we measure how similar would be documents sorted by readability metrics when different preprocessing methods are used. While the first experiment, shown in Section 5.3.1, provides insights for the researchers in the medical domain which assess the understandability of retrieved documents, the second, shown in Section 5.3.2 provides insights to information retrieval researchers.

For the further experiments, we consider the health/medical Information Retrieval task at CLEF eHealth 2015 and 2016 [77, 106] (see also Appendix A). The 2015 collection contains approximately one million Web documents related exclusively to the medical domain and is used as a resource to evaluate search engines tailored to health consumers. The 2016 collection contains medical related topics but used a much larger corpus of Web documents (ClueWeb 12-b, with more than 52 million documents), which is not limited to the health domain. We use these collections because of the importance the readability (and, more generally, the understandability) of Web documents presenting medical advice has within consumer health search [210, 224].

Table 5.1 reports the average number of words and sentences in the whole CLEF eHealth 2015 corpus and in the CLEF eHealth 2016 pool of 25,000 documents as extracted by the preprocessing methods studied here: *Naïve*, *Boilerpipe* and *JusText*. These statistics are at the basis of the readability measures of Table 2.1. We exclusively show in Table 5.1 the statistics of a part (the pooled documents) of the CLEF eHealth 2016 collection. This decision was motivated by the fact that it is more likely that these documents belong to the health domain, as the pool was created by pooling documents retrieved in answer of queries in the health domain by the participating teams of CLEF eHealth 2016 campaign.

Table 5.1: Number of words and sentences (mean and standard deviation) for documents in CLEF 2015 eHealth corpus (approximately 1 million documents) and CLEF 2016 pool of assessed documents (25.000 documents), as obtained by the three preprocessing tools and the two approaches to sentence boundaries (see Section 5.2).

Tool	# Words	# Sentences	
		ForcePeriod	DoNotForcePeriod
CLEF 2015 eHealth Collection			
Naïve	1001.5 ± 2062	137.2 ± 443	37.9 ± 93
Boilerpipe	364.2 ± 884	24.4 ± 55	18.6 ± 49
JusText	409.9 ± 1403	24.4 ± 82	19.4 ± 68
CLEF 2016 eHealth IR Task Pool			
Naïve	2106.4 ± 6103	255.0 ± 887	96.4 ± 434
Boilerpipe	1366.2 ± 5497	80.3 ± 423	73.8 ± 408
JusText	1728.5 ± 4968	93.3 ± 266	90.5 ± 252

Although all the experiments made in this thesis use the whole collection, as we use the same queries created in CLEF eHealth 2016, these documents are the most likely documents to be retrieved in our experiments.

From Table 5.1, we can observe that the *Naïve* method produces a much higher number of words and sentences for both collections than the other two methods. While small, the differences between *Boilerpipe* and *JusText* are still important in that they can influence the estimations of readability measures. Similarly, the use of the *ForcePeriod* approach for sentence boundary rather than the *DoNotForcePeriod* produces large differences among all text preprocessing approaches, in particular for the *Naïve* method.

In order to understand the effect of preprocessing methods on the retrieval of Web documents, we consider baseline runs used in both CLEF eHealth 2015 and 2016. For CLEF 2015, we use the default vector space retrieval model of Apache Lucene 4.8 to retrieve the top 100 documents per query, using the query titles of the 67 topics in the CLEF 2015 collection [77]. For CLEF 2016, we used the default BM25 retrieval model of Terrier 4.0 to retrieval the top 100 documents per query and as queries the 50 topics of the CLEF 2016 collection [106]. For each query, we compute the readability scores of the retrieved documents according to the different settings considered here in terms of preprocessing tools and the approaches to sentence boundaries.

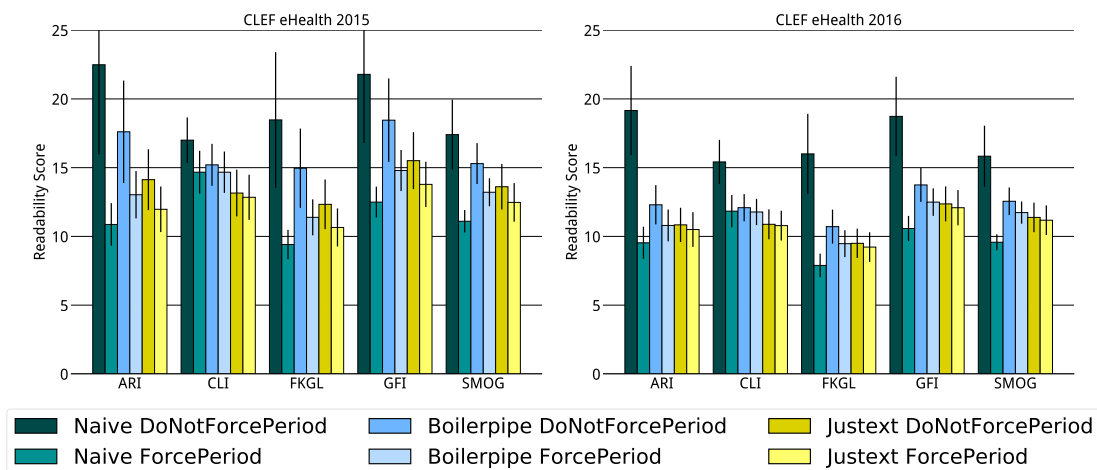


Figure 5.2: Readability scores for each measure based on various preprocessing and sentence boundary methods. Error bars indicate 95% confidence intervals around the mean.

5.3.1 The Influence of Preprocessing Methods on the Interpretation of Readability Formulas

For each retrieved document of queries in both CLEF eHealth 2015 and 2016, we calculate the readability score of various readability formulas (ARI, CLI, FKGL, GFI, SMOG) for each combination of preprocessing method and sentence boundary. Figure 5.2 reports the mean values of readability scores averaged across all retrieved documents. The results suggest that the choice of approach to use for sentence boundary has a significant influence on readability measures: the variance between the readability scores obtained with *ForcePeriod* and *DoNotForcePeriod* is large across all methods, apart for CLI that appears to be the most robust readability formula in this aspect. For example, consider the blue bars for CLEF eHealth 2015 representing the usage of the *Naïve* preprocessing method, the mean readability score of ARI can vary more than 100%, from 10.9 ± 0.4 , when using *ForcePeriod* sentences to 22.5 ± 1.8 , when using *DoNotForcePeriod* sentences. This high variability in the estimation of readability measures influences the conclusions one would infer about the difficulty of the retrieved documents: documents that could be readable by high school students (grade 11) – when sentence boundaries are detected with *ForcePeriod* – become suddenly intractable for people with level of education below that of a Ph.D. student (grade 22) – according to the readability measures computed using the *DoNotForcePeriod* method.

Note in Figure 5.2 that different preprocessing methods (i.e., *Naïve*, *Boilerpipe*, *JusText*) lead to different conclusions about the readability of text. For example, Figure 5.2 suggests that, when ARI is used as readability measure and the *DoNotForcePeriod* approach is employed to identify sentence boundaries, *Naïve* and *JusText* provide contrasting results,

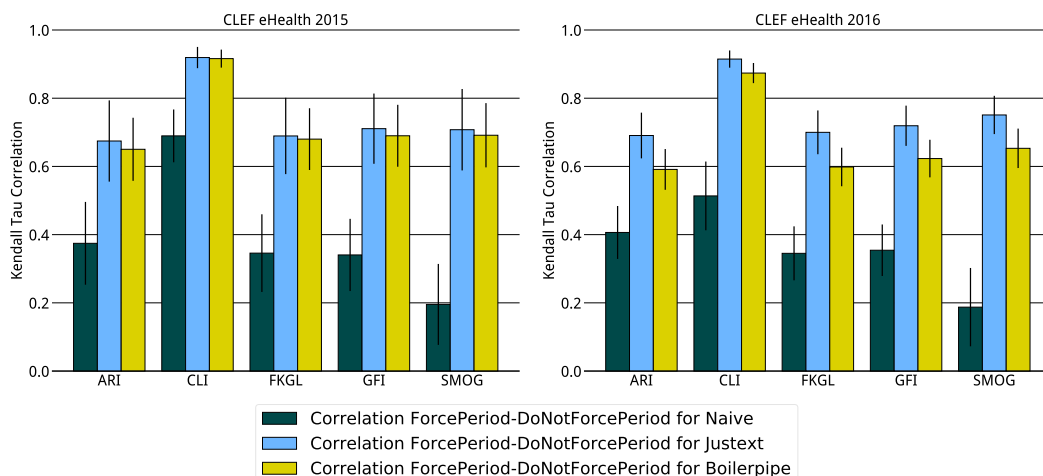


Figure 5.3: Kendall τ correlation and 95% confidence intervals between the *ForcePeriod* and *DoNotForcePeriod* approaches for sentence boundary identification.

with the mean readability of text assessed as being 22.5 ± 1.8 according to *Naïve* and 14.1 ± 0.6 according to *JusText*. These results highlight the significance that choices of the preprocessing tool and sentence boundary identification approach have on the estimation of readability scores for Web documents when using the commonly adopted readability measures considered in this study.

The results in Figure 5.2 also suggest that CLI is the most robust readability measure among those considered in this chapter. In particular, variations in preprocessing tool and sentence boundary identification have little impact on the estimated readability scores for this measure. The stability of CLI is due to the fact that $W \gg S$ and thus $\frac{S}{W} < 1$ (see CLI formula in Table 2.1), dampening the effect of the relation between the number of words and sentences (in our experiments, $1 < 30.0 \times \frac{S}{W} < 4$), and ensuring stability across different values of S . This is unlike measures such as ARI (see ARI formula in Table 2.1), where $3 < 0.5 \times \frac{W}{S} < 13$.

5.3.2 Ranking Correlations based on Readability Scores

Next, we consider how similar document rankings obtained from readability measure estimations are when using different preprocessing and sentence boundary approaches. This is interesting for information retrieval because it is often these differences between rankings, rather than the actual absolute value of the readability estimation, that are used to demote or promote Web documents when taking into account readability levels. For example, a high correlation between two different preprocessing settings would suggest that, although the actual readability scores may be very different, the preference ordering obtained by the readability measures (i.e., the ranking according to readability scores) are similar and therefore these two preprocessing settings would lead to little difference in terms of impact on retrieval.

5.3. The Influence of Preprocessing Methods on Retrieval Experiments

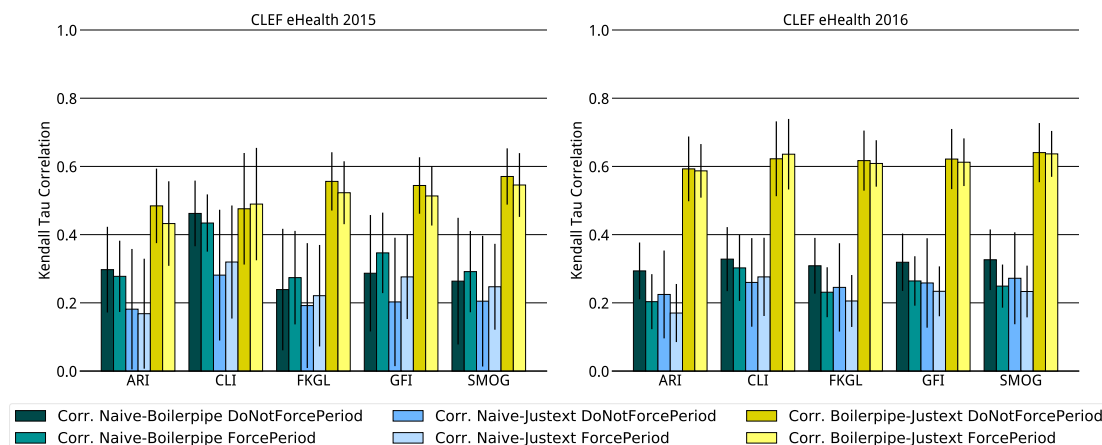


Figure 5.4: Kendall τ correlation and 95% confidence intervals between the approaches for HTML preprocessing, under different settings for sentence boundary identification.

To this aim, in the following experiment, we retrieve the top 100 documents for each query and rank them from the lowest to the highest readability score. We consider the Kendall τ ranking correlation between different settings of sentence boundary identification (Figure 5.3) and preprocessing tool (Figure 5.4). Figure 5.3 shows that, independently of the preprocessing tool used, the correlation between *ForcePeriod* and *DoNotForcePeriod* rankings obtained when using *Boilerpipe* or *JusText* as preprocessing tools is generally high, with the maximum correlation of 0.92 achieved using CLI. However, if the *Naïve* approach to text preprocessing is used, then correlations deteriorate, with the SMOG measure exhibiting a correlation coefficient as low as 0.20. CLI exhibits the least variance in correlation among the three preprocessing approaches (and indeed, the highest correlations) – stability that was already observed when analyzing Figure 5.2.

The results of Figure 5.4 suggest that different preprocessing tools produce different document rankings (when using readability to rank). Specifically, the highest correlation between two of these tools is achieved by the *Boilerpipe-JusText* pair – but these exhibit correlations of *only* about 0.5, independently of the readability measure or the sentence boundary approach (the highest correlation is achieved for SMOG by *Boilerpipe-JusText* with CLEF eHealth 2016 using *ForcePeriod*: $\tau = 0.64 \pm 0.08$). When comparing these methods to the *Naïve* approach, correlation decreases.

These results suggest that, whenever it is possible, advanced HTML cleansing tools, such as *Boilerpipe* and *JusText*, should be preferred over simple tools, such as JSoup or BeautifulSoup, represented here by the *Naïve* method. When the *Naïve* method is used, the influence of the sentence boundary identification is high, which is not seen when *JusText* and *Boilerpipe* are used.

5.4 Summary

This chapter analyzed the influence that preprocessing and sentence boundary identification choices have on the estimation of readability measures for Web documents. The experimental results show that these choices have **a significant impact on the estimation of readability scores**, which in turn can highly influence the order relations among documents that can be obtained from the readability scores. Our findings strongly suggest that attention should be directed to the choice of preprocessing settings when measuring readability for Web documents. Advanced HTML cleansing tools, such as *Boilerpipe* and *JusText*, provide more stable results across settings. In addition, the use of the Coleman-Liau Index (CLI) as readability measure leads to the most stable results across choices of preprocessing tools and sentence boundary identification strategies (although we have not yet assessed the quality of CLI for correctly estimating the readability of documents). In the following chapter, we study which combination of preprocessing settings and readability measure lead to estimations of readability that most agree with user assessment. Note that regardless of each preprocessing method used, the average readability level of results, as shown in Figure 5.2, is much higher than the recommended level of 8th grade [196, 203, 63].

Analysing Documents: understanding understandability through correlation analysis

Chapter 5 showed that preprocessing steps are crucial for the interpretation of readability formula results. Depending on the preprocessing steps used, the same readability formula may yield results that vary widely. Despite this fact, the understandability of documents in the Web is commonly estimated by readability formulas with little or no attention being paid to the preprocessing pipelines used (e.g., [81, 62, 210, 161, 9, 133]).

In this chapter, we introduce the understandability assessments made by humans in recent CLEF eHealth campaigns of 2015 [77] and 2016 [106]. These assessments were complimentary to the topicality assessments and are fundamental to the experiments we conduct in this chapter. With these understandability assessments (described in Section 6.1), we are now able to:

- create and evaluate new methods to estimate understandability of Web documents (methods are also described in Section 6.1). These methods explore a variety of approaches to estimate how understandable a Web page can be. These will posteriorly be used as features for machine learning (Section 6.5) and learning-to-rank models (Chapter 8).
- particularly evaluate the use of readability formulas to estimate understandability (Section 6.2);
- systematically evaluate the combination of different HTML preprocessing pipelines (described in Chapter 5) and methods to estimate the understandability of Web pages (Section 6.3);

- identify the methods that best correlate with human assessments to estimate understandability (Section 6.4);
- create and evaluate automatic machine learning methods to predict the understandability of Web documents (Section 6.5).

A summary of the main findings of this chapter is presented in Section 6.6.

6.1 Experimental Methodology

The main components of our evaluation are presented in this section in order: the dataset and ground truth data (Section 6.1.1), the methods of understandability estimation which can be used independently or act as features for machine learning algorithms (Section 6.1.2), the preprocessing of Web documents (Section 6.1.3), and the evaluation metrics (Section 6.1.4).

6.1.1 CLEF eHealth Data Collections

We make use in this chapter of the CLEF 2015 and 2016 eHealth collections [77, 106], as these collections have assessments for both topical relevance¹ and understandability. While a detailed description of CLEF 2015 and 2016 collections can be found in Appendix A, we briefly introduce these datasets here.

The CLEF 2015 collection contains 67 queries and 2,515 documents that have been assessed relevant by clinical experts and have an assessment for understandability [77]. Documents in this collection are a selected crawl of health Web sites, of which the majority are certified HON Web sites. The CLEF 2016 collection contains 300 queries and 3,298 relevant documents that also have been assessed with respect to understandability [106]. Documents in this collection belong to the ClueWeb12 B13 corpus, and thus are general English Web documents, not necessarily targeted to health topics, nor of a controlled quality (as are instead HON certified documents). Understandability assessments were provided on a 4-point Likert scale for CLEF 2015 (0: “*Content is very technical and difficult to read*” to 3: “*Content is very easy to read and understand*”), and on a [0, 100] range for CLEF 2016 (0 indicates the highest understandability). To illustrate how the assessment was conducted, we show in Figure A.3 the adaptations made on *Relevation!* [112], the assessment tool used in both CLEF campaigns.

To support the investigation on methods to automatically estimate the understandability of Web documents, we also considered correlations among multiple human assessors (inter-assessor agreement). For CLEF 2015, we used the publicly available additional assessments made by unpaid medical students and health consumers collected in our previous study of how medical expertise affects assessments [153]. For CLEF 2016,

¹We refer to this simply as relevance in the remainder of this thesis, when this does not cause confusion.

we collected understandability assessments for 100 documents. Three members of our research team were recruited to provide the assessments. Once again, *Relevation!* [112] was also used to assist with the assessments similarly to the original settings used in CLEF.

6.1.2 Understandability Estimators

As reviewed in Section 2.3, several approaches have been applied to estimate the understandability of health documents on the Web, with the most popular methods (at least in the biomedical literature) being readability formulas based on surface level characteristics of the text. In this section, we outline a large number of methods to estimate document understandability which we categorize into ten distinct groups described as follows. An overview of all methods is shown in Table 6.1.

Traditional Readability Formulas (TRF)

This group includes all the readability formulas introduced in Chapter 2 (Table 2.1). We implemented the Python package *ReadabilityCalculator*² with all the readability formulas used in this thesis.

Components of Readability Formulas (CRF)

This group is formed by the “building blocks” used in the traditional readability formulas. Examples of such building blocks include the average number of characters per word and the average number of syllables in a sentence. Words are divided into syllables using the Python package *Pyphen*³.

General Medical Vocabulary Features (GMV)

This group includes methods that count the number of words with a medical prefix or suffix, i.e., beginning or ending with Latin or Greek particles (e.g., amni-, angi-, algia-, arteri-), and text strings included in lists of acronyms or in medical vocabularies such as the International Statistical Classification of Diseases and Related Health Problems (ICD), Drugbank and the OpenMedSpel dictionary⁴. An acronym list from the ADAM database [222] was used. Methods in this group were matched with documents using simple keyword matching.

Consumer Vocabulary Features (CVF)

Previously used in Chapter 4, the Consumer Health Vocabulary (CHV) is a prominent medical vocabulary dedicated to mapping consumer (layperson) vocabulary to technical terms [218]. It attributes a score for each of its concepts with respect to their difficulty,

²<https://pypi.python.org/pypi/ReadabilityCalculator>

³<http://www.pyphen.org/>

⁴<http://extensions.openoffice.org/en/project/openmedspel-en-us>

6. ANALYSING DOCUMENTS: UNDERSTANDING UNDERSTANDABILITY THROUGH CORRELATION ANALYSIS

Table 6.1: Methods to estimate the understandability of Web documents. \star : raw values are used; \diamond : values normalized by the number of words in a document are used; \dagger : values normalized by the number of sentences in a document are used.

Group	Metric	Group	Metric
Traditional Readability Formulas (TRF)	Automated Readability Index (ARI) [180]	HTML Features (HF)	# of Abbr tags
	Coleman-Liau Index (CLI) [37]		# of A tags
	Dale Chall Index (DCI) [49]		# of Blockquote tags
	Flesch-Kincaid Grade Level (FKGL) [110]		# of Bold tags
	Flesch Reading Ease (FRE) [110]		# of Cite tags
	Gunning Fog Index (GFI) [83]		# of Div tags
Lasbarhetsindex (LIX) [19]	# of Forms tags		
Simple Measure of Gobbledygook (SMOG) [129]	# of H1 tags		
Components of Readability Formulas (CRF)	# of Characters $\star\dagger$		# of H2 tags
	# of Words $\star\dagger$		# of H3 tags
	# of Sentences $\star\dagger$		# of H4 tags
	# of Difficult Words (Dale Chall list [49]) $\star\dagger$		# of H5 tags
	# of Words Longer than 4 chars $\star\dagger$		# of H6 tags
	# of Words Longer than 6 chars $\star\dagger$		# of Hs (any H above)
	# of Words Longer than 10 chars $\star\dagger$		# of Img tags
	# of Words Longer than 13 chars $\star\dagger$		# of Input tags
	# of Number of Syllables $\star\dagger$		# of Link tags
General Medical Vocabularies Features (GMV)	# of Polysyllable Words (>3 Syllables) $\star\dagger$		# of DL tags
	# of Words with Medical Prefix $\star\dagger$		# of UL tags
	# of Words with Medical Suffix $\star\dagger$		# of OL tags
	# of Acronyms $\star\dagger$		# of List (DL + UL + OL)
	# of ICD Concepts $\star\dagger$	# of Q tags	
Consumer Medical Vocabulary Features (CMV) [218]	# of Drugbank $\star\dagger$	# of Scripts tags	
	# of Words in medical dict. (OpenMedSpel) $\star\dagger$	# of Spans tags	
	CHV Mean Score for all Concepts $\star\dagger$	# of Table tags	
	# of CHV Concepts $\star\dagger$	# of P tags	
	CHV Mean Score for Symptom Concepts $\star\dagger$	Word Frequency Features (WFF)	
# of CHV Symptom Concepts $\star\dagger$	25th percentile English Wikipedia		
CHV Mean Score for Disease Concepts $\star\dagger$	50th percentile English Wikipedia		
# of CHV Disease Concepts $\star\dagger$	75th percentile English Wikipedia		
Expert Medical Vocabulary Features (EMV)	# of MeSH Concepts $\star\dagger$		Mean Freq. Percentile English Wikipedia
	Average Tree of MeSH Concepts $\star\dagger$		Mean Freq. Percentile English Wikipedia - Includes OV
	# of MeSH Symptom Concepts $\star\dagger$		25th percentile Medical Reddit
	Average Tree of MeSH Symptom Concepts $\star\dagger$		50th percentile Medical Reddit
	# of MeSH Disease Concepts $\star\dagger$		75th percentile Medical Reddit
Natural Language Features (NLF)	Average Tree of MeSH Disease Concepts $\star\dagger$		Mean Freq. Percentile Medical Reddit
	Positive Words $\star\dagger$		Mean Freq. Percentile Medical Reddit include OV
	Negative Words $\star\dagger$		25th percentile Pubmed
	Neutral Words $\star\dagger$	50th percentile Pubmed	
	# of verbs $\star\dagger$	75th percentile Pubmed	
	# of nouns $\star\dagger$	Mean Freq. Percentile Pubmed	
	# of pronouns $\star\dagger$	Mean Freq. Percentile Pubmed - Includes OV	
	# of adjectives $\star\dagger$	25th p. Wikipedia+Reddit+Pubmed	
	# of adverbs $\star\dagger$	50th p. Wikipedia+Reddit+Pubmed	
	# of adpositions $\star\dagger$	75th p. Wikipedia+Reddit+Pubmed	
	# of conjunctions $\star\dagger$	Mean R. Wiki.+Reddit+Pubmed	
	# of determiners $\star\dagger$	Mean R. Wiki.+Reddit+Pubmed - w. OV	
	# of cardinal numbers $\star\dagger$	Regressor (MLR)	
	# of particles or other function words $\star\dagger$		Linear Regressor
	# of other POS (foreign words, typos) $\star\dagger$		eXtreme Gradient Boosting (XGB) Regressor
	# of punctuation $\star\dagger$		Multi-layer Perceptron Regressor
	Height of part-of-speech parser tree $\star\dagger$	Random Forest Regressor	
# of Entities $\star\dagger$	Support Vector Machine Regressor		
# of Stopwords $\star\dagger$	Classifier (MLC)		
# of words not found in Aspell Eng. dict. $\star\dagger$		Logistic Regression	
		eXtreme Gradient Boosting (XGB) Classifier	
		Multi-layer Perceptron Classifier	
		Random Forest Classifier	
	Support Vector Machine Classifier		
	Multinomial Naive Bayes		

with lower/higher scores for harder/easier concepts. The MetaMap [6] tool was used to map the text content of Web documents to entries in CHV. We also explore MetaMap options to filter only concepts identified as symptoms or diseases, using the same definitions from Section 3.4.2.

Expert Medical Vocabulary Features (EMV)

The hierarchy of Medical Subject Headers (MeSH) was previously used in the literature to identify difficult concepts, assuming that a concept that is deep in the hierarchy is more difficult than a shallow one [213]. We use MeSH as an example of expert vocabulary and, as done with CHV, we used MetaMap to map the content of Web documents to MeSH entities. We also study mappings to symptoms and disease concepts separately.

Natural Language Features (NLF)

This group includes commonly used natural language heuristics such as the ratio of part-of-speech (POS) classes, the height of the POS parser tree, the number of entities in the text, the sentiment polarity [158] and the ratio of words found in English vocabularies. The Python package NLTK 3.2⁵ was employed for sentiment analysis and POS tagging. The GNU Aspell⁶ dictionary was used as a standard English vocabulary and a stopword list was built by merging the stopword lists of the Indri⁷ and Terrier⁸ toolkits. Discourse features, such as the distribution of POS classes and proportion of entity in a document, were previously studied in the task of understandability estimation [60] and found superior to complex features such as entity co-reference and entity grid [15]. Our intuition when using sentiment polarity is that the content produced by laypeople in patient forums or blogs (easy-to-read) might be potentially more emotional than scientific publications (hard-to-read).

HTML Features (HF)

This group includes the identification of a large number of HTML tags, which were extracted with the Python library BeautifulSoup v4.4⁹. The intuition for these features is that Web documents with many images and tables might explain and summarize health content better, thus providing more understandable content to the general public.

Word Frequency Features (WFF)

Generally speaking, common and known words are usually frequent words, whereas unknown and obscure words are generally rare. This idea is implemented in readability formulas such as the Dale-Chall Index, which uses a list of common words and counts the number of words that fall outside this list (complex words) [49] (see Section 2.3.1). We extended these observations by studying corpus-wide word frequencies. Three very distinct auxiliary corpora, but all in the health domain, were analyzed to extract word frequencies:

⁵<http://www.nltk.org/>

⁶<http://www.aspell.net/>

⁷<http://www.lemurproject.org/indri/>

⁸<http://www.terrier.org/>

⁹<https://www.crummy.com/software/BeautifulSoup/>

6. ANALYSING DOCUMENTS: UNDERSTANDING UNDERSTANDABILITY THROUGH CORRELATION ANALYSIS

- Medical Reddit: Reddit¹⁰ is a Web forum with a sizeable user community which is responsible for generating and moderating its content. Any user can start a discussion or reply to a discussion. This forum is intensively used for health purposes: for example, in the Reddit community AskDocs¹¹, licensed nurses and doctors (subject to user identity verification) advise help seekers free of charge. We selected six of such communities (*medical*, *AskDocs*, *AskDoctorSmeeee*, *Health*, *WomensHealth*, *Mens_Health*) and downloaded all user interactions available until September 1st, 2017 using the Python library PRAW¹²), v5.1. In total 43,019 discussions were collected.
- Medical English Wikipedia: after obtaining a recent Wikipedia dump¹³ (May 1st 2017), we filtered articles to only those containing an Infobox¹⁴ in which at least one of the following words appeared as a property: *ICD10*, *ICD9*, *DiseasesDB*, *MeSH*, *MeSHID*, *MeshName*, *MeshNumber*, *GeneReviewsName*, *Orphanet*, *eMedicine*, *MedlinePlus*, *drug_name*, *Drugs.com*, *DailyMedID*, *LOINC*. Figure 6.1 illustrates a Wikipedia document that is marked as medical because of its Infobox entries. In doing so, we followed the method by Soldaini et al. [181], which favors precision over recall when identifying a health-related article. This resulted in a collection of 11,868 articles.
- PubMed Central: PubMed Central (PMC)¹⁵ is an online digital database of full-text biomedical literature. We used the collection distributed for the TREC 2014¹⁶ and 2015¹⁷ Clinical Decision Support Track [167, 168], consisting of 733,191 articles.

A summary of the corpora statistics is reported in Table 6.2. Instead of working with the absolute frequency of a term in a corpus, we use its frequency percentile. Let L be a list of occurring terms sorted by their frequency in a corpus, the frequency percentile of a term t is the position of t in L divided by the length of L , multiplied by 100. Intuitively, this is a linearization of the typically exponential distribution of frequencies of terms in a corpus. This way, for example, the consumer term ‘*heart*’ is at percentile 99.75 in the Reddit corpus, while the expert term ‘*myocardium*’ is only found at percentile 66.03 in the same corpus.

Using the frequency percentiles, we defined different features for a document. For example, the *Mean Frequency Percentile English Wikipedia* of a document is the sum of the frequency percentile of each word of the document found in the English Wikipedia corpus divided by the total number of words in the document. The *50th percentile*¹⁸

¹⁰<http://www.reddit.com/>

¹¹<http://www.reddit.com/r/AskDocs/>

¹²<https://praw.readthedocs.io/>

¹³<https://dumps.wikimedia.org/enwiki/>

¹⁴A Wikipedia infobox is a structured template that appears on the right of Wikipedia documents summarizing key aspects of articles.

¹⁵<https://www.ncbi.nlm.nih.gov/pubmed/>

¹⁶<http://trec-cds.appspot.com/2014.html>

¹⁷<http://trec-cds.appspot.com/2015.html>

¹⁸Not to be confused with the frequency percentile.

Table 6.2: General statistics for the auxiliary corpora used in this work

Statistic	Medical Wikipedia	Medical Reddit	PubMed Central
Number of Docs.	11,868	43,019	733,191
Number of Words	10,655,572	11,978,447	144,024,976
Number of Unique Words	467,650	317,106	2,933,167
Avg. Words per Doc.	898.90 \pm 1351.76	278.45 \pm 359.70	227.22 \pm 270.44
Avg. Char per Doc.	5107.81 \pm 7618.57	1258.44 \pm 1659.96	1309.11 \pm 1447.31
Avg. Char per Word	5.68 \pm 3.75	4.52 \pm 3.52	5.76 \pm 3.51

Article [Talk](#) [Read](#) [Edit](#) [View history](#)

Hyperthermia

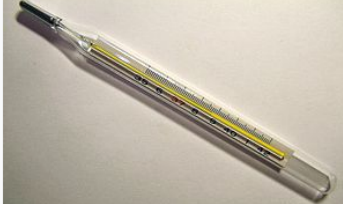
From Wikipedia, the free encyclopedia

Hyperthermia is elevated body temperature due to failed [thermoregulation](#) that occurs when a body produces or absorbs more [heat](#) than it dissipates. Extreme temperature elevation then becomes a [medical emergency](#) requiring immediate treatment to prevent disability or death.

The most common causes include [heat stroke](#) and adverse reactions to drugs. The former is an [acute temperature elevation](#) caused by exposure to excessive heat, or combination of heat and humidity, that overwhelms the heat-regulating mechanisms. The latter is a relatively rare side effect of many drugs, particularly those that affect the [central nervous system](#). [Malignant hyperthermia](#) is a rare complication of some types of [general anesthesia](#).

Hyperthermia differs from [fever](#) in that the body's temperature set [point](#) remains unchanged. The opposite is [hypothermia](#), which occurs when the temperature drops below that required to maintain normal metabolism. The term is from [Greek](#) ὑπέρ, *hyper*, meaning "above" or "over", and θερμός, *thermos*, meaning "hot".

Hyperthermia



An analog [medical thermometer](#) showing a temperature of 38.7 °C (101.7 °F)

Classification and external resources

Specialty	Critical care medicine
ICD-9-CM	780.6 ↗
DiseasesDB	18924 ↗
MeSH	D005334 ↗

[\[edit on Wikidata\]](#)

Figure 6.1: Wikipedia document on hyperthermia. The rectangular red box identifies the Infobox on the right-hand side. This document is filtered as medical because it contains entries for *ICD-9*, *DiseasesDB* and *MeSH*.

English Wikipedia of a document is the median frequency percentile among the frequency percentile of all the words in the document (including duplicates).

Machine Learning on Text: Regressors (MLR) and Classifiers (MLC)

This group includes machine learning methods for estimating Web document understandability based on word counts. While it has been noted in the literature that machine learning methods are promising for estimating understandability, an open challenge is identifying the background corpus to be used for training [38]. In this work, we propose the use of the three corpora detailed above as background corpus, and assumed understandability labels according to the expected difficulty of documents in these

collections:

- Medical Reddit (**label 1**): Documents in this collection are expected to be written in a colloquial style, and thus the easiest to understand. All the conversations are in fact explicitly directed to assist inexperienced health consumers;
- Medical English Wikipedia (**label 2**): Documents in this collection are expected to be less formal than scientific articles, but more formal than a Web forum like Reddit, thus somewhat more difficult to understand;
- PubMed Central (**label 3**): Documents from this collection are expected to be written in a highly formal style, as the target audiences are physicians and biomedical researchers.

Based on the labels of each class above, models were learned using all documents from these corpora after features were extracted using Latent Semantic Analysis (LSA) with ten dimensions on top of TF-IDF calculated for each word. We modeled a classification task as well as a regression task using these corpora. In the classification task, the first step is to train a classifier on documents belonging to these three collections with the three different classes shown above. The second step is to use the classifier to estimate which of these three possible classes an unseen document from the CLEF 2015 or CLEF 2016 would belong. Similarly, in the regression task, after training, a regressor has to estimate an “understandability” value to an unseen CLEF document. We hypothesize that documents that are more difficult to read are more similar to PubMed documents than to Wikipedia or Reddit ones.

6.1.3 Preprocessing Pipelines for Web Documents

Except for the methods in the HTML features group (HF in Table 6.1), the methods in all other nine groups, in order to estimate an understandability score, require a preprocessing pipeline to remove the HTML markups of a Web document. It is expected the preprocessing pipelines are more important to some groups, such as the group of readability formulas, than others. In this chapter, we systematically evaluate the preprocessing pipelines introduced in Chapter 5 comparing preprocessing pipelines and methods to estimate understandability within and across groups.

To preprocess the documents in the CLEF 2015 and 2016 collections, extracting the main content of Web documents from the HTML source, we use the same preprocessing pipeline from Chapter 5: BeautifulSoup [166] (*Naïve*), which just naively removes HTML tags, Boilerpipe [111] (*Boi*) and JusText [164] (*Jst*), which eliminate boilerplate text together with HTML tags. Also, we experiment with the same two preprocessing heuristics devised in Chapter 5: *ForcePeriod (FP)*, which forces a period at the end of each extracted HTML field, and *DoNotForcePeriod (DNFP)* which does not.

6.1.4 Evaluation Metrics

As part of our study, we investigate the influence that the preprocessing of Web documents has on the estimation of understandability. We do so by comparing the combination of a number of preprocessing pipelines and heuristics (Section 6.1.3), and understandability estimation methods (Section 6.1.2) with human assessments of Web document understandability from the CLEF eHealth collections (Section 6.1.1). For that, we use Pearson, Kendall and Spearman correlations to compare the understandability assessments made by human assessors in the CLEF collections with estimations obtained by the considered approaches, under all combinations of preprocessing pipelines and heuristics. Pearson correlation is used to calculate the strength of the linear relationship between two variables, whereas Kendall and Spearman measure the rank correlations among the variables. We opted to report all three correlation coefficients to allow for a thorough comparison with other work, as they are equally used in the literature.

Whenever appropriate, we use either the two-tailed paired t-test [118] or the Analysis Of Variance (ANOVA) test [61] to compare whether the differences found in our experiments are statistically significant. Tests are performed at the 5% level of significance, i.e., differences are significant if $p < .05$.

6.2 Evaluation of Readability Formulas

We begin our evaluation by focusing our analysis on the readability formulas alone due to their importance. We assume that readability score estimated by the readability formulas can be directly used as proxies for document understandability, i.e., they can measure how understandable a document is.

In Chapter 5, we studied the stability of each readability formula in the face of different preprocessing strategies. We identified that the use of *Naïve* preprocessing was associated with more larger variances in the (understandability) score estimated by readability formulas, as shown in Figure 5.2. Now we investigate how strong is the correlation between the scores of readability formulas preprocessed with different methods with the human ground truth. For that, the correlation scores of each traditional readability formula with the human assessments made in CLEF 2015 and 2016 are shown in Figures 6.2 and 6.3, respectively.

Once more, the *Naïve* preprocessing is the worst choice. No matter which correlation measure or which readability formula is used, the *Naïve* preprocessing shows the lowest correlation coefficients. To summarize the analysis of Figures 6.2 and 6.3, we show in Table 6.3 the average correlation for each preprocessing pipeline across the results of the various readability formulas. As shown in Table 6.3, *JusText* has the highest average correlation coefficient in both CLEF eHealth 2015 and 2016 for all correlation measure studied. Experiments with CLEF 2015 show that correlations with human assessments are significantly higher when using *JusText* compared to the other methods, but from CLEF 2016, *JusText* is not significantly different from *Boilerpipe* when Spearman or

6. ANALYSING DOCUMENTS: UNDERSTANDING UNDERSTANDABILITY THROUGH CORRELATION ANALYSIS

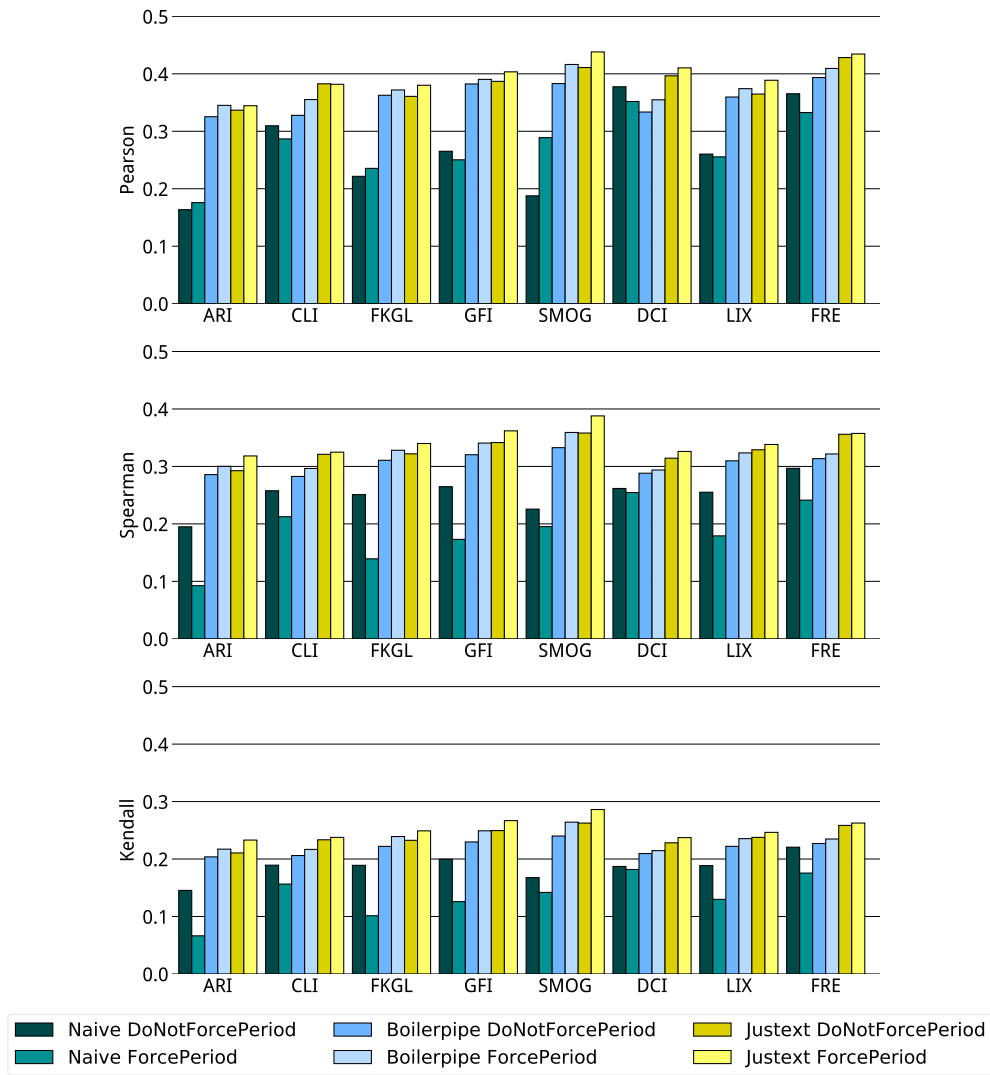


Figure 6.2: Correlation between the scores of various readability formulas and the understandability scores assessed by humans collected in CLEF eHealth 2015 considering different HTML preprocessing pipelines

Kendall are used. Nevertheless, correlations are significantly lower when using the *Naive* pipeline in all experiments.

To study the impact of using *ForcePeriod* or *DoNotForcePeriod*, similarly to Table 6.3, in Table 6.4, we measured the mean absolute correlation between human assessments and the score predicted by the readability formulas across readability formulas and preprocessing pipelines. No differences were found when comparing the use of *ForcePeriod* and *DoNotForcePeriod*.

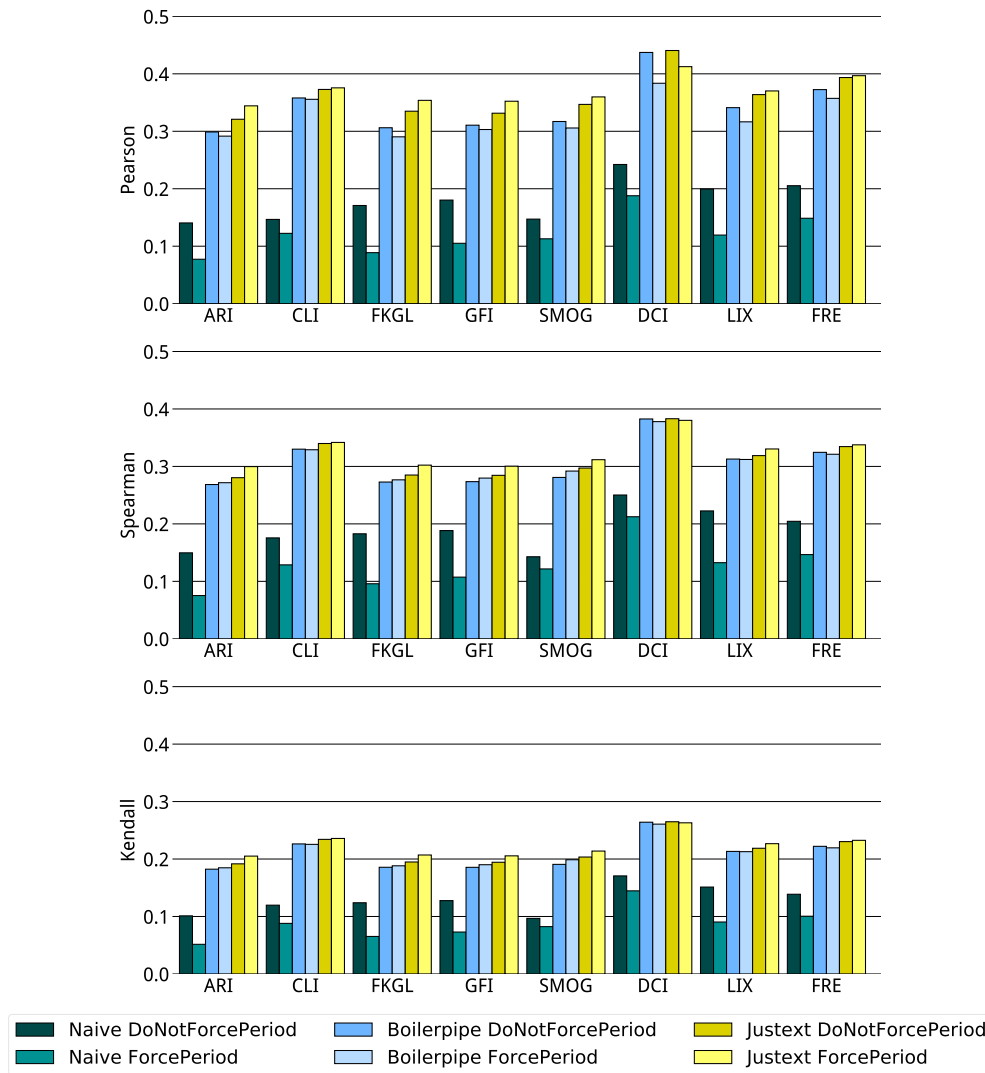


Figure 6.3: Correlation between the scores of various readability formulas and the understandability scores assessed by humans collected in CLEF eHealth 2016 considering different HTML preprocessing pipelines

Finally, in Table 6.5, we show the average absolute correlation for each readability formula across the different preprocessing pipelines, heuristics and correlation measures. Results show that the Flesch Reading Ease formula had the highest average correlation in CLEF 2015 and the second highest in CLEF 2016, although it was statistically different only from the ARI Index in CLEF 2015. The Dale-Chall Index had the highest average correlation in CLEF 2016, significantly higher than four other readability formulas (ARI Index, Flesch Kincaid Grade Level, Gunning Fox and Smog Index), and third highest in CLEF 2015 (only .007 less than the runner-up formula, Smog Index). The ARI Index

6. ANALYSING DOCUMENTS: UNDERSTANDING UNDERSTANDABILITY THROUGH CORRELATION ANALYSIS

Table 6.3: Correlation results (Mean \pm Std) each correlation measure across different preprocessing pipelines for readability formula in Figures 6.2 (CLEF eHealth 2015) and 6.3 (CLEF eHealth 2016). The symbols B , J and N are used to show significant differences between the current value and the one obtained by using, respectively, *Boilerpipe*, *JusText* or the *Naïve* preprocessing pipeline.

Pipeline	CLEF eHealth 2015				CLEF eHealth 2016			
	Pearson	Spearman	Kendall	Average	Pearson	Spearman	Kendall	Average
Naïve	.27 \pm .06 ^{BJ}	.22 \pm .05 ^{BJ}	.16 \pm .04 ^{BJ}	.20 \pm .06 ^{BJ}	.15 \pm .04 ^{BJ}	.16 \pm .05 ^{BJ}	.11 \pm .03 ^{BJ}	.16 \pm .04 ^{BJ}
Boilerpipe	.37 \pm .03 ^{NJ}	.31 \pm .02 ^{NJ}	.23 \pm .02 ^{NJ}	.28 \pm .04 ^{NJ}	.33 \pm .04 ^{NJ}	.31 \pm .04 ^N	.21 \pm .03 ^N	.28 \pm .06 ^N
JusText	.39 \pm .03 ^{NB}	.34 \pm .02 ^{NB}	.25 \pm .02 ^{NB}	.31 \pm .05 ^{NB}	.37 \pm .03 ^{NB}	.32 \pm .03 ^N	.22 \pm .02 ^N	.30 \pm .07 ^N

Table 6.4: Mean absolute difference between the average correlation coefficient when using *ForcePeriod* and *DoNotForcePeriod* for each correlation measure across different preprocessing pipelines for each readability formula in Figures 6.2 (CLEF eHealth 2015) and 6.3 (CLEF eHealth 2016). No significant differences were found when comparing *ForcePeriod* to *DoNotForcePeriod*.

Pipeline	CLEF eHealth 2015				CLEF eHealth 2016			
	Pearson	Spearman	Kendall	Average	Pearson	Spearman	Kendall	Average
ForcePeriod	.34 \pm .07	.30 \pm .04	.22 \pm .03	.28 \pm .07	.29 \pm .09	.30 \pm .04	.18 \pm .05	.26 \pm .09
DoNotForcePeriod	.35 \pm .07	.28 \pm .08	.21 \pm .06	.28 \pm .09	.27 \pm .11	.28 \pm .08	.17 \pm .07	.24 \pm .11

had the lowest average correlation for both datasets and the only readability formula significantly worse than the best formulas of CLEF 2015 and 2016.

Results of Chapter 5 indicated that Coleman Liau Index (CLI) was the most stable metric across different preprocessing pipelines. However, our experiments in this section show that other formulas correlated better with human assessments than CLI, although the differences between CLI and the best formulas are not significant. While we do not explicitly advocate in favor of any readability formula in particular, we advocate against the use of ARI index, which was the least stable readability formula among the ones evaluated in Chapter 5 (i.e., changing the preprocessing method highly impacts the estimated understandability score by the ARI index, see Figure 5.4) and obtained the lowest average absolute correlation among all metrics, as shown in Table 6.5.

It is important to note the large difference between the correlation coefficients reported in Chapter 5 and the ones reported in this section. In Figures 5.3 and 5.4, we calculated the correlation coefficient among the different preprocessing strategies, but never against human assessments. The correlation coefficients for the comparisons shown in Figures 5.3 and 5.4 were as high as 0.90. Instead, in Figures 6.2 and 6.3, we compared preprocessing strategies against human assessments, obtaining correlation coefficients that were never higher than 0.45. Assessing the difficulty of a text is a hard and subjective task, and human assessments for this task are naturally noisy. Although the correlation coefficients shown in this section seem low when compared to those obtained in Chapter 5, they

Table 6.5: Correlation results (Absolute Mean \pm Std) for each readability formula across different preprocessing pipelines and heuristics. These results summarize those of Figures 6.2 (CLEF eHealth 2015) and 6.3 (CLEF eHealth 2016). Highest results for each collection are shown in bold.

Readability Formula	CLEF eHealth 2015	CLEF eHealth 2016
ARI Index	.236 \pm .085	.207 \pm .091
Coleman Liau Index	.277 \pm .065	.250 \pm .099
Flesch Kincaid Grade Level	.270 \pm .079	.218 \pm .086
Gunning Fog Index	.289 \pm .079	.222 \pm .083
SMOG Index	.297 \pm .090	.223 \pm .089
Dale-Chall Index	.290 \pm .070	.303 \pm .093
LIX Index	.278 \pm .072	.247 \pm .086
Flesch Reading Ease	.313 \pm .076	.260 \pm .093

are actually not far from the correlation coefficient measured by different sets of human assessors as we will see in Section 6.3 (e.g., the highest Kendall- τ 's correlation between two sets of human assessors for CLEF eHealth 2015 is only 0.35 as shown in Figure 6.4).

6.3 Evaluation of Preprocessing Pipelines and Heuristics

We further compare the correlation of other methods for understandability estimation introduced in Section 6.1.2 (summarized in Table 6.1) with human assessments made in CLEF eHealth 2015 and 2016. The results for each group of methods are aggregated and shown with boxplots in Figures 6.4 (CLEF 2015) and 6.5 (CLEF 2016). Note that we kept the results separated with respect to the preprocessing pipelines used. For instance, the first boxplot on the top of Figure 6.4 represents the distribution of Pearson correlations with human assessments across all methods in the category Traditional Readability Features (Table 6.1), obtained with the *Naïve ForcePeriod* preprocessing, for CLEF 2015. Each box extends from the lower to the upper quartile values, with the red marker representing the median value for that category. Whiskers show the range of the data in each category and circles represent values considered outliers for the category. The last four boxes are the summary results across all understandability assessment methods and sentence-ending heuristics for each of the preprocessing pipelines (named *Preprocessing Accumulator*), and the inter-assessor correlation (last box) when multiple assessors provided assessments about the understandability of Web documents (details about this data in Section 6.1.1). This indicates the range of variability and subjectiveness when assessing understandability, along with the highest correlation we measured between human assessors.

The correlations between human assessments and readability formulas are once again shown in both Figures 6.4 and 6.5 to provide a comparison with other methods to estimate understandability. The choice of preprocessing method strongly impacts readability formulas, but they are not the only ones: in general, differences in correlation strength

6. ANALYSING DOCUMENTS: UNDERSTANDING UNDERSTANDABILITY THROUGH CORRELATION ANALYSIS

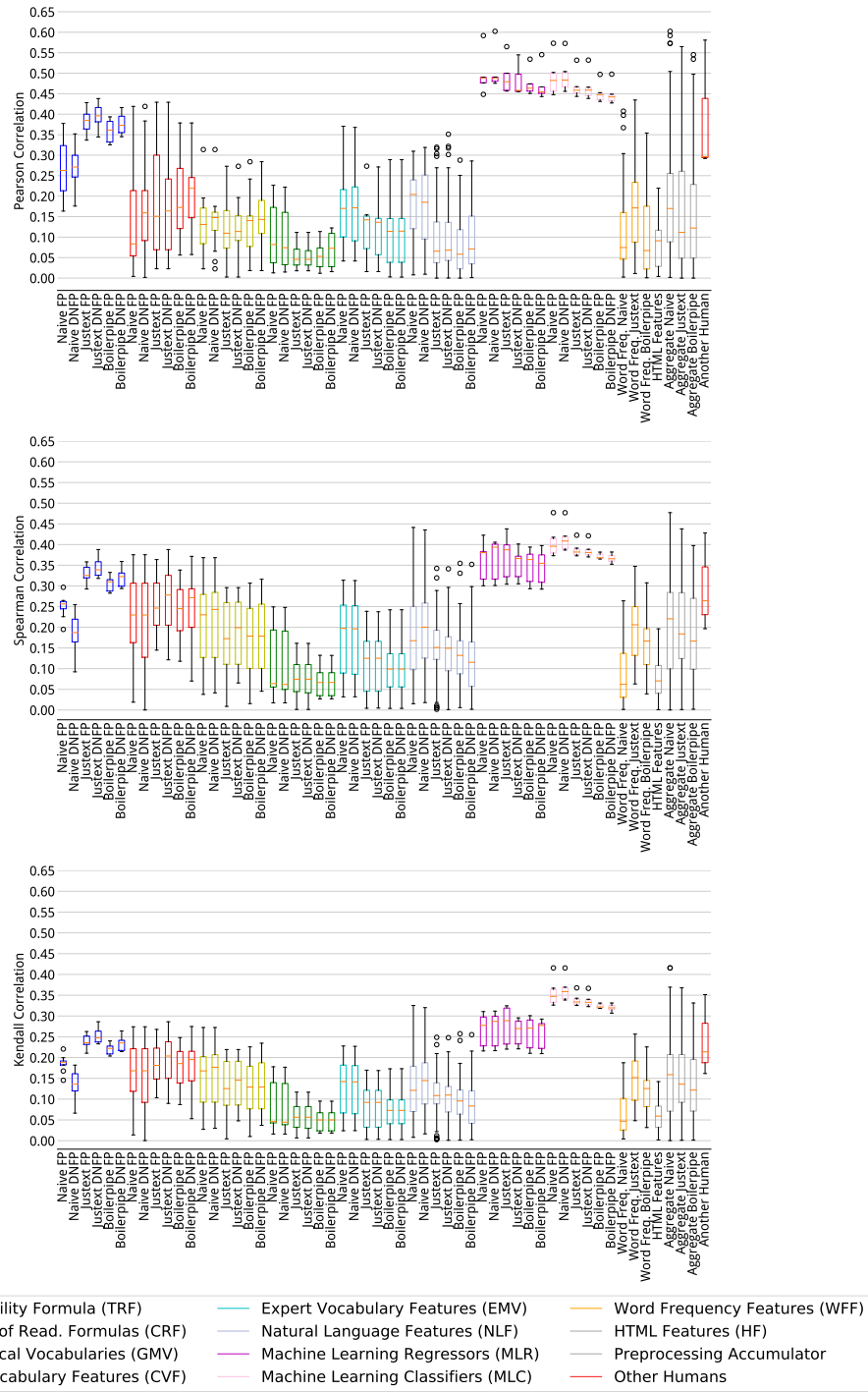


Figure 6.4: Box plots divided by feature groups. Correlations are calculated using understandability labels from relevant documents assessed in CLEF eHealth 2015

6.3. Evaluation of Preprocessing Pipelines and Heuristics

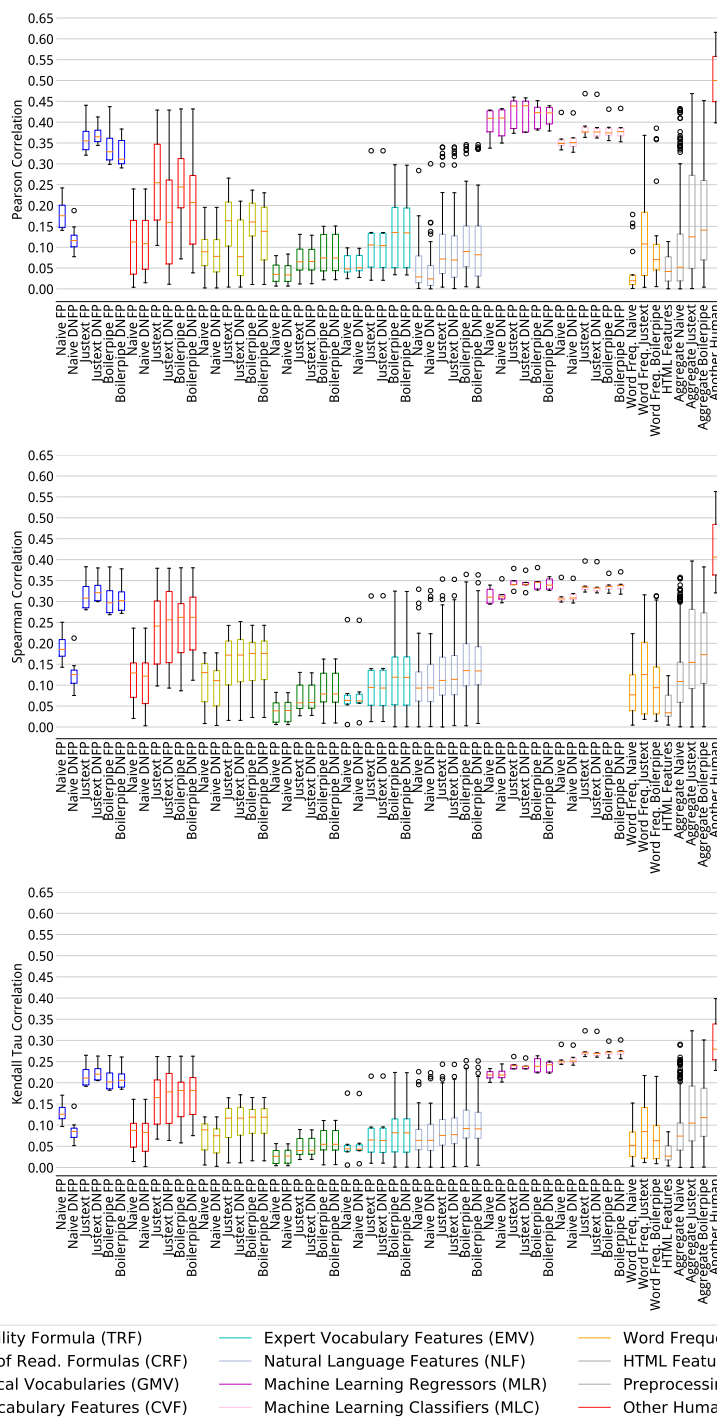


Figure 6.5: Box plots divided by feature groups. Correlations are calculated using understandability labels from relevant documents assessed in CLEF eHealth 2016

are seen in various groups when comparing the strategies that use advanced HTML preprocessing methods (*JusText* and *Boilerpipe*) and those that do not (*Naïve*).

When considering methods beyond those based on readability formulas, we found that the highest correlations were achieved by the regressors (MLR) and classifiers (MLC) trained on our auxiliary corpora, independently of the preprocessing method used. There is little difference in terms of effectiveness of methods in these categories: the largest variances were found for the MLR on CLEF 2015 in which the Neural Network Regressor the Pearson correlation was 0.44 while the Support Vector Regressor was 0.30. This difference of 0.14 is small if compared with the difference of 0.42 found for the Pearson correlation of the NLF group also in CLEF 2015.

A common trend when comparing preprocessing pipelines is that the *Naïve* pipeline provided the weakest correlations with human assessments for CLEF 2016, regardless of estimation methods and heuristics. This result, however, was not confirmed for CLEF 2015, where the *Naïve* preprocessing negatively influenced correlations for the readability formula category (RF), but not for other categories, although it was generally associated with larger variances of the correlation coefficients (i.e., larger differences between the best and the worst method in a boxplot).

To provide a full analysis of our data, we run an Analysis of Variance (ANOVA) test to study the influence of four variables: (1) the collection (CLEF 2015 vs. CLEF 2016); (2) the group of methods to estimate understandability; (3) the use of *ForcePeriod* vs. *DoNotForcePeriod*; and (4) the use of *Naïve* vs. *Boilerpipe* vs. *JusText*. We show the results of the ANOVA test using only the Spearman correlation, but the same significant differences are found when conducting these experiments with Pearson and Kendall correlation. The results of the ANOVA test comparing the Spearman correlation (mean \pm standard deviation) show that:

- **CLEF 2015 vs. CLEF 2016:** the absolute Spearman correlation results in CLEF 2015 ($.17 \pm .10$) are significantly higher than those in CLEF 2016 ($.14 \pm .10$), $p < .001$.
- **Groups of methods to estimate understandability:** we found that the variance due to different groups is statistically significant ($p < .001$). We conducted a post-hoc-test with the Tukey's honest significance test [195] to detect which means are significantly different from each other. Figure 6.6 shows the mean and confidence interval values for each of the groups studied here. MLR and MLC are not significantly different from each other (the confidence interval of these group groups overlap), but their Spearman correlation is significantly higher than any other group. Also, note that the group with the third highest mean is the Traditional Readability Formulas (TRF), followed by the group of Components of Readability Formulas (CRF).
- **ForcePeriod vs. DoNotForcePeriod:** the absolute Spearman correlation results using *ForcePeriod* ($.167 \pm .104$) are not significantly different from the results using *DoNotForcePeriod* ($.166 \pm .106$), $p = .855$. This result reaffirms the comparison of *ForcePeriod* Vs. *DoNotForcePeriod* made in Section 6.2.

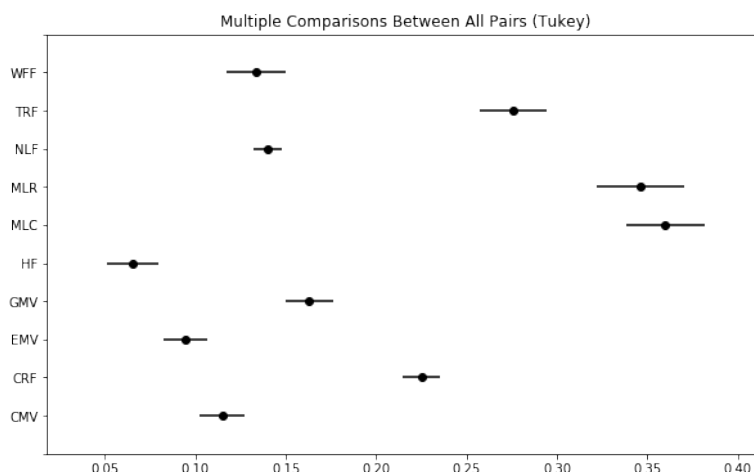


Figure 6.6: Average difference and confidence interval when comparing the Spearman correlation for each group. Groups whose confidence intervals overlap are not significantly different.

- **Naïve vs. Boilerpipe vs. JusText:** we also found that the variance due to the different preprocessing pipelines is statistically significant, $p = .025$. In Figure 6.7, we show that the absolute Spearman correlation when using the *JusText* preprocessing is significantly higher than when using the *Naïve* approach. No statistically significant differences were found between the use of *Boilerpipe* and *JusText* and *Boilerpipe* and *Naïve*. This result is different from the one found in Section 6.2, as now we also consider the other groups of methods and the preprocessing is less important for them than for the group of readability formulas. Still, results point to the fact that the *Naïve* method should be avoided.

6.4 The Best Understandability Estimators

We report in Table 6.6 the methods for understandability estimation of each group with the highest correlation coefficient. For example, the method that counts the number of words not found in Aspell dictionary when the documents were preprocessed with *JusText DoNotForcePeriod* showed the highest Pearson correlation (0.351) among the Natural Language Processing methods in CLEF 2015. For the same group and collection, the method that counts the number of pronouns per word with documents processed with *Naïve ForcePeriod* obtained the highest Spearman (0.441) and Kendall- τ (0.325) correlations.

For the methods in the Traditional Readability Formula (TRF) group, SMOG had the highest correlations for CLEF 2015 and DCI for CLEF 2016, regardless of correlation measure. These results resonate with those obtained for the group of Components of

6. ANALYSING DOCUMENTS: UNDERSTANDING UNDERSTANDABILITY THROUGH CORRELATION ANALYSIS

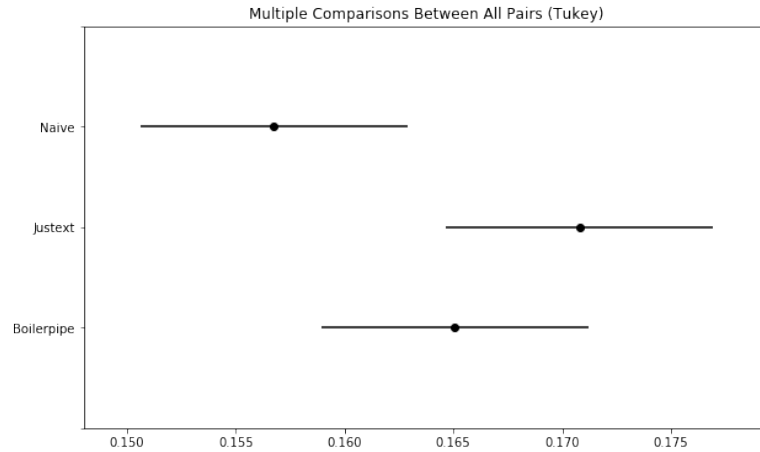


Figure 6.7: Average difference and confidence interval when comparing the Spearman correlation of the three preprocessing HTML tools we validated. The mean absolute Spearman correlation of the JusText preprocessing is significantly higher than the Naïve preprocessing. *Naïve* and *Justext* are the only groups that are significantly different, i.e., their confidence intervals do not overlap.

Table 6.6: Methods with highest correlation per group. In bold are the methods that achieved the highest correlation for a correlation measure.

Group	CLEF 2015					CLEF 2016				
	Method	Prep.	Pears.	Spear.	Kend.	Method	Prep.	Pears.	Spear.	Kend.
TRF	SMOG Index	Jst NFP	.438	.388	.286	Dale Chall Index	Jst FP Boi FP	.439 .437	.381 .382	.264 .264
CRF	Avg. N. of Polysyl. Words per Word Avg. N. of Polysyl. Words per Sentence	Jst FP Jst NFP	.429 .192	.364 .388	.268 .286	Avg. Difficult Words Per Word	Boi FP	.431	.379	.262
GMV	Avg. N. Medical Prefixes per Word N. of Medical Prefixes	Nai FP	.314 .131	.312 .368	.229 .272	Avg. Prefixes per Sentence ICD Concepts Per Sentence	Jst FP Jst NFP	.263 .014	.242 .253	.164 .172
CMV	CHV Mean Score for all Concepts	Nai FP	.371	.314	.228	CHV Mean Score for all Concepts	Jst FP Boi FP	.329 .329	.313 .325	.216 .224
EMV	N. of MeSH Concepts	Nai FP	.227	.249	.178	N. of MeSH Concepts N. of MeSH Disease Concepts	Nai NFP Boi NFP	.201 .179	.166 .192	.113 .132
NLF	N. of words not found in Aspell Dict. N. of Pronouns per Word	Jst NFP Nai FP	.351 .271	.276 .441	.203 .325	Avg. Stopword Per Word N. of Pronouns	Boi FP Boi FP	.344 .341	.312 .364	.213 .252
HF	N. of P Tags	None	.219	.196	.142	N. of Lists N. of P Tags	None	.114 .110	.021 .123	.015 .084
WFF	Mean Percentile Medical Reddit - Inc. OV 25th percentile Pubmed	Jst NFP	.435 .330	.277 .347	.197 .256	Mean Percentile Medical Reddit 50th percentile Medical Reddit	Boi NFP Jst NFP	.387 .351	.312 .315	.214 .216
MLR	XGB Regressor	Boi NFP Jst FP	.602 .565	.394 .438	.287 .324	XGB Regressor Random Forest Regressor	Jst NFP Boi NFP	.454 .389	.373 .355	.258 .264
MLC	Multinomial Naive Bayes	Nai FP	.573	.477	.416	Multinomial Naive Bayes	Jst FP	.461	.391	.318

Readability Formulas (CRF). In fact, the method that counts the number of polysyllable words in documents, which is the main component of the SMOG readability formula, had the highest correlation for CLEF 2015 among CRF methods. Similarly, the number of difficult words, which is the main component used in DCI, had the highest correlation for CLEF 2016 among CRF methods.

For the General Medical Vocabulary (GMV) group, the counts of medical prefixes and

ICD concepts per sentence obtained the highest correlation. Note that, although the Spearman and Kendall correlation for ICD concepts per sentence was high in CLEF 2016, its Pearson correlation was near zero. Overall, Spearman and Kendall correlations obtained similar results (in terms of which methods exhibited the highest correlations): this was expected as, unlike Pearson, they are both rank-based correlations.

When examining the Expert Medical Vocabulary (EMV) group, we found that the number of MeSH concepts obtained the highest correlations with human assessments; however, its correlations were considerably lower than those achieved by the best method from the consumer medical vocabulary group, i.e., the scores of CHV concepts. For the Natural Language Feature (NLF) group, we found that the number of pronouns, the number of stop words and the number of out of vocabulary words had the highest correlations – and these were even higher than those obtained with MeSH and CHV based methods. In turn, the methods that obtained the highest correlations among the HTML features group (counts of P tags and list tags) exhibited overall the lowest correlations compared to methods in the other groups. P tags are used to create paragraphs in a Web document, being thus a rough proxy for text length. Among methods in the Word Frequency Features (WFF) group, the use of Medical Reddit (but also of PubMed) showed the highest correlations, and these were comparable with those obtained by the readability formulas. Finally, the groups with the highest correlated estimators are the regressors and classifiers trained on the auxiliary corpora, with top estimators being the eXtreme Gradient Boosting regressor and the multinomial Naïve Bayes.

Our intention in this section is to explicitly document what worked better in each group. Next, the methods described in this chapter will be used as the features to represent a document. With this, we will be able to build classifiers and regressors to estimate the understandability of a document automatically.

6.5 Predicting Document Understandability

The methods to estimate understandability described in Section 6.1.2 can also be combined to build models with even better estimations of document understandability. The typical machine learning approach is to use each method of Section 6.1.2 as a feature to represent a document. These features are then utilized to build automatic models to predict understandability.

Note that machine learning has already been used in this chapter when developing the methods belonging to the MLC and MLR groups in Section 6.1.2. The goal of the methods in the MLC and MLR is the same as the machine learning methods in this section: predict understandability. The difference is the input of these models. For the methods in the MLC and MLR groups, we extracted features from the auxiliary corpora. For the machine learning methods of this section, we use as features the various methods of Section 6.1.2, including the predictions made by methods in the MLC and MLR groups.

The scale of assessment labels of CLEF 2015 and 2016 collections are different. Documents in CLEF 2015 were assessed using a 4 level scale, while documents in CLEF 2016 were assessed using a 101-level scale. We take advantage of these differences to experiment with both classification and regression tasks. For that, in Section 6.5.1, with the CLEF 2015 collection, we propose a *multiclass classification task* to estimate the understandability of documents using the four different labels as four distinct classes (i.e., on one end, we have a class for documents that are very hard to understand and, on the other end, a class for documents that are very easy to understand). In Section 6.5.2, with the CLEF 2016 collection, we propose a *regression task* in which an understandability value ranging from 0 to 100 is estimated for each document exploring the original assessments created by human assessors in CLEF 2016.

For each collection, we perform ten-fold cross-validation so that 10% of the documents are separated in a test set and the rest 90% is used for training, tuning hyper-parameters and feature selection. Each time, a different and disjoint set of documents is used in the test set and results are average across the ten experiments. We evaluate the regressors and classifiers with standard evaluation metrics. For the classification task with the CLEF 2015 collection, we used the accuracy score (Acc.)¹⁹ and the macro F_1 score ($Mac.F_1$)²⁰. For the regression task with the CLEF 2016 collection, we used the coefficient of determination (R^2)²¹ and the mean absolute error (MAE)²². We concatenate the predictions of each one of the ten folds to calculate the (Pearson, Spearman and Kendall) correlation between the predicted understandability scores and the human assessed scores for the documents. This allows us to compare the correlation of the machine learning predictions with the correlation of other methods previously shown in Figures 6.2 and 6.3.

We experiment with various methods for feature selection. We take advantage that the features from Section 6.1.2 were divided into groups to evaluate the importance of each group separately. For that, we train models using only one group at the time and models excluding only one group at the time (ablation study). We also experiment with the three following approaches to select features across groups. The first approach is to use only the *best* features of each group of features; those were shown in Table 6.6 for both CLEF 2015 and 2016. The second approach is to use only features which obtained a (Pearson or Spearman or Kendall) correlation degree with the human assessments in the training set higher than a pre-defined threshold (we experiment with both .30 and .40 as thresholds). The third approach is based on tree methods: the idea is that after fitting a

¹⁹Accuracy counts the number of correctly classified documents.

²⁰The macro F_1 is the average of the F_1 score ($F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$) calculated for each class separately.

²¹ R^2 is the proportion of the variance in the predicted scores that is predictable from the model. It usually varies from 0.0 (model predictions are independent of the data) to 1.0 (model predictions perfectly fit the data), but it can also yield negative values meaning that even a constant value is better than the predictions made by the model.)

²²This is the only measure used in the experiment of this chapter in which lower values mean better predictions.

tree-based machine learning method, such as Random Forest [22] or Extra Tree [73], we can calculate how important features are by counting how often a feature is used as split points of the trees in the model. We fit an initial model with all features and assign a value to every feature according to their importance. A feature with importance greater than the mean importance of all features is kept, otherwise discarded.

Significance tests are conducted with a paired two-tail t-test with Bonferroni correction [69]. The traditional $p < 0.05$ is adjusted per collection according to the number of experiments. In total, we make 28 t-tests when we compare the machine learning model using all features with all other settings, resulting in a $p = .0018$ ($0.05/28$) according to the Bonferroni correction.

To keep our analysis concise, we show only the results of the Extra Tree classifier (CLEF 2015) and regressor (CLEF 2016) implemented in the sklearn Python package²³. The results of other machine learning algorithms are similar to the ones shown next.

6.5.1 Classification Task with CLEF 2015 Collection

We first define four baselines to use in the experiments in this section: (1) *Most Frequent Class Classifier*: always outputs the most frequent class in the dataset; (2) *Stratified Class Classifier*: generates predictions respecting the training set’s class distribution; (3) *Thresholds*: three thresholds are learned from the distribution of scores of a readability formula. The readability formula is then applied to the documents in the test set and documents are classified depending on the thresholds learned in the training set. We used the values at the 25th, 50th and 75th percentile of the distribution of the readability formula’s scores in the training set. For example, if the thresholds were $th_1 = 10.29$, $th_2 = 11.23$ and $th_3 = 13.13$, a document with a score of 9.42 would be classified as “*Very Easy to Understand*”. (4) *ML on Single Feature*: an Extra Tree classifier trained only on the scores predicted by a single readability formula. We selected the SMOG Index as the readability formula for baselines (3) and (4) because it showed the best correlation with human assessments among all readability formulas for topically relevant documents in CLEF 2015 (Table 6.6).

In Table 6.7, we report the results for the classification task with the CLEF 2015 collection. The best baseline method in terms of accuracy is the *Most Frequent Class Classifier*, but its predictions are not useful for IR systems. Retrieval methods would not be able to promote easy-to-read documents, as all documents are assigned to the same class.

The *ML on Single Feature (SMOG)* baseline presents the highest accuracy and Macro F_1 among all baseline methods. The *Threshold (SMOG)* shows the highest correlation results post-classification. Note that for these two baselines, the correlation results obtained here (e.g., Kendall correlation of .33) are similar to the ones obtained when directly correlating the SMOG index scores with the human assessments (in Table 6.6, Pearson correlation of .28). The baselines show that the scores of readability formulas can be

²³<http://scikit-learn.org/0.18/modules/ensemble.html#forest>

6. ANALYSING DOCUMENTS: UNDERSTANDING UNDERSTANDABILITY THROUGH CORRELATION ANALYSIS

Table 6.7: Results for the document understandability classification task using the relevant documents from CLEF 2015 collection. Accuracy and Macro F_1 (both Mean \pm Std) were calculated across the ten folds in a 10-fold cross-validated experiment. Results of each fold are accumulated and correlation with human assessments is shown in the last three columns. Bold is used to show the best values per experiment.

Experiment	Features	Class. Res.		Correlation Analysis		
		Acc.	Mac. F_1	Pears.	Spear.	Kend.
Baselines	Most Frequent Class Classifier	.44 \pm .06	.15 \pm .02	-	-	-
	Stratified Class Classifier	.37 \pm .03	.24 \pm .03	.03	.03	.03
	Thresholds (SMOG)	.19 \pm .04	.14 \pm .02	.41	.38	.33
	ML on Single Feature (SMOG)	.54\pm.04	.44\pm.06	.37	.36	.33
Only one Feature Group at time	Trad. Readability Formulas	.60 \pm .04	.56 \pm .06	.56	.50	.47
	Components of Read. Formulas	.61 \pm .04	.56 \pm .05	.57	.51	.48
	General Medical Voc. Features	.60 \pm .04	.53 \pm .06	.49	.46	.43
	Consumer Vocabulary Features	.60 \pm .04	.53 \pm .06	.47	.44	.42
	Expert Vocabulary Features	.60 \pm .04	.51 \pm .06	.43	.41	.38
	Natural Language Features	.62\pm.04	.58 \pm .06	.58	.52	.48
	Machine Learning Regressors Feats.	.61 \pm .04	.60\pm.05	.61	.53	.49
	Machine Learning Classifiers Feats.	.52 \pm .04	.51 \pm .05	.53	.42	.39
	Word Frequency Features	.61 \pm .04	.58 \pm .05	.59	.52	.49
HTML Features	.61 \pm .04	.59 \pm .05	.60	.52	.49	
All Features		.64\pm.04	.63\pm.05	.65	.57	.53
Ablation Analysis	Trad. Readability Formulas	.63 \pm .04	.63\pm.05	.65	.57	.53
	Components of Read. Formulas	.64\pm.04	.63\pm.05	.65	.57	.53
	General Medical Vocabulary	.64\pm.04	.63\pm.05	.65	.57	.53
	Consumer Vocabulary	.64\pm.04	.63\pm.05	.65	.57	.53
	Expert Vocabulary	.63 \pm .04	.63\pm.05	.65	.56	.53
	Natural Language Features	.64\pm.04	.63\pm.05	.65	.57	.54
	Machine Learning Regressors	.63 \pm .04	.63\pm.05	.64	.56	.53
	Machine Learning Classifiers	.63 \pm .04	.62 \pm .05	.64	.56	.53
	Word Frequency Features	.63 \pm .04	.63\pm.05	.65	.57	.53
HTML Features	.63 \pm .04	.63\pm.05	.65	.56	.53	
Feature Selection	Only Best Features	.64\pm.04	.63\pm.05	.65	.57	.53
	Only Features Correlation $>$.40	.63 \pm .04	.62 \pm .05	.64	.55	.52
	Only Features Correlation $>$.30	.63 \pm .04	.62 \pm .05	.64	.56	.52
	Feature Selection Trees	.64\pm.04	.63\pm.05	.65	.57	.53

directly used even without machine learning (as in the *Threshold* approach), but the use of machine learning, even with a unique feature can significantly improve the results.

The second part of Table 6.7 shows the results of only training the Extra Tree classifier with features of one group at the time. Our intent with this experiment is to measure the contribution of each group of features individually. The best results were obtained by a classifier trained with features of the *Natural Language Features* group, while the worst were obtained using the features of the *Machine Learning Classifiers Features* group.

The factors that help to explain the performance of the classifiers with both *Natural Language Features* and *Machine Learning Classifier Features* groups are linked to the characteristics of the features in each group. One large difference between these two groups is the number of features in each group. *Natural Language Features* group had a total of 342 features (= 19 features \times 3 variants (i.e., raw counts and counts normalized by the number of words and sentences in the documents) \times 6 preprocessing pipelines (e.g., *JusText ForcePeriod*, *JusText DoNotForcePeriod*, etc)), while the *Machine Learning Classifier Features* had only 36 features. A larger feature set is more likely to have more variety among the features in the group and that is the case in here. The features in the *Natural Language Features* group are diverse, with features that could indicate how long a document is (e.g., number of nouns or verbs in the document) and how rare the vocabulary used in a document is (e.g., number of words not found in the Aspell dictionary). However, the features in the *Machine Learning Classifier* group are limited to only 3 possible values (1, 2 or 3) and with a substantial overlap among themselves, as those were extracted from classifiers predicting if a document should be classified as *Medical Reddit*, *Medical English Wikipedia* or *PubMed Central* (see Section 6.1.2). Note that the *Machine Learning Regressor Features* group also had a small number of features, 30, but the fact that the predictions were made in continuous scale²⁴ seems to have allowed the Extra Tree classifier to explore the search space better and provide more accurate predictions than using the *Machine Learning Classifier Features* group. Also note that, although the features in group of *HTML Features* in Table 6.6 did not correlate well with the human assessments (i.e., the number of p tags, its best method, showed a weaker correlation than the best method of any other group of features), they could successfully be used by a machine learning classifier showing results as good as the other groups of features. Finally, note that the Macro F_1 of the classifier using the *Expert Medical Vocabulary Features* is, together with the classifier using the *Machine Learning Classifier Features*, the worst one. The post-classification results for the *Expert Medical Vocabulary Features* are also the worst one for all correlation measures. This group focus on exploring features related to the MeSH hierarchy, which has already been used successfully in the literature to classify documents according to their expertise [213]. However, our results indicate that features extracted from the MeSH hierarchy are not particularly useful.

Although the classification accuracy of .64 between the classifier using all features and the human assessments of CLEF 2015 seems low (i.e., one out of three predictions are wrong), it is important to compare this accuracy with the ones obtained by comparing different sets of human assessments among each other. In fact, if the other sets of human assessments were used as understandability predictions, their accuracy would not be higher than .45, even though they would still be useful to rank retrieval systems [153]. Apart from it, the correlation coefficients obtained by the classifier using all features (Pearson correlation of .65, Spearman of .57 and Kendall of .54), which are more important for a retrieval task, are as high as the highest ones obtained by human assessors (Figure 6.4). Both accuracy

²⁴That is, the values from the MLR features could assume any real number value although the set of training labels was the same as in the MLC: 1, 2 and 3.

score and correlation coefficients indicate that the classifier using all features is as good as humans in ranking documents regarding their understandability.

6.5.2 Regression Task with CLEF 2016 Collection

We also define four baselines in the regression experiments: (1) *Mean Value*: always outputs the mean understandability score from the training set; (2) *Median Value*: always outputs the median understandability score from the training set; (3) *MinMax Scaling*: the scores obtained by a readability formula are converted in a scale between 0 (minimum) and 100 (maximum). (4) *ML on Single Feature*: a machine learning regressor trained only on the scores of a single readability formula. For baselines (3) and (4), we use the Dale-Chall Index as readability formula due to its correlations with human assessments in the previous experiments with CLEF 2016 (Table 6.6). We only show the results of the Extra Tree Regressor in this section, but note that similar results are obtained by other machine learning regressors.

The first part of Table 6.8 shows the results of the baseline methods. The predictions of the *Mean Value* and *Median Value* models are constant for each fold, resulting in an R^2 close to 0. The post regression analysis also shows that correlations with the human assessments are close to 0. While the use of Dale-Chall in the *MinMax Scaling* and *ML on Single Feature* methods resulted in predictions that are equivalent or worse than simply predicting the mean or median value, e.g., $R^2 = -.23$ for the *ML on Single Feature* method, the post regression analysis of *MinMax Scaling* shows higher correlations than the other baselines, with correlation coefficients as high as the ones seen in Figure 6.4.

The second part of Table 6.8 shows the contribution of each group of features individually. As seen in the experiments in Section 6.5.1, the results using the *Natural Language Features* were the best ones, while the ones using *Machine Learning Classifiers Features* were among the worst ones. The worst results in terms of both R^2 and MAE were seen in the experiments using only the *Expert Vocabulary Features* group, showing once more that the MeSH hierarchy features are not particularly useful for this task.

Inline with our findings in Section 6.5.1, the results of a regressor that uses all features are significantly better than a regressor that uses only one group of features (for all groups, $p < .0018$) and are not significantly different from a regressor that uses all group of features except one (for all groups, $p > .0018$).

The last part of Table 6.8 shows the attempts to select features across different groups. The best results are accomplished by the feature selection using the feature importance of the Extra Tree regressor, however these results are not statistically different from the results of the regressor using all features ($p > .0018$). The other methods resulted in regressors that are significantly worse than the regressor with all features (for all comparisons, $p < .0018$).

Finally, note the results of the correlation analysis made with the outcomes of the best regressor (with all features): Pearson correlation of .59, Spearman of .50 and Kendall of

Table 6.8: Results for the document understandability classification task using the relevant documents from CLEF 2016 collection. R^2 and Mean Absolute Error (both Mean \pm Std) were calculated across the ten folds in a 10-fold cross-validated experiment. Results of each fold are accumulated and correlation with human assessments is shown in the last three columns. Bold is used to show the best values per experiment.

Experiment	Features	Reg. Results		Correlation Analysis		
		R^2	MAE	Pears.	Spear.	Kend.
Baseline	Mean Value	.00 \pm .00	17.45 \pm .80	.05	.05	.03
	Median Value	.04 \pm .02	17.02\pm.86	.03	.03	.03
	MinMax Scaling (Dale-Chall)	.05\pm.07	17.26 \pm .44	.44	.38	.26
	ML on Single Feature (Dale-Chall)	-.23 \pm .11	18.68 \pm .85	.31	.23	.16
Only one Feature Group at time	Trad. Readability Formulas	.23 \pm .05	14.84 \pm .62	.50	.40	.28
	Components of Read. Formulas	.26 \pm .05	14.59 \pm .70	.53	.44	.31
	General Medical Voc. Features	.19 \pm .05	15.15 \pm .71	.46	.39	.27
	Consumer Vocabulary Features	.22 \pm .05	14.93 \pm .76	.48	.42	.29
	Expert Vocabulary Features	.10 \pm .05	15.99 \pm .72	.37	.31	.21
	Natural Language Features	.28\pm.04	14.28\pm.67	.54	.45	.31
	Machine Learning Regressors Feats.	.24 \pm .07	14.75 \pm .73	.51	.42	.29
	Machine Learning Classifiers Feats.	.14 \pm .07	15.70 \pm .82	.44	.34	.23
	Word Frequency Features	.26 \pm .06	14.64 \pm .67	.52	.40	.28
HTML Features	.11 \pm .05	15.87 \pm .78	.39	.32	.22	
All Features		.33\pm.04	13.62\pm.59	.59	.50	.35
Ablation Analysis	Trad. Readability Formulas	.34\pm.05	13.62 \pm .66	.59	.50	.35
	Components of Read. Formulas	.34\pm.05	13.62 \pm .62	.59	.50	.35
	General Medical Voc. Features	.34\pm.05	13.58 \pm .61	.59	.50	.36
	Consumer Vocabulary Features	.33 \pm .05	13.69 \pm .65	.59	.49	.35
	Expert Vocabulary Features	.34\pm.05	13.57\pm.65	.59	.51	.36
	Natural Language Features	.33 \pm .06	13.64 \pm .62	.59	.50	.36
	Machine Learning Regressors Feats.	.33 \pm .05	13.65 \pm .68	.59	.49	.35
	Machine Learning Classifiers Feats.	.34\pm.05	13.61 \pm .62	.59	.50	.36
	Word Frequency Features	.34\pm.04	13.60 \pm .54	.59	.50	.35
HTML Features	.33 \pm .05	13.65 \pm .70	.59	.50	.35	
Feature Selection	Only Best Feats.	.29 \pm .04	14.15 \pm .64	.56	.45	.31
	Only Features Corr. > .40	.24 \pm .05	14.65 \pm .65	.52	.43	.30
	Only Features Corr. > .30	.30 \pm .05	14.07 \pm .62	.56	.47	.33
	Feature Selection Trees	.34\pm.05	13.58\pm.64	.59	.50	.36

.35. We can directly compare these values with the boxplots shown in Figure 6.5. For all results shown in Figure 6.5, only the highest correlation made by human assessors is higher than the correlation results shown by the regressor with all features. This shows that the model using all features is as good as humans in ranking documents according to their understandability. In Chapter 8 we investigate how to integrate the machine learning methods investigated in this chapter into retrieval systems in order to boost understandable documents among the topically relevant ones.

6.6 Summary

There is an abundance of factors that affect how users perceive understandability. In this chapter, we devised and studied a large number of understandability estimators, ranging from traditional readability formulas extensively used in the past 50 years to state-of-the-art machine learning algorithms build with the auxiliary corpora. We grouped them into semantically related groups in order to facilitate the investigation of their correlation with human assessments collected during CLEF eHealth campaigns in 2015 and 2016.

Complementary to Chapter 5, we evaluated how preprocessing steps impact the understandability estimation in traditional readability formulas and other modern estimators. We empirically learned the importance of preprocessing steps when applying readability formulas, as the highest correlations happen when proper HTML cleaning methods are used. For the most modern estimators, such as the ones based on machine learning methods, the correlation is less sensitive to the preprocessing steps.

We also studied the correlation of each readability formula with the human assessment to provide insights on which formula should be preferred. Our analysis did not conclude that one single formula is better than the others in particular. In fact, we could only find one formula that is significantly worse than the others: the ARI Index. The correlation results between the various methods to estimate understandability and human assessments are higher when preprocessing text with *JusText* than when using *Boilerpipe* or the *Naiïve* strategies, although they are not significantly different from using *Boilerpipe*. We also did not find significant differences between the use of *ForcePeriod* and *DoNotForcePeriod*.

Although we did not show how well each method in Section 6.1.2 correlated with the human assessments due to the vast number of methods tested in this chapter, we explicitly show what are the best methods in each group, providing an overview of what works better in each group. Finally, we showed that machine learning algorithms can be successfully used to combine methods to estimate understandability, reaching results that are significantly higher than the methods used individually. The experiments in this chapter serve as a basis for the following chapters of this thesis, as the learning-to-rank methods take advantage of the estimators devised and analyzed here.

Part IV

Understandability in Search Engines

Multidimensional Evaluation of Search Engines

The topicality of a document to a query or information need is central to the notion of relevance, but other factors (also called dimensions) that influence the relevance of a document do exist. In fact, researchers have long established that the notion of relevance in information retrieval is multidimensional [173, 20]. These dimensions include novelty, diversity, timeliness, scope, understandability and trustworthiness, among others [159, 173]. In particular, in the context of consumer health search, the relevance dimensions of understandability and information trustworthiness are fundamental [90]. It means that health information is only valuable to users, allowing them to make appropriate health decision, if it is understandable and correct. It is therefore important to take into account these additional relevance dimensions, along with topicality, when evaluating the effectiveness of search systems in the context of consumer health search tasks, and in general in other tasks with similar requirements.

An evaluation framework that integrates understandability into information retrieval evaluation has been recently devised [224, 223] and it has been largely adopted to evaluate systems for consumer health search [154, 226, 155]. The framework, named *Understandability-Biased IR Evaluation* (UBIRE), builds upon the gain-discount framework of evaluation measures used in information retrieval (measures like normalized Discounted Cumulative Gain (nDCG), Expected Reciprocal Rank (ERR), Rank Biased Precision metric (RBP) belong to this framework) [26]. UBIRE uses a discount based on the rank position at which documents are retrieved, and a gain function that integrates contributions from both topicality and understandability (see Section 7.1).

A limitation of the approach used to model multidimensional relevance in UBIRE is that it is not trivial to identify how different dimensions of relevance affect the final evaluation score. This is because in UBIRE gains produced by documents for each of the considered

dimensions of relevance are combined early on in the evaluation measure. This limitation makes the interpretation of evaluation results using UBIRE difficult as it is impossible to determine whether improvements (deteriorations) are due to more (less) understandable or more (less) topical documents being retrieved.

In this chapter, we propose an alternative to UBIRE, called the *MM* (for **M**ultidimensional **M**etric), which overcomes the interpretability limitation of UBIRE, while still enabling the combination of multidimensional relevance evidence when evaluating information retrieval systems (Section 7.2). Using small synthetic data, we show the intuitive differences between UBIRE and *MM* and demonstrate how *MM* overcomes UBIRE’s limitation (Section 7.3). We further empirically compare specific measures instantiated from the two frameworks using real data to study system ranking correlations across UBIRE and *MM* (Section 7.4). The results show that while system correlations measured with *MM* are aligned with UBIRE, *MM* provides richer information to researchers, allowing them to assess and control how each relevance dimension contributes to the evaluation score of a system.

7.1 Incorporating Understandability into Evaluation Metrics

The understandability based framework of Zuccon [223] is based on the gain-discount framework by Carterette [26], which can be generically defined as an evaluation metric \mathcal{M} as:

$$\mathcal{M} = \frac{1}{\mathcal{N}} \sum_{k=1}^K \mathbf{d}(k) \mathbf{g}(d@k) \quad (7.1)$$

where $\mathbf{g}(d@k)$ and $\mathbf{d}(k)$ are respectively the *gain function* computed for the (relevance of the) document at rank k (i.e., $d@k$) and the *discount function* computed for the rank k . K is the depth of assessment at which measure \mathcal{M} is evaluated, and $1/\mathcal{N}$ is an optional normalization factor, which serves to bound the value of the sum into the range $[0,1]$ (details in [26]).

The gain-discount framework encompasses measures such as the normalized Discounted Cumulative Gain (nDCG) [102] with $\mathbf{g}(d@k) = 2^{P(R|d@k)} - 1$ and $\mathbf{d}(k) = 1/(\log_2(1+k))$; the expected reciprocal rank (ERR) [29] with $\mathbf{g}(d@K) = (2^{P(R|d@k)} - 1)/2^{\max(P(R|d))}$ and $\mathbf{d}(k) = 1/k$; and the Rank Biased Precision (RBP) with $\mathbf{g}(d@k)$ equal to 1 if $d@k$ is relevant and 0 otherwise and $\mathbf{d}(k) = \rho^{k-1}$ (with ρ representing the user persistence).

The gain provided by a document at rank k can be expressed as a function of its probability of relevance. Without loss of generality, $\mathbf{g}(d@k) = f(P(R|d@k))$, where $P(R|d@k)$ is the probability of relevance given the document at k . When only topical relevance is modeled, $P(R|d@k) = P(T|d@k)$, i.e., the probability that the document at k is topically relevant. For binary relevance, this probability can simply be 1 for relevant documents and 0 for non-relevant documents. For non-binary cases, this probability can be distributed according to the number of relevance levels.

UBIRE extends this framework to consider cases where relevance is modeled beyond topicality to explicitly model other dimensions, such as understandability. This is done by modeling the probability of relevance $P(R|d@k)$ as the joint distribution over all considered dimensions, $P(\delta_1, \dots, \delta_n|d@k)$, where each $\delta_i \in \mathcal{D}$ represents a dimension of relevance, e.g., topicality, understandability. The computation is simplified by assuming that dimensions are compositional events and their probabilities independent (see [223] for more details). The gain function with respect to different dimensions of relevance can then be expressed as:

$$\mathbf{g}(d@k) = f(P(R|d@k)) \quad (7.2)$$

$$= f(P(\delta_1, \dots, \delta_n|d@k)) \quad (7.3)$$

$$= f\left(\prod_{i=1}^n P(\delta_i|d@k)\right) \quad (7.4)$$

Evaluation metrics developed within this framework differ through the instantiations of $f(P(\delta_1, \dots, \delta_n|d@k))$, other than by which dimensions are modeled. Zuccon provided an instantiation that considers both topicality (T) and understandability (U) [223]:

$$\mathbf{g}(d@k) = f(P(R|d@k)) = f(P(T, U|d@k)) = f(P(T|d@k) \cdot P(U|d@k)) \quad (7.5)$$

with $P(R|d@k)$ as the joint $P(T, U|d@k)$ that is in turn computed as the product $P(T|d@k) \cdot P(U|d@k)$ following the assumptions discussed above.

Specific implementations of the UBIRE framework that have been developed in previous work considered the basic gain and discount functions from RBP [134]; an instantiation with understandability [224, 223] has been later extended by jointly considering also trustworthiness [155]. For ease of explanation, we consider the formulation with topicality and understandability; similar considerations apply when also trustworthiness is modeled (as well as other dimensions). In this case, the understandability-biased RBP, $uRBP$, is defined as:

$$uRBP(\rho) = (1 - \rho) \sum_{k=1}^K \rho^{k-1} P(T|d@k) \cdot P(U|d@k) \quad (7.6)$$

$$= (1 - \rho) \sum_{k=1}^K \rho^{k-1} \mathbf{g}_{RBP}(d@k) \cdot \mathbf{g}_U(d@k) \quad (7.7)$$

In the $uRBP$, the function $\mathbf{g}_{RBP}(d@k)$ is the same as the gain in RBP and transforms relevance values into the corresponding gains and, likewise, $\mathbf{g}_U(d@k)$ transforms understandability values into the corresponding gains. If $\mathbf{g}_U(d@k) = 1$ for every document, then only topical relevance affects retrieval evaluation, i.e., every document is considered as having equal understandability (and its highest value) and we obtain the original RBP. Two instantiations of the gain function $\mathbf{g}_U(d@k)$ have been explored in previous work: one binary ($uRBP$) and the other graded ($uRBPgr$). In the binary version $\mathbf{g}_U(d@k) = 1$

if $P(U|d@k) \geq th_U$, where th_U is a threshold on the assessments of understandability (every assessment that is greater than or equal to th_U would generate a gain of 1), and $g_U(d@k) = 0$ otherwise. In the graded version, we rely on the possibility to transform the understandability assessment collected into estimations of $P(U|d@k)$. For example, assessments collected in a Likert scale of 5 levels can be easily converted into estimations ranging from 0.0 to 1.0 with steps of 0.25 (0.0, 0.25, 0.50, 0.75 and 1.0). Assessments collected in a scale from 0 to 100 (e.g., 0 being the lowest level of understandability), could be directly used as estimations of $P(U|d@k)$ or modified by a smoother function. These estimations are then plugged into the metric.

7.2 A new Framework for Multi-Dimension IR Evaluation

A limitation of UBIRE is that it prematurely combines the gains contributed by each dimension of relevance in **one** single step, providing a unique evaluation score [224, 223]. While this allows for the comparison of systems, it does not permit to understand the contribution each dimension had on the evaluation measure. To overcome this limitation, we aim to create a framework which, while still allowing the modeling of multidimensional relevance, is of easy interpretation and for which it is straightforward to track the contribution each relevance dimension had on the final effectiveness score. This is achieved by separating the evaluation of each dimension such that a value for each dimension is calculated separately with respect to its gain and discount, and then these are combined into a unique effectiveness measure. Note that we assume that it is possible to evaluate each measure separately.

The evaluation of each relevance dimension separately is trivial, as it consists in applying the discount and gain function of the underlying evaluation measure, e.g. RBP, to each relevance dimension $\delta \in \mathcal{D}$, where the gains are those associated with the criteria for that specific dimension.

While the outputs of each relevance dimension could be combined with a linear or geometric combination of values, we opt to use the weighted harmonic mean, as it is particularly sensitive to a single lower-than-average value. The same intuition is used to combine recall and precision in the widely used F -measure. Given a (discount-gain) evaluation measure \mathcal{M} and a particular dimension δ , we apply the measure \mathcal{M} to evaluate a list of documents l_δ which have been labeled with respect to dimension δ (i.e., we compute $\mathcal{M}(l_\delta)$). Then, to compute the proposed measure $MM_{\mathcal{M}}$, we combine all $\mathcal{M}(l_\delta)$ for each relevance dimension using the harmonic mean, where each dimension is weighted according to a preferential weight w_δ assigned to each dimension; formally:

$$MM_{\mathcal{M}} = \left(\frac{\sum_{\delta=1}^n w_\delta \cdot \mathcal{M}(l_\delta)^{-1}}{\sum_{\delta=1}^n w_\delta} \right)^{-1} = \frac{\sum_{\delta=1}^n w_\delta}{\sum_{\delta=1}^n \frac{w_\delta}{\mathcal{M}(l_\delta)}} \quad (7.8)$$

Without loss of generality, we instantiate $\mathcal{M} = RBP$ and define the following modification of RBP [134] for each dimension:

- $RBP_t(\rho)$: uses binary topicality assessments (i.e., the usual RBP).
- $RBP_u(\rho)$: uses understandability assessments (either graded or binary; see below for specific instantiations).

Thus Equation 7.8 becomes (we assumed $w_t = w_u$):

$$MM_{RBP(\rho)} = 2 \cdot \frac{RBP_t(\rho) \cdot RBP_u(\rho)}{RBP_t(\rho) + RBP_u(\rho)} \quad (7.9)$$

7.3 Comparing frameworks Through System Simulations

To understand the behavior of UBIRE and MM when facing different IR systems, we first employed synthetic systems to have fine-grained control over our experiments. This allowed knowing a priori what has changed between two system instances and study the effect these changes had on evaluation. In our experiments, along with topicality, we considered understandability, leaving the (trivial) extension to other dimensions to later work. In the following simulations, we controlled the number of topical documents and understandable documents retrieved. We did so by following this two-phase procedure:

1. **Topicality Phase:** we controlled the number of topical documents in a simulated run using a random variable T , $0 \leq T \leq 1$. We constructed a synthetic run by drawing a real number N_i , $0 \leq N_i \leq 1$, for each position i in a ranking. If $N_i \leq T$, we marked the document at position i as relevant, otherwise, we marked it as not relevant. It is expected that a run generated with $T = 0.1$ has 10% of the documents assessed as relevant (90% as non-relevant), while a run with $T = 0.5$ has as many relevant as non-relevant documents.
2. **Understandability Phase:** we controlled the level of understandability of the documents in a synthetic run. In order to create and control the randomness of our synthetic systems, we generated understandability labels using a Gaussian distribution with pre-defined mean μ and variance σ . As previously done in consumer health search collections [226, 155], we forced the understandability labels to be in the interval $[0, 100]$. We fixed a relatively large variance, $\sigma = 40$, to mimic results of previous collections in which the understandability labels had a large variance [226], and we varied the mean μ of the Gaussian from 0 to 100. Figure 7.1 shows the expected label distribution for $\mu = 20, 50, 80$, i.e., $\mathcal{N}(20, 40)$, $\mathcal{N}(50, 40)$ and $\mathcal{N}(80, 40)$. In Figure 7.1 we also included the threshold U used to compute RBP_u (Section 7.2).

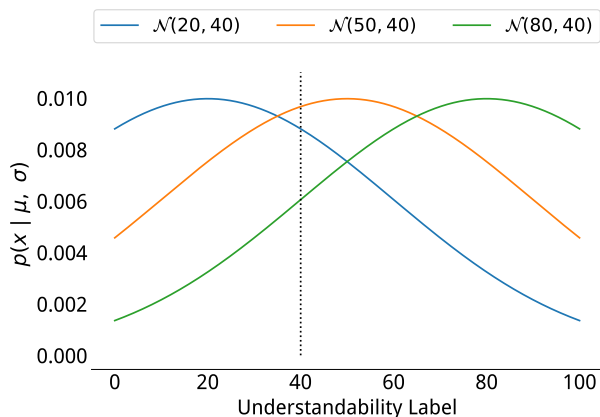


Figure 7.1: Gaussian distribution for different μ : higher μ generates higher understandability labels (more difficult documents were retrieved). In the experiments in this chapter, only documents with understandability lower than 40 are considered easy-to-understand (understandability threshold shown as dotted line).

We executed these two phases in succession. In total, we generated 1,000 runs for each value of T (topicality phase) and value of μ (understandability phase).

We calculated $uRBP$ (using UBIRE) and MM_{RBP} for each synthetic system. The average result of the synthetic runs is shown in Table 7.1. Each row shows the results of simulations with different values for T , i.e., different expected number of topical documents retrieved. We varied μ which was used to create the understandability labels, and show the results for $\mu = 50, 40, 30$. A smaller μ means that more understandable documents were retrieved. The results show that as the expected number of topical documents (T) increases, RBP increases. Likewise, $uRBP$ increases, as it is bounded by topical relevance. In turn, increasing T does not affect RBP_u , but increases MM_{RBP} , as it is also directly dependant on RBP . When the number of understandable documents retrieved is increased (i.e., μ decreased), RBP stays constant, as it does not measure how understandable documents are. In turn, $uRBP$, RBP_u and MM_{RBP} increase. These are the expected behaviors of the considered measures.

We further focused our attention on the results of specific experiments highlighted in blue and yellow in Table 7.1 as they show the advantage of MM framework compared to UBIRE. These cases simulated an initial system S1 that exhibited the results in blue (condition $T = 0.6$ and $\mathcal{N}(40, 40)$) being modified to improve the understandability of retrieved documents ($\mathcal{N}(30, 40)$) at the expenses of topicality ($T = 0.5$), producing a new system S2. The effectiveness of S2 is highlighted in yellow.

If RBP and $uRBP$ were used to decide whether S2 should be preferred over the original system S1, then S2 would be discarded and S1 preferred, as S2 produced an 16% reduction in RBP and a 13% reduction in $uRBP$. With these results, an IR researcher would conclude that the modifications in S2 did not pay off.

Table 7.1: We varied T , the expected proportion of topically relevance documents (rows), and the mean μ of Gaussian distribution used to generate understandability labels (columns). A smaller μ means that easier to read documents are retrieved. We showed the average and standard deviation of each experiment. S1 (blue) and S2 (yellow) represent two systems with similar MM_{RBP} , but very different $RBP(=RBP_t)$ and RBP_u results.

T	Understandability $\mathcal{N}(50,40)$				Understandability $\mathcal{N}(40,40)$				Understandability $\mathcal{N}(30,40)$			
	RBP	uRBP	RBP_u	MM_{RBP}	RBP	uRBP	RBP_u	MM_{RBP}	RBP	uRBP	RBP_u	MM_{RBP}
.3	.29 ± .15	.15 ± .09	.39 ± .17	.30 ± .12	.29 ± .15	.17 ± .11	.50 ± .16	.34 ± .14	.29 ± .15	.19 ± .12	.61 ± .16	.36 ± .15
.4	.39 ± .17	.20 ± .11	.40 ± .17	.36 ± .14	.39 ± .17	.22 ± .12	.48 ± .17	.40 ± .13	.39 ± .17	.25 ± .13	.60 ± .16	.44 ± .14
.5	.50 ± .17	.25 ± .11	.42 ± .16	.42 ± .13	.50 ± .17	.29 ± .12	.50 ± .17	.47 ± .13	.50 ± .17	.33 ± .14	.60 ± .17	.52 ± .13
.6	.60 ± .16	.30 ± .12	.41 ± .16	.46 ± .14	.60 ± .16	.35 ± .12	.50 ± .17	.52 ± .13	.60 ± .16	.40 ± .13	.61 ± .17	.58 ± .13
.7	.70 ± .15	.36 ± .12	.41 ± .17	.49 ± .15	.70 ± .15	.41 ± .13	.51 ± .17	.56 ± .14	.70 ± .15	.46 ± .13	.59 ± .16	.62 ± .12

If MM_{RBP} was used instead, the IR researcher would have been able to gain more insights about system effectiveness and the trade-off between understandability and topicality. To use MM_{RBP} , $RBP_t (=RBP)$ and RBP_u needed to be computed. Between S1 and S2, there was a decrease in RBP_t of 16%; but conversely RBP_u increased by 20%: this clearly allows the trade-off between topicality and understandability to be gauged.

When RBP and RBP_u were combined within MM_{RBP} , if both dimensions were given equal weight, then systems S1 and S2 obtained the same MM_{RBP} . Note that MM can be adapted to specific circumstances: if topicality is more important than understandability, then the weights of each dimension can be changed accordingly in the harmonic mean computation.

7.4 Rank Correlations

Next, we compared the behaviors of MM and UBIRE using real data. For this, we used the systems participating in the CLEF eHealth IR Lab evaluations in 2015 and 2016 [154, 226]. In both these evaluation challenges, systems were evaluated using $uRBP$ – we further evaluated each system using MM and studied the correlations among system rankings obtained using RBP (thus considering topicality only), $uRBP$ (UBIRE), and our proposed RBP_u (thus considering only understandability) and MM_{RBP} . This investigation of correlations is a common approach to compare and understand the relative behavior of evaluation measures [223].

Specifically, we studied a setting where understandability was binary, akin to topicality, which also was considered as binary. For topicality, this was achieved using the common gain function for RBP that only models binary relevance: graded relevance labels were conflated to binary such that highly relevant and relevant assessments were mapped to relevant, and the rest to irrelevant. For understandability, the binarization of the assessments was dependant on the year of the challenge. For 2015, understandability assessments were made on a 4-point scale (very easy, easy, hard and very hard) [154]: we made this binary by assuming that a document marked as very easy and easy was understandable, while we made the remaining as not-understandable. For 2016,

Table 7.2: Kendall- τ correlation for systems participating in CLEF eHealth 2015 and 2016.

	CLEF 2015				CLEF 2016			
	RBP	uRBP	RBP_u	MM_{RBP}	RBP	uRBP	RBP_u	MM_{RBP}
RBP	1.000	0.901	0.483	0.843	1.000	0.948	0.497	0.850
uRBP	0.901	1.000	0.563	0.901	0.948	1.000	0.456	0.866
RBP_u	0.483	0.563	1.000	0.610	0.497	0.524	1.000	0.633
MM_{RBP}	0.843	0.901	0.610	1.000	0.850	0.866	0.633	1.000

understandability assessments were made on an integer scale ranging from 0 (very easy) to 100 (very hard) [226]: we made this binary by assuming that documents with an assessment lower than or equal to 40¹ were understandable, while we made the remaining as not-understandable.

Table 7.2 shows the Kendall- τ rank correlations of systems according to RBP, $uRBP$, RBP_u and MM_{RBP} . Rank correlation between RBP and $uRBP$ was high for both 2015 and 2016 data. This emphasizes the tight relation between RBP and $uRBP$. On the other hand, MM_{RBP} exhibited the strongest rank correlation with RBP_u , while the correlation between RBP_u and RBP or $uRBP$ is marginal. In addition, we found that MM_{RBP} strongly correlated with RBP, but not as strongly as $uRBP$ does. Finally, MM_{RBP} and $uRBP$ showed a generally high correlation among themselves, highlighting that the two measures provided similar evaluations of system effectiveness; however, MM_{RBP} had the advantage that the trade-off between topicality and understandability could be clearly identified and studied.

7.5 Summary

In this chapter, we proposed a new framework, called MM , to evaluate search engines when multidimensional relevance should be considered. Using both synthetic and real data, we compared MM to the understandability-biased information retrieval evaluation framework (UBIRE), which has recently been used to evaluate search systems in the consumer health search domain.

Our experiments showed that while MM correlated well with UBIRE and that both had an equivalent power to rank and distinguish good systems, MM has the advantage of allowing experimenters to easily understand how each relevance dimension affects their system performance, as well as carefully tune the trade-off between topical relevance and other relevance dimensions. While our empirical experiments only considered understandability as an additional dimension to relevance, this was done for directly comparing with UBIRE, and by definition MM naturally accommodates for an unlimited number of relevance dimensions.

¹This threshold was arbitrary chosen because of the semantic and colors used in the Relevance! assessment tool in CLEF 2016 (see Figure A.3). The green part of the slider, which is likely understood as “easy-to-read documents” started at the label 40, thus we re-used this threshold here.

Integrating Understandability into Search Engines

In this chapter, we investigate how understandability estimations can be integrated into retrieval methods to increase the quality of search results. Similarly to Chapter 6, we make extensive use of both CLEF eHealth 2015 and 2016 collections in this chapter. Please refer to the Appendix A for an overview of the CLEF eHealth collections.

We describe in Section 8.1 our experimental methodology and approaches to integrate understandability into retrieval. Three approaches exploring the understandability estimators from Chapter 6 are proposed. They are evaluated with measures described in Section 8.2, which are based on the frameworks for multidimensional evaluation studied in Chapter 7. The result of our experiments is shown in Section 8.3. A summary of our main findings is in Section 8.4.

8.1 Methods to Integrate Understandability into Retrieval

We consider three different strategies to integrate understandability into retrieval: (1) *re-ranking with an understandability estimator*, (2) *rank fusion*, and (3) *learning-to-rank*. *Re-ranking with readability formula* and *rank fusion* are approaches that can be applied to any ranking list, i.e., these two approaches re-rank documents according to their understandability assuming that a retrieval system, seen as a black box, retrieved documents exclusively according to their topical relevance to the query. The *learning-to-rank* approach, however, in a single step integrates the topical relevance and understandability of the documents. For that, the retrieval system cannot act as a black box and, as we investigate in this chapter, needs to be modified to consider the understandability of the documents. Any of these strategies can be directly employed in websites such as

HealthOnNet.org (HON) or HealthDirect.gov.au, allowing the health consumers of these websites to have access to documents that are relevant and easy-to-understand at the same time.

Re-ranking with an understandability estimator, presented in Section 8.1.1, is a straightforward strategy to increase the understandability level of the retrieved results. Given that a (complex) search engine is capable of retrieving highly relevant results, an understandability estimator (such as a readability formula) can be used to re-rank these results according to their understandability. This approach assumes that re-ranking the top documents returned by such a search engine can improve the user satisfaction by increasing the understandability of the top results without hurting their topicality. *Rank fusion*, presented in Section 8.1.2, aims to balance topicality and understandability automatically. In this strategy, two ranking lists are automatically combined: (1) the results retrieved by the search engine with the focus on topicality, and (2) the re-ranked results from the *Re-ranking with an understandability estimator* focusing on understandability. The assumption is that documents that are at the same time topically relevant and understandable will be favored when the two ranking lists are merged. Finally, the *learning-to-rank* strategy, presented in Section 8.1.3, is based on machine learning methods. These methods create retrieval models taking into account features representing both topicality and understandability of documents.

The *Re-rank with an understandability estimator* and the *rank fusion* require an initial retrieval method. We consider three alternatives for initial retrieval methods in this chapter. These include the best two runs submitted to each CLEF task, and a plain BM25 baseline (default Terrier parameters: $b = 0.75$ and $k_1 = 1.2$). The best submission of CLEF 2015, created by team ECNU [183], was a system based on Terrier’s TF-IDF model that expanded the queries by re-issuing them in Google and collecting words from the titles and snippets associated with the top ten Google results. The runner-up system of CLEF 2015 was built by KISTI [138] and based on Lucene’s Dirichlet Language Model. After an initial retrieval, their best system explored two approaches for re-ranking: concept-based document centrality (CBDC) and cluster-based external expansion model (CBEEM). The best system of CLEF 2016 was created by team GUIIR [182] and was based on a Terrier with Divergence from Randomness model. Their system generated reformulations of the queries exploring synonyms and hypernyms from UMLS and then merged the original query and the generated ones using the Borda rank aggregation algorithm. The runner-up system of CLEF 2016 was again created by ECNU [184] and once more based on query expansion with Google. This time, however, their base system was Terrier’s BM25 model, the same that we use as our main baseline system.

As understandability estimators for the strategies *re-ranking with an understandability estimator* and *rank fusion*, we used the SMOG Index for CLEF 2015 and Dale-Chall Index for CLEF 2016 as they were the best performing readability formulas investigated in Table 6.6. We also use the XGB Regressor as an alternative understandability estimator that, apart of not being a readability formula, showed correlation coefficients higher than those of the best readability formulas for both CLEF collections (this is also shown in

Table 6.6, MLR group).

8.1.1 Re-Ranking with an understandability estimator

To integrate understandability estimators into the retrieval process, we first investigate re-ranking search results retrieved by the initial runs only based on an understandability estimator. If all the search results from a run were to be considered, then such a re-ranking method would place at early ranks web pages highly likely to be understandable, but possibly less likely to be topically relevant. To balance relevance and understandability, we only re-ranked the first k documents. We explored rank cut-offs $k = 15, 20, 50$. Because the evaluation is performed with respect to the first $n = 10$ rank positions, the setting $k = 15$ provides a conservative re-ranking of search results, while $k = 50$ provides a less conservative re-ranking approach. Results are presented in Section 8.3.1.

8.1.2 Rank Fusion

As an alternative to the previous two-step ranking strategy for combining topical relevance and understandability, we explore the *fusion* of two search result lists separately obtained for relevance and understandability. For this, we used the Reciprocal Rank Fusion (RRF) method [41], which was shown effective for combining two lists of search results based on their documents *ranks*, rather than scores. Given a set D of documents to be ranked and a set of rankings T and U , respectively the search result lists for topical relevance and understandability, we compute:

$$RRF(d \in D) = \frac{1}{C + r_T(d)} + \frac{1}{C + r_U(d)}$$

with $r_X(d)$ representing the rank position of document d in the result list X , and C is a constant which mitigates the impact of high rankings by outlier systems. In this work, we use $C = 60$ as in the original formulation of RRF [41]. It is important to note that score-based fusion methods, such as CombMNZ or CombSUM [64], are not recommended for a fusion task such as the one made in this work, as the distribution of topical relevance scores and the understandability scores widely differs.

In our experiments, we used, separately, three retrieval systems for each collection. For CLEF 2015, we used BM25, ECNU [183] and KISTI [138], while for CLEF 2016, we used GUIR [182], ECNU [184] and also BM25. To consider the understandability of the documents retrieved by the retrieval systems, we used the method with the highest correlation to human assessments (according to Pearson correlation for CLEF 2015 - shown in Table 6.6), eXtreme Gradient Boosting (XGB) regressor, which is also studied by the strategy of *re-ranking with an understandability estimator*. Also for this approach, we studied limiting the ranking of results to be considered by the methods across the cut-offs $k = 15, 20, 50$. Results are presented in Section 8.3.2.

8.1.3 Learning-to-Rank

Finally, we considered a third alternative to combine relevance and understandability using *learning-to-rank*. A typical learning-to-rank approach is based on 3 components: (1) a learning-to-rank algorithm and a loss function; (2) a labeling strategy to combine relevance and understandability; (3) features to describe a document. We describe these components next. The results are presented in Section 8.3.3.

Learning-to-rank algorithm and Loss Function

A large number of learning-to-rank algorithms have been recently developed. While details of each algorithm can be found elsewhere [124], they are traditionally divided into three categories: pointwise, pairwise and listwise, depending on how the learning-to-rank problem is modeled.

In the pointwise approach, the learning-to-rank problem is modeled as a regression problem. After a query is issued, a score is predicted for each document and the re-ranked result is simply the ordered list according to this predicted score. The PRanking [44] is an example of an algorithm in this category in which a perceptron algorithm predicts the score of each document.

In the pairwise approach, a pair of documents is compared with a binary classifier built to decide whether a document is more relevant than the other. The Multiple Additive Regression Trees (MART) [68], the RankBoost [67], the RankNet [24] and XGB [31] are popular pairwise approaches.

Finally, the listwise approach directly optimizes a whole list of documents, usually directly evaluating the quality of a ranking with IR measures. The AdaRank [212] and ListNet [25] are examples of this approach.

In this work, we experimented with the algorithms implemented in the Ranklib¹ framework (which includes the aforementioned MART, RankBoost, RankNet, AdaRank and ListNet), and the pairwise learning-to-rank algorithm based on tree boosting (XGB²). We only report the results of the XGB approach in this chapter as it reached the best results among all learning-to-rank approaches.

The loss function used in our experiments to train the learning-to-rank model was *NDCG@10*.

Labeling Strategies

Assigning a label for each document-query pair is an important step of a learning-to-rank setup. Any of the approaches discussed above heavily rely on these labels to infer the importance of a document for a query. Usually, these labels are simply the topical relevance assessments available in a collection, but in our case, we would also like to

¹<https://sourceforge.net/p/lemur/wiki/RankLib/>

²<https://github.com/dmlc/xgboost/tree/master/demo/rank>

consider the understandability labels of the documents. In this work, we explore four strategies to combine the topical relevance and understandability labels with the CLEF 2015/2016 collection.

Let $R_{d,q}$, or just R , be the topical relevance of document d for a query q , and U_d , or just U , be the understandability score of document d , then we can define the different labeling strategies as a function F of U and R . This way, the first labeling strategy, named *relevance only*, can be defined as $F(R,U) = R$. This strategy, akin to how learning-to-rank is used when understandability assessments of the documents do not exist, simply uses the topical relevance of a document as its label. The possible values of R in both CLEF 2015 and CLEF 2016 are 0 for documents that were explicitly assessed as not relevant or documents that were not assessed for the query, i.e., documents that were not in the pool set of the collection; 1 for documents assessed as somewhat relevant for the query; and 2 for documents assessed as highly relevant for the query.

The second strategy, named *proportional*, is to assign labels to documents not only proportionally to their topical relevance, but also to their understandability score. This is done by considering how far the understandability score is from the score of the easiest-to-understand document in the collection. Remember that in CLEF 2015, the easiest-to-read documents were assigned a score $U = 3$ while the hardest to read documents got $U = 0$. Thus, for CLEF 2015 this strategy can be defined as $F(R,U) = R * U / 3$. This way, even a document that is highly relevant for a query can still get a label **0** if it this document was assessed as very hard to read. Similarly, for CLEF 2016, the easiest-to-read documents were assigned a score $U = 0$ and the hardest-to-read documents got $U = 100$. Thus, for CLEF 2016, this strategy can be defined as $F(R,U) = R * (100 - U) / 100$. Another way to see this strategy is that a penalty is proportionally assigned to documents according to how bad is their understandability score. For example, a document with understandability 0 receives no penalty, as 0 is the easiest level of understanding in the 2016 dataset, while another with understandability 50 received a 50% penalty, meaning that its relevance score is halved.

The third and fourth strategies assume the existence of an understandability threshold that divides the documents into two groups: the easy-to-read and the hard-to-read documents. This understandability threshold was set to $U = 2$ for CLEF 2015 and documents with $U \geq 2$ are considered easy-to-read. For CLEF 2016, we use $U = 40$ and documents in which $U \leq 40$ are considered easy-to-read³.

The third labeling strategy, named *threshold*, assigns label 0 to hard-to-read documents and the topical relevance label to easy-to-read documents. Formally, it can be defined as $F(R,U) = \begin{cases} R & \text{if } U \geq 2 \\ 0 & \text{otherwise} \end{cases}$ for CLEF 2015 and as $F(R,U) = \begin{cases} R & \text{if } U \leq 40 \\ 0 & \text{otherwise} \end{cases}$ for CLEF 2016.

³Similarly to Section 7.4, this threshold was arbitrary chosen based on the semantic and colors used in the Relevance! assessment tool in CLEF 2016 (see Figure A.3). The green part of the slider, which is likely understood as “easy-to-read documents” goes from understandability scores 0 to 40, thus we re-used this threshold here.

Lastly, the fourth labeling strategy, named *boost*, avoids assigning zeros to hard-to-read documents. Instead, it boosts the label of easy-to-understand documents. In our experiments, we double the value of label of easy-to-understand documents. Formally, we define as $F(R, U) = \begin{cases} 2 \times R & \text{if } U \geq 2 \\ R & \text{otherwise} \end{cases}$ for CLEF 2015 and as $F(R, U) = \begin{cases} 2 \times R & \text{if } U \leq 40 \\ R & \text{otherwise} \end{cases}$ for CLEF 2016.

Feature Set

The last part of a typical learning-to-rank experiment consists in defining the representation of the documents. We define three options for the feature set to represent documents: the use of information retrieval (IR) features; the use of understandability features; and the combination of IR and understandability features.

As IR features, we explore different retrieval models: eight retrieval models implemented in the Terrier toolkit were used - BM25, PL2, DirichletLM, Hiemstra_LM, LemurTF_IDF, TF_IDF,DFRee and D1 - this is a representative subset of all families of retrieval models⁴. Specifically, we devised 24 IR features using the Terrier framework. The score of the eight retrieval models listed above was extracted from a multi-field index composed of title, body and whole document. Although simple, this is a typical learning-to-rank setting. As understandability features, we explore the understandability estimators investigated in Section 6.1.2. Considering all preprocessing variations and features listed in Table 6.1, a total of 1,082 features were used for understandability.

Experiment Settings

Each of the four labeling strategies can be combined with one option of a feature set. This results in twelve possible experimental settings. In order to keep our experiments concise, we experiment with only five combinations of labeling strategy and feature set.

The experimental setting named *LTR 1* is the most common learning-to-rank setting in which only topical relevance is considered for labeling strategy and only IR features are extracted from the documents. In *LTR 2*, while using only the topical relevance as labeling strategy, we expand the document representation to include understandability features as well as IR features. *LTR 3*, *LTR 4* and *LTR 5* use both understandability and IR features, but different labeling strategies. *LTR 2* uses the *proportional* strategy, while *LTR 3* uses the *threshold* strategy and *LTR 4* uses the *boost* strategy.

These combinations are listed in Table 8.1, with R being the relevance of documents and U their understandability estimation.

⁴Details on each method can be found online at http://terrier.org/docs/v4.2/configure_retrieval.html

Table 8.1: Learning to rank settings.

Name	Features	Labeling Strategy	
		CLEF 2015	CLEF 2016
LTR 1	Only IR features	$F(R,U) = R$	$F(R,U) = R$
LTR 2	IR and Understandability features	$F(R,U) = R$	$F(R,U) = R$
LTR 3	IR and Understandability features	$F(R,U) = R \times U/3$.	$F(R,U) = R \times (100 - U)/100$
LTR 4	IR and Understandability features	$F(R,U) = \begin{cases} R & \text{if } U \geq 2 \\ 0 & \text{otherwise} \end{cases}$	$F(R,U) = \begin{cases} R & \text{if } U \leq 40 \\ 0 & \text{otherwise} \end{cases}$
LTR 5	IR and Understandability features	$F(R,U) = \begin{cases} 2 \times R & \text{if } U \geq 2 \\ R & \text{otherwise} \end{cases}$	$F(R,U) = \begin{cases} 2 \times R & \text{if } U \leq 40 \\ R & \text{otherwise} \end{cases}$

8.2 Evaluation Measures

Apart from the rank-biased precision (RBP) measure [134], for the retrieval experiments in Section 8.3, we used evaluation measures that rely on both (topical) relevance and understandability. From the UBIRE framework [224, 223], we used $uRBP$, and from the MM framework, we used both RBP_u and MM_{RBP} . Both frameworks and their measures were described and studied in Chapter 7. For all measures, we set the persistence parameter of RBP to 0.80, as done in both CLEF 2015 and 2016 [77, 106].

Shallow pools were used in both CLEF collections, i.e., only a limited number of documents were selected to be assessed for relevance. Because of that, we focused on evaluating the top 10 search results for all measures (e.g., $RBP_r@10$). To cope with unassessed documents, we also report the *condensed* version of the evaluation measures, represented with a superscripted “*” (e.g., RBP_r^* or MM_{RBP}^*). The condensed approach, proposed by Sakai as a way to deal with unassessed documents [170], is nothing else than the corresponding measure calculated by ignoring unassessed documents. Finally, we recorded the number of unassessed documents in the top 10 ($una@10$) and the RBP residuals for RBP-based measures (e.g., $RBP_r@10$, $uRBP@10$ or MM_{RBP}) [134]. The residuals represent the error bound of the RBP measures, i.e., the maximum contribution to the RBP score if the unassessed documents were assessed as relevant documents.

Our special care with unassessed documents aims to minimize pool bias since the pools built in CLEF were of limited size, and the investigated methods retrieved a substantial number of unassessed documents. Pool bias refers to the possible bias in the evaluation towards systems that have contributed documents to the assessment pool: these erroneously receive higher evaluation scores compared with systems that did not contribute to the pool (i.e., that were not sampled to create the set of documents to be judged for relevance).

8.3 Evaluating Understandability Aware Retrieval

Results for our experiments are shown in Table 8.2 for CLEF 2015 and Table 8.3 for CLEF 2016. The effectiveness of the top two submissions to CLEF 2015/2016 and the

BM25 baseline are reported at indices 1–3 of Tables 8.2 and 8.3. Statistically significant differences compared with the best run in CLEF 2015, ECNU, and CLEF 2016, GUIR, are indicated with \diamond ; differences between an original run (indices 1–3) and its modifications are indicated with \dagger . Statistically significant differences using a paired two-tail t-test with Bonferroni correction [69] were calculated between the result of an experiment and (1) the BM25 baseline and (2) the best method of each collection (ECNU for CLEF 2015 and GUIR for CLEF 2016). For the Bonferroni correction, as 294 t-tests were conducted per collection (14 experiments [3 using the best readability formula + 3 using the best machine learning method + 3 using the combination of the best readability formula and machine learning method + 5 learning-to-rank approaches] \times 7 evaluation metrics [RBP_r + RBP_u + MM_{RBP} + $uRBP$ + RBP_r^* + RBP_u^* + MM_{RBP}^*] \times 3 systems [top 2 from each collection + BM 25] = 294), the actual p-value used was $.05/294 = .00017$, i.e., differences were considered statistically significant only if $p < .00017$.

Next, we report the results of each sub-experiment: *Simple re-ranking* (indices 4–21), *Fusion Experiments* (indices 22–30), *Learning-to-rank* (indices 31–35).

8.3.1 Simple Re-ranking

Indices 4–12 of Tables 8.2 and 8.3 report the results of re-ranking methods applied to the runs listed at indices 1–3. Re-ranking was applied based on the SMOG score for CLEF 2015 (preprocessing made with *JusText-ForcePeriod*) and Dale-Chall Index (DCI) score for CLEF 2016 (preprocessing made with *JusText-ForcePeriod*) of each document. Readability formulas and preprocessing combination were chosen based on their Pearson correlation to human assessments, previously examined in Table 6.6, but any other understandability estimator studied in Chapter 6 could be used instead. We found that the topical relevance of the re-ranked runs (as measured by RBP and RBP_r^*) significantly decreased, compared with the original runs: e.g., in CLEF 2016, re-ranking the top 15 search results using DCI made RBP_r decreasing from 25.28 to 23.22 for the BM25 system. However, these re-ranked results were significantly more understandable: for the previous example, RBP_u passed from 42.08 to 47.09. As described in Chapter 7, note the limitation of $uRBP$ to reveal such an effect, as both topical relevance and understandability are tied together in one single score.

In the experiments, we also studied the influence of the numbers of documents considered for re-ranking (cut-off). Indices 4–6 refer to re-ranking only the top $k = 15$ documents from the original runs; 7–9 refer to the first $k = 20$; and 10–12 to the first $k = 50$. The results show that the more documents are considered for re-ranking, the more degradation in RBP_r effectiveness. Similarly to RBP_r , $uRBP$ and RBP_u , metrics that take into account understandability, degraded when k increased.

It is important to note that with the increase in the number of documents considered for re-ranking, there is an increase in the number of unassessed documents being considered by the evaluation measures. Both the RBP residuals and the column *Una* quantify the effect that unassessed documents have on evaluation. In particular, when we exclude

Table 8.2: Results obtained by integrating understandability estimations within retrieval methods on CLEF 2015. Baseline runs are reported at table indices 1–3 (the index column is labeled Idx). Re-ranking experiments are reported at indices 4–21. Fusion experiments are reported at indices 22–30. Learning to rank experiments are reported at indices 31–35. All measures were calculated up to rank $n = 10$. The highest result of each set of experiments is reported in boldface for each measure. Statistically significant differences compared to ECNU are indicated with \diamond , while differences between an original run (indices 1-3) and its modifications are indicated with \dagger (paired, two-tail t-test with Bonferroni correction, $p < .00017$ (.05 / 294)).

Idx	Rerank	Run	CLEF 2015 Measures			New Measures to Evaluate the Understandability in Retrieval								
			RBP_r	Res.	uRBP	Res.	RBP_u	Res.	MM_{RBP}	Res.	Una@10	RBP_r^*	RBP_u^*	MM_{RBP}^*
1	No Rerank	ECNU [183] (1st)	51.57	8.95	50.51	8.95	59.55	10.09	46.22	8.62	0.00	51.57	59.55	46.22
2		KISTI [138] (2nd)	36.72	8.06	35.92	7.32	64.50	11.54	37.56	7.89	0.03	37.07	65.31	37.96
3		BM25 Baseline	31.20 \diamond	8.76	30.51 \diamond	7.65	67.60	12.20	35.75	8.76	0.03	31.57 \diamond	68.94	36.42
4	SMOG Top 15	Based on ECNU	38.16 \diamond	20.09	37.36 \diamond	8.95	55.11	22.14	37.59 \diamond	18.98	0.14	45.45	62.58	43.65
5		Based on KISTI	31.28 \diamond	10.21	30.59 \diamond	7.48	67.05	13.95	34.23	10.05	0.05	33.12 \diamond	69.10	36.02 \diamond
6	SMOG Top 20	Based on BM25	24.39 \diamond	11.46	23.83 \diamond	7.16	67.69	17.66	28.97 \diamond	11.46	0.10	26.28 \diamond	72.82	31.53
7		Based on ECNU	34.88 \diamond	25.96	34.13 \diamond	8.95	54.15	27.91	35.34 \diamond	25.02	0.20	44.87	66.20	44.88
8	SMOG Top 50	Based on KISTI	28.01 \diamond	11.05	27.39 \diamond	7.32	66.61	15.44	30.93 \diamond	10.82	0.08	30.87 \diamond	69.99	33.81
9		Based on BM25	22.83 \diamond	13.48	22.31 \diamond	7.32	63.61	22.44	26.52 \diamond	13.48	0.15	26.22 \diamond	72.72 \diamond	30.87 \diamond
10	SMOG Top 15	Based on ECNU	21.83 \diamond	36.63	21.34 \diamond	8.95	41.71 \diamond	46.89	23.65 \diamond	36.63	0.45	39.88 \diamond	71.61	43.86
11		Based on KISTI	21.00 \diamond	18.67	20.53 \diamond	7.32	59.72	26.67	24.88 \diamond	18.67	0.23	27.20 \diamond	72.47 \diamond	31.90
12	SMOG Top 20	Based on BM25	15.20 \diamond	18.73	14.86 \diamond	6.51	49.87 \dagger	36.41	17.79 \dagger	18.73	0.32	21.22 \diamond	73.17	25.78 \diamond
13		Based on ECNU	39.77 \diamond	21.59	38.90 \diamond	9.11	54.84	24.22	38.35 \diamond	20.72	0.15	47.00	64.28	45.22
14	XGB Top 15	Based on KISTI	31.36 \diamond	8.95	30.63 \diamond	6.83	68.63	13.63	33.78	8.78	0.05	32.80	70.87	35.24
15		Based on BM25	23.38 \diamond	11.36	22.84 \diamond	7.16	66.79	19.43	27.60 \diamond	11.36	0.10	26.35 \diamond	73.48	31.35
16	XGB Top 20	Based on ECNU	34.91 \diamond	27.19	34.12 \diamond	9.11	52.69	30.19	34.89 \diamond	25.95	0.25	46.71	67.26	45.81
17		Based on KISTI	29.04 \diamond	10.10	28.35 \diamond	6.51	69.10	15.86	32.00	9.94	0.07	31.82	73.19	34.84
18	XGB Top 50	Based on BM25	21.83 \diamond	12.88	21.32 \diamond	6.51	64.16	25.29	25.47 \dagger	12.88	0.16	26.56 \diamond	76.26 \diamond	31.45
19		Based on ECNU	22.75 \diamond	33.65	22.22 \diamond	8.79	43.36 \diamond	49.18	24.90 \diamond	33.65	0.45	40.69 \diamond	74.39 \diamond	44.65
20	XGB Top 15	Based on KISTI	18.69 \diamond	17.21	18.25 \diamond	6.67	56.87	33.37	21.82 \dagger	17.21	0.29	27.69 \diamond	76.44 \diamond	32.53
21		Based on BM25	17.47 \diamond	20.95	17.00 \diamond	6.02	47.21 \dagger	43.81	19.22 \dagger	20.98	0.41	27.09 \diamond	79.77 \diamond	31.35
22	RRF (XGB & Orig.) Top 15	Based on ECNU	47.81 \diamond	12.85	46.78	8.95	60.04	15.04	44.69	12.34	0.10	50.09	62.98	46.85
23		Based on KISTI	33.78	7.71	33.02	6.83	68.57	12.02	35.72	7.55	0.03	34.34	69.67	36.26
24	RRF (XGB & Orig.) Top 20	Based on BM25	26.85 \diamond	12.13	26.23 \diamond	7.48	66.64	15.98	31.52	12.13	0.07	28.35 \diamond	70.65	33.70
25		Based on ECNU	46.49 \diamond	15.16	45.48 \diamond	9.11	59.95	17.12	43.69	14.54	0.12	49.95	64.30	47.01
26	RRF (XGB & Orig.) Top 50	Based on KISTI	32.72	8.45	31.97	7.00	69.06	12.77	35.14	8.28	0.04	33.83	70.83	36.29
27		Based on BM25	25.64 \diamond	12.51	25.05 \diamond	7.16	66.55	17.31	30.14	12.51	0.09	27.64 \diamond	71.79	32.96
28	RRF (XGB & Orig.) Top 15	Based on ECNU	38.97 \diamond	21.40	38.08 \diamond	8.79	57.37	25.44	39.54	19.83	0.24	47.00	67.95	47.12
29		Based on KISTI	27.78 \diamond	11.24	27.13 \diamond	6.83	67.83	16.75	31.07 \diamond	11.24	0.09	31.33 \diamond	73.06 \dagger	34.85
30	Based on BM25	19.28 \diamond	17.11	18.86 \diamond	7.00	57.40	27.02	22.78 \dagger	17.11	0.19	25.06 \diamond	71.56	30.26 \diamond	
31	XGB LeToR	LTR 1 on BM25	24.86 \diamond	17.39	24.32 \diamond	7.81	55.60 \dagger	24.11	28.89 \diamond	17.39	0.22	29.67 \diamond	66.41	34.76
32		LTR 2 on BM25	30.72 \diamond	21.25	30.08 \diamond	8.46	48.87 \diamond	28.82	31.76 \diamond	18.99	0.26	37.09 \diamond	61.89	39.17
33		LTR 3 on BM25	28.92 \diamond	24.35	28.32 \diamond	8.46	49.02 \diamond	32.11	30.14	23.83	0.31	37.32 \diamond	63.86	39.84
34		LTR 4 on BM25	25.65 \diamond	25.72	25.09 \diamond	8.30	49.00 \dagger	33.39	27.45 \diamond	24.40	0.33	35.82 \diamond	66.14	38.21
35		LTR 5 on BM25	30.21 \diamond	20.79	29.59 \diamond	8.62	48.47 \diamond	27.88	30.95 \diamond	19.99	0.25	37.11 \diamond	61.25	39.15

the unassessed document from the evaluation and look at the understandability of the retrieved documents, i.e., RBP_u^* , we noticed that increasing k increases the amount of easy-to-read documents retrieved. The topical relevance of these documents, measured by RBP_r^* , however, decreases when k increases.

Indices 13–21 refer to using the XGB regressor trained with all features listed in Table 6.1 to estimate understandability (as described in Section 6.5). Similarly to when using SMOG or DCI, as the cut-off increased, e.g., from $k = 15$ to $k = 50$, the documents returned were more understandable but less relevant (in particular when the condensed measure RBP_u^* is taken into account). For the same cut-off value, e.g., $k = 15$, the machine learning method used for estimating understandability consistently yielded more understandable results than SMOG or DCI (i.e., higher RBP_u and RBP_u^*).

Table 8.3: Results obtained by integrating understandability estimations within retrieval methods on CLEF 2016. Baseline runs are reported at table indices 1–3 (the index column is labeled Idx). Re-ranking experiments are reported at indices 4–21. Fusion experiments are reported at indices 22–30. Learning to rank experiments are reported at indices 31–35. All measures were calculated up to rank $n = 10$. The highest result of each set of experiments is reported in boldface for each measure. Statistically significant differences compared to GUIR are indicated with \diamond , while differences between an original run (indices 1-3) and its modifications are indicated with \dagger (paired, two-tail t-test with Bonferroni correction, $p < .00017$ (.05 / 294)).

Idx	Rerank	Run	CLEF 2016 Measures				New Measures to Evaluate the Understandability in Retrieval							
			RBP_r	Res.	uRBP	Res.	RBP_u	Res.	MM_{RBP}	Res.	Una@10	RBP_r^*	RBP_u^*	MM_{RBP}^*
1	No Rerank	GUIR [182] (1st)	28.11	7.65	18.12	7.19	45.69	8.86	25.61	6.50	0.01	28.29	46.03	25.79
2		ECNU [184] (2nd)	27.70	7.37	17.55	7.34	43.89	8.66	25.35	6.26	0.01	27.77	44.18	25.48
3		BM25 Baseline	25.28	8.24	16.05 \diamond	6.94	42.08 \diamond	10.97	22.97 \diamond	7.19	0.06	26.01	43.89	23.93
4	Dale-Chall Top 15	Based on GUIR	24.29 \dagger	8.89	16.62	7.27	48.98 \dagger	11.05	24.76	7.61	0.03	24.83 \dagger	50.56 \dagger	25.38
5		Based on ECNU	24.38 \dagger	8.19	16.45	7.16	48.96 \dagger	10.10	24.68	6.86	0.03	24.88 \dagger	49.99 \dagger	25.24
6	Dale-Chall Top 20	Based on BM25	23.22 \dagger	8.78	15.85 \diamond	6.94	47.09 \dagger	11.83	24.01	7.42	0.07	24.04 \diamond	48.60 \dagger	24.82
7		Based on GUIR	21.86 \dagger	10.04	15.15 \dagger	7.05	48.54	13.13	22.94	8.68	0.06	23.23 \dagger	51.78 \dagger	24.44
8	Dale-Chall Top 50	Based on ECNU	22.61 \dagger	9.24	15.41 \dagger	6.94	48.45 \dagger	12.31	23.43	8.06	0.05	23.62 \dagger	50.86 \dagger	24.58
9		Based on BM25	21.58 \dagger	9.51	14.83 \diamond	7.02	46.99 \dagger	13.00	22.89	8.06	0.09	22.93 \dagger	49.55 \dagger	24.26
10	Dale-Chall Top 15	Based on GUIR	16.07 \dagger	15.31	11.50 \dagger	6.80	41.41	24.34	17.95 \dagger	14.48	0.23	20.80 \dagger	53.07 \dagger	23.14
11		Based on ECNU	16.72 \dagger	17.64	11.67 \dagger	7.27	40.46 \diamond	24.18	18.13 \dagger	15.87	0.24	21.38 \dagger	52.10 \dagger	23.41
12	XGB Top 15	Based on BM25	15.06 \dagger	15.35	10.55 \dagger	6.62	40.03 \diamond	23.88	16.55 \dagger	13.83	0.24	19.42 \dagger	51.69 \dagger	21.59 \dagger
13		Based on GUIR	25.16 \dagger	8.09	17.27 \dagger	7.12	50.96 \dagger	10.11	25.16	6.89	0.02	25.61 \dagger	52.00 \dagger	25.68
14	XGB Top 20	Based on ECNU	24.18 \dagger	7.69	16.54	7.09	50.00 \dagger	9.91	24.56	6.65	0.02	24.56 \dagger	50.74 \dagger	25.01
15		Based on BM25	22.79 \diamond	7.94	15.46	6.84	47.88 \dagger	11.65	23.10	7.08	0.07	23.46 \diamond	49.12 \dagger	23.80
16	XGB Top 50	Based on GUIR	22.38 \dagger	9.49	15.61 \dagger	7.05	50.45 \dagger	12.08	23.30	8.16	0.05	23.62 \dagger	52.98 \dagger	24.68
17		Based on ECNU	22.95 \dagger	8.82	15.95	7.02	50.42 \dagger	11.70	23.97	7.56	0.04	23.68 \dagger	52.15 \dagger	24.73
18	XGB Top 15	Based on BM25	20.77 \dagger	9.26	14.51 \dagger	6.76	47.85 \dagger	13.24	22.04	8.24	0.08	22.10 \dagger	50.28 \dagger	23.32
19		Based on GUIR	16.65 \dagger	15.73	12.39 \dagger	6.84	43.49	23.63	18.70 \dagger	13.74	0.22	21.13 \dagger	55.07 \dagger	23.58
20	XGB Top 50	Based on ECNU	16.19 \dagger	17.01	11.82 \dagger	7.27	43.05	24.75	18.27 \dagger	14.41	0.24	20.16 \dagger	54.70 \dagger	22.96
21		Based on BM25	15.55 \dagger	14.92	11.44 \dagger	6.59	42.29	23.21	17.60 \dagger	13.00	0.25	19.58 \dagger	53.94 \dagger	22.19
22	RRF (XGB & Orig.) Top 15	Based on GUIR	27.23	7.76	18.31	7.23	49.69 \dagger	9.18	26.46	6.62	0.01	27.46	50.07 \dagger	26.69
23		Based on ECNU	26.60	7.41	17.81	7.19	48.67 \dagger	8.80	26.02	6.09	0.01	26.76	49.10 \dagger	26.27
24	RRF (XGB & Orig.) Top 20	Based on BM25	24.64 \diamond	8.13	16.50	6.98	46.57 \dagger	11.02	24.14	7.11	0.06	25.39 \diamond	48.25 \dagger	25.04
25		Based on GUIR	26.21 \dagger	7.96	17.73	7.19	50.29 \dagger	9.58	25.89	6.73	0.03	26.53 \dagger	50.98 \dagger	26.25
26	RRF (XGB & Orig.) Top 50	Based on ECNU	26.15	7.64	17.69	7.09	49.70 \dagger	9.28	26.07	6.39	0.02	26.38	50.32 \dagger	26.35
27		Based on BM25	24.10 \diamond	8.26	16.34	6.91	47.62 \dagger	11.34	24.10	7.37	0.06	24.92 \diamond	49.37 \dagger	25.03
28	RRF (XGB & Orig.) Top 15	Based on GUIR	24.09 \dagger	9.44	16.85	7.02	50.55 \dagger	11.76	24.76	8.01	0.07	25.08 \dagger	52.84 \dagger	25.84
29		Based on ECNU	24.17 \dagger	8.67	16.75	7.12	50.63 \dagger	11.66	25.00	7.61	0.07	24.90 \dagger	52.56 \dagger	25.84
30	Based on BM25	22.32 \dagger	8.82	15.51	6.73	48.78 \dagger	12.84	23.14	7.78	0.10	23.45 \diamond	51.85 \dagger	24.53	
31	XGB LeToR	LTR 1 on BM25	20.42 \dagger	17.61	13.00 \diamond	7.41	32.17 \dagger	24.61	18.39 \dagger	14.41	0.28	25.25	43.19	23.83
32		LTR 2 on BM25	25.16 \diamond	19.95	15.85 \diamond	8.09	34.31 \dagger	24.95	21.69 \dagger	17.43	0.28	30.60	45.37	27.57
33		LTR 3 on BM25	26.35	20.48	15.88 \diamond	8.16	34.73 \dagger	24.69	21.81	17.41	0.22	32.25	45.44	28.22 \dagger
34		LTR 4 on BM25	16.16 \dagger	19.48	10.76 \dagger	7.27	36.75 \dagger	28.51	16.77 \dagger	17.80	0.29	22.20 \dagger	50.06 \dagger	23.32
35		LTR 5 on BM25	26.76	20.48	16.19 \diamond	8.34	35.26 \dagger	24.13	22.96	17.59	0.22	32.60	45.87	29.20 \dagger

8.3.2 Rank Fusion

Next, we report the results of automatically combining topical relevance and understandability through rank fusion (indices 22 to 30). We used the XGB method for estimating understandability, as it was the one yielding highest effectiveness for the re-ranking method. Runs were thus produced by fusing the re-ranking with XGB and the original run.

As for re-ranking, also for the rank fusion approaches, we found that, in general, higher cut-offs were associated to higher effectiveness in terms of understandability measures on the one hand, but higher losses in terms of relevance-oriented measures on the other. Overall, results obtained with rank fusion were superior to those obtained with re-ranking

only, though most differences were not statistically significant.

8.3.3 Learning-to-Rank

Finally, we analyze the results obtained by the learning-to-rank methods (indices 31-35). Unlike with the previous methods, it was not necessary to impose a rank cut-off on the learning-to-rank approach as the document representation used here always includes information retrieval features. Consequentially, it considers the signals regarding topical relevance that are encoded in the IR features. Learning-to-rank was only applied to the BM25 baseline, as we had no access to the IR features for the runs submitted at CLEF (i.e., GUIR and ECNU for CLEF 2016).

When considering RBP_r and $uRBP$, learning-to-rank exhibited effectiveness that was significantly inferior to that of the GUIR, KISTI and ECNU baseline runs, though higher than those for the BM25 baseline (for some configurations). The examination of the RBP residuals (and the number of unassessed documents) revealed that this might have been because measures were affected by the large number of unassessed documents retrieved in the top 10 ranks. For example, the RBP_r residual for learning-to-rank methods was about double that of the baselines or other approaches. In fact, among the documents retrieved in the top 10 results by learning-to-rank, there were at least 22% that were unassessed, compared with an average of 3% for the other methods.

We thus should carefully account for unassessed documents by considering the residuals of RBP measures, as well as the condensed measures that ignore unassessed documents (i.e., RBP_r^* , RBP_u^* and MM_{RBP}^*). When this was done, we observed that learning-to-rank methods overall provided substantial gains over the original runs and other methods (when considering RBP_r^* , RBP_u^* and MM_{RBP}^*), or large potential gains over these methods (when considering the residuals). Next, we analyze these results in more detail.

No improvements over the baselines were found for LTR 1 (index 31) in both Tables 8.2 and 8.3. The high residuals for RBP_r were not matched by other residuals or by considering only assessed documents. Remember LTR 1 was a simple method that used only IR features and was trained only on topical relevance.

Compared with LTR 1, LTR 2 (index 32) included the understandability features listed in Table 6.1. This inclusion was as beneficial to the understandability measures as to the relevance measures, with RBP_r^* , RBP_u^* and MM_{RBP}^* all showing gains over the baselines. LTR 3 obtained similar MM_{RBP}^* values, though with higher effectiveness for relevance measures (RBP_r^*) than for understandability (RBP_u^*). In particular, for the experiments with CLEF 2015, LTR 3 reached the highest MM_{RBP}^* of 39.84, an improvement of 3.4 percentile points over the baseline.

LTR 4 and 5 were devised based on a set understandability threshold: $U \geq 2$ for CLEF 2015 and $U \leq 40$ for CLEF 2016. Although LTR 4 took into consideration only documents that are easy-to-read, LTR 5 considered all documents, but boosted the relevance score of easy-to-read documents. LTR 4 reached the highest understandability

score for the learning-to-rank approaches for CLEF 2016 ($RBP_u^* = 50.06$), but it failed to retrieve a substantial number of relevant documents ($RBP_r^* = 22.20$). In turn, LTR 5 reached the highest understandability-relevance trade-off for CLEF 2016 ($MM_{RBP}^* = 29.20$) and the third for CLEF 2015 ($MM_{RBP}^* = 39.15$, just 0.69 smaller than LTR 3 and 0.02 smaller than LTR 2). In particular for CLEF 2016, compared to the BM25 baseline (on which it was based), LTR 5 largely increased both relevance (RBP_r^* went from 26.01 to 32.60 – an increase of 6.59 percentile points) and understandability (RBP_u^* went from 43.89 to 45.87 – an increase of 1.97 percentile points). Note that LTR 5 was also better than GUIR, the best run submitted to CLEF 2016, for both RBP_r^* (increase of 4.3 percentile points) and MM_{RBP}^* (increase of 3.41 percentile points).

8.4 Summary

This chapter aimed to demonstrate approaches to integrate understandability into retrieval systems empirically.

A simple method such as re-ranking the retrieved results according to an understandability estimator, such as DCI or SMOG, was our first step towards increasing the understandability of results without losing retrieval effectiveness. We showed that as we increased the number of documents to be re-ranked, we traded topical relevance for understandability. In our experiments, the use of machine learning methods was shown superior to the use of readability formulas as understandability estimators for re-ranking.

We then showed how it is possible to find a balance between topical relevance and understandability automatically. For that, we proposed the use of ranking fusion methods. In particular, we used the reciprocal ranking fusion method as it is based only on the ranking position of documents and not on their scores.

Finally, our experiments with learning-to-rank methods suggested that models can be specially trained to promote more understandable search results, while still providing an effective trade-off with topical relevance.

Part V
Closure

Discussion and Conclusion

We reach the end of this work by discussing, in Section 9.1, the research questions introduced in Chapter 1 and by listing, in Section 9.2, the limitations of this work and future directions.

9.1 Discussion of Research Questions

The research questions of Parts II, III and IV are respectively reviewed and answered, and conclusions are drawn from the answers, in the following sections.

9.1.1 User Search Behavior in the Health Domain

Part II of this thesis started with a detailed study of health information search behavior through query logs. We analyzed what users search for and how users search in the health domain, comparing our results to other search log based studies. For that, we used MetaMap to annotate the queries. As only a few studies evaluated MetaMap for mapping short queries to health concepts, we performed a full evaluation with a part of our dataset. In Part II, we also built a classifier to distinguish between health consumers and health experts based on their search queries. Search engines, such as Google, HON or PubMed, could make use of the classifier proposed here to personalize the results for their users, e.g., promoting content that, while topically relevant, has the highest level of understandability.

The research questions of Part II were:

1. **Chapter 3:** *What and how do consumers and health professionals search in the health domain?* We found that the way experts and health consumers search is different. On average in our experiments with search logs, health experts issued

longer queries both regarding characters and words per query (with a small effect as measured by the Cohen's d value). We also found that search sessions of health experts were longer both in terms of the mean number of queries per session as well as the amount of time spent per session. These longer sessions might indicate that health experts are more persistent, i.e., they interact more often with the search engine reformulating their queries to find relevant information, or might indicate that they often struggle to find relevant information for their queries.

A significant difference concerning what is published in the literature was noticed when analyzing what the users are searching for. Our analysis showed that diseases were the focus of the largest number of sessions (20.9%–58.2%), as opposed to symptoms (63.8% in [27]). This difference is mainly due to the criteria used by other researchers to extract medical queries from the search logs of a general purpose search engine, which skewed the results toward symptoms. This result suggests that the occurrence of cyberchondria [208] is less prevalent, especially on domain-specific medical search engines.

2. **Chapter 3:** *How suitable is an automatic health text annotator, such as UMLS MetaMap, to analyze and annotate short Web queries?* Our study relied on the accuracy of MetaMap to identify the intent of the searchers. As MetaMap was designed for annotation of documents and not queries, we evaluated its performance for short queries. Using an existing dataset of 10,000 manually annotated queries [136], we evaluated MetaMap on two of the categories used in our experiments: cause and remedy. The category symptom was not evaluated as it is not included in the dataset used. We found that MetaMap can annotate the cause category with an F1 of 78% and the Remedy category with an F1 of 72%. While these values are not directly comparable to other results published, they correspond to the level of accuracy measured for related tasks: MetaMap was shown to map disease concepts in queries with an F1 of 70% [137], and a mapping into five classes in [140] on 1000 queries was also done with an F1 of 70%. Most importantly, the inter-annotator agreement for the manual annotation of the query corpus by Névéal et al. [137] was 73%. This demonstrates that the results obtained by annotating the queries by MetaMap are at the same level as those obtained by manual annotation, implying that the MetaMap annotations are sufficiently accurate for this study.
3. **Chapter 4:** *Can we automatically infer user health expertise through user search behavior?* The feature set devised from the query-log analysis made in Chapter 3 was directly applied in Chapter 4 showing that it is possible to distinguish between health experts and consumers based on search behavior characteristics.
4. **Chapter 4:** *What are the most useful features to infer user health expertise through search behavior automatically?* We grouped the features devised in this work into five distinct sets: user behavior features, part-of-speech tagging features, semantic features, UMLS features, and consumer health vocabulary features. Features from the three last feature sets are all domain-specific ones. Our experiments showed

that while classifiers exploring each feature set alone are not highly effective in distinguishing between health consumers and health experts, a classifier using all features combined can be effective for the task. Our feature importance analysis concluded that the average number of characters (behavior feature) was the most influential feature. Domain-specific features, such as the features based on health vocabularies, were important for the task. Two crucial domain-specific features were the average number of UMLS sources and the average CHV scores of the terms in a query. Both features potentially measure how common a term is: the latter directly measures how easy to understand a term is, while the former indirectly measures how popular a term is, as a term that occurs in multiple health vocabularies is potentially a common one (i.e., *heart attack* is in 26 vocabulary sources while the rare disease *Morgellons* appears only in two).

9.1.2 Understandability Estimation of Web Documents

Part III analyzed methods to estimate the understandability of Web documents, with special care for the role that preprocessing and sentence boundary identification choices have on the estimation of understandability of such Web documents.

The research questions of Part III were:

5. **Chapter 5:** *What is the effect of preprocessing pipelines on readability formulas when estimating the understandability of Web documents?* Our experimental results indicate that the choice of preprocessing pipelines has a significant impact on the estimation of understandability scores of Web documents. This impacts on the order relations among documents that can be obtained from readability formulas. Our findings suggest that more attention should be given to the choice of preprocessing settings when measuring the understandability of Web documents. In particular, advanced boilerplate removal tools, such as *Boilerpipe* and *JusText*, should be preferred over simpler HTML cleansing methods that only get rid of HTML tags, as they provide more stable results across different preprocessing settings.
6. **Chapter 5:** *Among the readability formulas, what are the most and the least robust ones? Which readability formula should we use?* Our experiments revealed that the Coleman-Liau Index (CLI) was the most robust readability formula across choices of preprocessing tools and sentence boundary identification strategies. On the other extreme, the Automated Readability Index (ARI) was the least robust one with the choice of preprocessing methods deeply impacting its score estimations. The results from the correlation with human assessments introduced in Chapter 6 did not find a unique best readability formula to use, however it identified that Automated Readability Index (ARI) should be avoided, as in both CLEF 2015 and CLEF 2016 it was the least correlated formula with human assessments and significantly worse than the best formula for each collection.

7. **Chapter 6:** *What are the best understandability estimators among the various studied?*

We experimented with a large number of understandability estimators to infer the understandability level of documents from both CLEF 2015 and CLEF 2016 collections, as human assessments for the understandability of relevant documents in these collections were provided. We divided our experiments into two parts: in the first part, we focus on the correlation of each estimator with the human assessments. In the second part, we combine all these *weak* estimators to build *strong* machine learning models that predict the understandability of documents. In the first part of our experiments, we found that the understandability estimators that better correlated with human assessments of understandability were based on machine learning techniques trained on the auxiliary data extracted from Reddit, Wikipedia and PubMed. These machine learning methods correlated better with human assessments of understandability than other groups of methods including the traditional readability formulas, which are often used by health information providers on the Web as references to measure whether or not a text is too hard to understand. In the second part of our experiments, each estimator becomes a feature for a machine learning model. We grouped the devised features into ten semantically related groups and we trained machine learning models with various settings. In particular, to measure how strong each group of features was, we trained models using only one single group at the time and compared their performance to models trained with all features. The best results were achieved by classifiers and regressors trained with features from the natural language group, which included features such as the distribution of part-of-speech classes and the percentage of stopwords in a text.

8. **Chapter 6:** *How do preprocessing pipelines affect methods of understandability estimation?* Our experiments demonstrated that the preprocessing pipelines affect the various understandability estimation methods. We first evaluated the preprocessing pipelines only on the readability formulas and then on all the methods to estimate understandability. In both experiments, we found that the highest correlations with human assessments were achieved by preprocessing documents with the *JusText* approach, the differences were significantly higher than using the *Boilerpipe* or the *Naïve* preprocessing. No significant differences were found when comparing *ForcePeriod* to *DoNotForcePeriod*. We empirically learned that the *Naïve* method should be avoided as in both experiments it was the preprocessing strategy that yielded the lowest correlation with human assessments, being significantly worse than *Boilerpipe* when preprocessing documents to apply readability formulas.

9.1.3 Understandability Integrated into Search Engines

Part IV of this thesis focused on the evaluation of information retrieval system when more than one relevance dimension is taken into account. For that, we studied frameworks that consider other relevance dimensions, such as understandability, and, with such

frameworks, we evaluated various strategies to integrate understandability estimation into search engines.

The research questions of Part IV were:

9. **Chapter 7:** *How can we incorporate other relevance dimensions (e.g., understandability) into existing system evaluation metrics?* UBIRE, the understandability-biased IR evaluation framework, builds upon the gain-discount framework of evaluation measures used in information retrieval, such as nDCG or RBP. It extends the gain-discount framework to consider cases where relevance is modeled beyond topicality to contemplate other dimensions, such as understandability, explicitly. UBIRE assumes that dimensions are independent and do not affect each other. Based on the same assumptions, we defined the *MM* framework, which calculates a value for each dimension separately with respect to its gain and discount, and then combines these values into a unique effectiveness measure with a weighted harmonic mean.
10. **Chapter 7:** *What is the limitation of the state-of-the-art multidimensional evaluation framework and how can we overcome its limitation?* The UBIRE framework, which is the current state-of-the-art multidimensional evaluation framework, lacks interpretability as it does not permit the quantification of the contribution that each dimension has on the evaluation measure. In turn, the *MM* framework proposed here has the advantage of allowing experimenters to easily understand how each relevance dimension affects their system performance, as well as carefully tune the trade-off between topical relevance and other dimensions. We conducted experiments with the systems submitted to both CLEF 2015 and 2016 to evaluate whether UBIRE and *MM* would rank the system similarly and we found that the system rankings correlation of these two frameworks was higher than 0.90, meaning that while they both perform a very similar job in ranking systems, *MM* provides explainability to the experimenters on why a system performed better than another.
11. **Chapter 8:** *How can understandability estimations be integrated into retrieval methods to enhance the quality of the retrieved health information?* We experimented with three different strategies to integrate understandability into retrieval methods. The first was a simple re-ranking in which the retrieved results were ranked according to an understandability estimator, such as DCI, SMOG or even a machine-learned model to predict understandability of documents. We showed that as we increased the number of documents to be re-ranked, we traded topical relevance for understandability. We also showed that machine learning methods were superior to readability formulas as understandability estimators for re-ranking. The second approach, the fusion methods, aimed to find a balance between topical relevance and understandability automatically. The third method, the learning-to-rank experiments, yielded the best results. In particular, the combinations *LTR 3* and *LTR 5* that explored both the topical and understandability features, reached the

best results for CLEF 2015 and CLEF 2016, respectively, highly suggesting that models can be specially trained to promote more understandable search results while still providing an effective trade-off with topical relevance.

9.2 Limitations and Future Work

In Part II of this thesis, a limitation of our query log study was the lack of click-through information, which would have allowed us to perform a more detailed analysis of search behavior. Another limitation was that MetaMap can only annotate English text. Health consumers, in particular, prefer to query in their own language, as clearly noted by the high number of non-English queries that were removed from the Health on the Net search logs for this study. There is undoubtedly a vast amount of work to be done for supporting a multilingual scenario. A query analysis for languages other than English has many additional challenges, such as the lack of detailed language resources for many languages. A possible solution might be to increase the effort on automatic translation of resources such as the MeSH vocabulary. Another challenge identified concerns the automatic language detection of very short queries, which could be accomplished by exploring geolocation clues and user search history. The age of the query logs used in this thesis is yet another limitation. One of the consequences of the fast evolution of the Web is that user behavior analysis, such as the ones made in this thesis, might quickly become outdated (for example, the demographics of the search engine users have very likely changed in the last 15 years). Although we could not replicate our experiments in more recent search logs, we advocate that other researchers do so, assessing to what extent our findings are still valid.

Part III relied on data collected through the CLEF 2015 and CLEF 2016 evaluation campaigns to evaluate the effectiveness of methods that estimate the understandability of the Web documents. These assessments were obtained by asking ratings from medical experts and practitioners. Although they were asked to estimate the understandability of the content as if they were the patients they treat, there might have been noise and imprecisions in the collection mechanism because of the subjectivity of the task. Section 6.3 highlights this by showing that the agreement between assessors is relatively low. As future work, we may directly recruit health consumers: the task would still be subjective but could capture real ratings, rather than inferred or perceived ones. Despite this, our previous work had shown that no substantial differences were found in the downstream evaluation of retrieval systems when we acquired understandability assessments from health consumers for a subset of the CLEF 2015 collection [153]. Comparing our results with the newest developments in the natural language processing field, such as word embeddings and recurrent neural networks [80, 28], is left as future work.

A question left open in Part IV is whether *MM* correlates with human preferences and how it compares with *UBIRE* in this respect. To answer this, future work needs to consider user-based validation and comparison of the two multidimensional evaluation

frameworks. As seen in the experiments, the topical relevance assessments on the CLEF 2015 and 2016 collections are incomplete [77, 106], i.e., not all top ranked Web documents retrieved by the investigated methods have an explicit relevance assessment. This is often the case in information retrieval, where the validity of experiments based on incomplete assessments has been thoroughly investigated [171]. Nonetheless, we carefully controlled for the impact unassessed documents had in our experiments by measuring their number and using measures such as RBP that account for residuals and condensed variants. An important task left as future work is to investigate the effect of incomplete assessments on multidimensional evaluation frameworks.

There has been an increased concern regarding the inherent bias present in machine learning and retrieval models [45, 53, 132, 95, 42]. While this thesis aims to build retrieval systems for excluded populations, e.g., those with low-literacy rate, our models are not exempt from bias. For example, our results in Chapter 6 and our related research [153] show how assessing the understandability of documents can be highly subjective. We consider that having a more diverse group of assessors when collecting document understandability assessments would help to alleviate the bias in our systems. However, this task is left as future work.

The machine learning experiments described in Parts III and II, as well as the retrieval experiments described in Part IV of this thesis are strictly technical evaluations. Real users were involved in the collection of the ground truth assessments, but not in measuring how these systems impact their lives. Assessing how our findings translate to actual user experience is left as future work. An experimental setup similar to the one described by Pogacar et al. [163], in which a controlled laboratory study is conducted to investigate the effect of search results on choices of medical treatment, could be employed to assess the impact on people's decision when facing retrieved documents with different level of difficulty.

Finally, the methods investigated here do not provide a fully personalized search, with respect to how much of the health content consumers with different health knowledge might be able to understand. Instead, we focus on making the results understandable by anyone, and thus promote in the search results content that has the highest level of understandability. Although easy-to-read documents can be retrieved for the purpose of a direct handout (i.e., as a printout) to patients during a consultation, consumers with a more than average medical knowledge might benefit more from specialized content. We leave this personalization aspect, i.e., the tailoring of the understandability level of the promoted content with respect to the user's knowledge and abilities, to further work.

CLEF eHealth Data

A.1 Introduction

CLEF eHealth¹ is an evaluation lab organized within the Conference and Labs of the Evaluation Forum (CLEF). The aim is to build resources and methods to support health consumers, their next-of-kin, clinical staff, and health scientists in understanding, accessing, and authoring eHealth information in a multilingual setting. The lab has been running annually since 2013 [189, 107, 77, 106, 78] and historically has been built upon three main tasks in the health domain: information extraction, information management and information retrieval. We cover in this Appendix the latter task.

The information retrieval (IR) tasks, which started in 2013 and currently is running its fifth edition in 2018 [75, 76, 154, 226, 155, 103], embraces the TREC-style evaluation process, with a shared collection of documents and queries. Every year a task is proposed with a set of queries; teams submit runs in answer to the queries and their runs contribute to the subsequent formation of relevance assessments and evaluation. Tasks investigate the problem of retrieving Web pages to support the information needs of health consumers that are confronted with a health problem or medical condition and that use a search engine to seek a better understanding of their health. The use of the Web as a source of health-related information is a widespread practice among health consumers [130] and search engines are commonly used as a means to access health information available online [65].

The earlier iterations of this task, i.e., the 2013 and 2014 CLEF eHealth Lab Task 3 [75, 76], aimed at evaluating the effectiveness of search engines to support people when searching for information about their conditions, e.g., to answer queries like “thrombocytopenia treatment corticosteroids length”. These two evaluation exercises have provided valuable

¹<https://sites.google.com/site/clefehealth/>

resources and an evaluation framework for developing and testing new and existing techniques.

In 2015, the IR task took a different focus, explicitly centering on supporting consumers searching for self-diagnosis information [154], an important type of health information seeking activity [65]. The motivation for the shift was recent research showing that current commercial search engines are still far from being effective in answering such unclear and underspecified queries [225].

An important innovation of 2015’s task consisted of assessing for the first time how difficult to understand were the content of retrieved Web pages. Previous research has shown that exposing people with no or scarce medical knowledge to complex medical language may lead to erroneous self-diagnosis and self-treatment and that access to medical information on the Web can lead to the escalation of concerns about common symptoms (e.g., cyberchondria) [17, 207].

In 2016, the IR task expanded on the 2015 task, by considering not only self-diagnosis information needs but also needs related to treatment and management of health conditions [226]. Apart from understandability assessments, assessors were asked to assess pages on how trustworthy the content of the page was.

Finally, in 2017, the IR task continued the growth path identified in past years and focused on conducting assessments on deeper pooled sets than was possible in previous years of the task. As the assessments were just released, we do not include them in this thesis.





In the rest of this Appendix, we study the assessments made in CLEF eHealth 2015 and 2016.

A.2 Query Sets

In CLEF eHealth 2015, queries aimed to simulate the situation of health consumers seeking information to understand symptoms or conditions they may be affected by. This was achieved by using imagery or video stimuli of 23 symptoms or conditions as prompts for the query creators. A cohort of 12 query creators was used and each query creator was given the imagery of 10 conditions for which they were asked to generate up to 3 queries per condition. Each condition/image pair was presented to more than one person.

The task collected a total of 266 possible unique queries; of these, 67 queries (22 conditions with 3 queries and 1 condition with 1 query) were selected to be used as part of the CLEF 2015 task. We selected the three query variants using the following procedure: a initial query was randomly chosen for each condition (this query was called the *pivot query*), then five native English speakers were recruited to select, for each query, the *most* and the *least* similar query variant compared to the pivot one. Examples of queries, query variations and imagery material used for the query creation are provided in Table A.1. After the query set was released, one typo was found in query 62, which could compromise the

Table A.1: Examples of queries from the CLEF eHealth 2015.

Image	Information Need	Query Type	QueryId	Query Variation
	Ringworm	Pivot	03	dry red and scaly feet in children
		Most similar	38	scaly red itchy feet in children
		Least similar	45	dry feel with irritation
	Scabies	Pivot	04	itchy lumps skin
		Most similar	43	itchy raised bumps skin
		Least similar	21	common itchy skin rashes
	Onycholysis	Pivot	61	fungernail bruises
		Most similar	19	bruised thumb nail
		Least similar	44	nail getting dark
	Rocky Mountain Spotted Fever	Pivot	27	return from overseas with mean spots on legs
		Most similar	01	many red marks on legs after traveling from us
		Least similar	58	39 degree and chicken pox

translations. In order to keep consistency between the English query and all translations made by the experts, query 62 was excluded. See [154] for more details on the query creation method.

In CLEF eHealth 2016, we considered real health information needs expressed by the general public through posts published in public health web forums. Forum posts were extracted from the *AskDocs* section of Reddit². This section of Reddit allows users to post a description of a medical case or ask a medical question seeking medical information such as diagnosis, or details regarding treatments. Users can also interact through comments.

We selected posts that were descriptive, clear and understandable. Posts with information regarding the author or patient (in case the post author sought help for another person), such as demographics (age, gender), medical history and current medical condition, were preferred. Figure A.1 shows an example of post extracted to generate queries. Each of the selected forum posts was presented to 6 query creators with different medical expertise: these included 3 medical experts (final year medical students undertaking rotations in hospitals) and 3 lay users with no prior medical knowledge. A total of 300 queries were created, 6 for each one of the 50 forum posts.

In Figure A.2 we show variants 1, 2 (both generated by laypeople) and 4 (generated by an expert) created for post number 103 (posts started from number 101), shown in Figure A.1. The assessments for different query variants were made separately but pooled together in the analysis made in this Appendix. See [226] for more details on the query creation method.

²<https://www.reddit.com/r/AskDocs/>

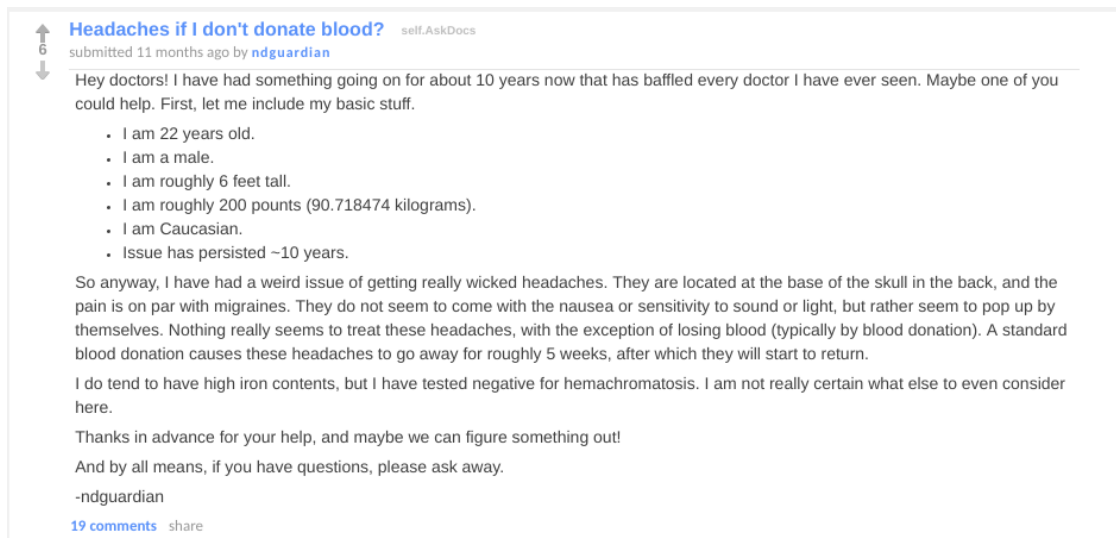


Figure A.1: Post from Reddit’s Section AskDocs. It was used to generate queries ranging from 103001 to 103006 in CLEF eHealth 2016.

```

<queries>
  ...
  <query>
    <id> 103001 </id>
    <title>headaches relieved by blood donation</title>
  </query>
  <query>
    <id> 103002 </id>
    <title>high iron headache</title>
  </query>
  ...
  <query>
    <id> 103004 </id>
    <title>headaches caused by too much blood or
      "high blood pressure"</title>
  </query>
  ...
</queries>

```

Figure A.2: Extract from the query set released in CLEF eHealth 2016. The assessments for all variants were pooled together in the analysis of this Appendix.

A.3 Assessments

Assessments made for CLEF eHealth 2015 and 2016 were performed by paid final year medical students or medical doctors who had access to queries, documents, and relevance criteria. In all the years that the IR task ran, documents were assessed with respect to their relevance using the following grades: **0** for *Not Relevant*, **1** for *Somewhat Relevant* and **2** for *Highly Relevant*.

Since 2015, understandability assessments are collected for the documents in the assessment pool. In 2015, these assessments were obtained using the following 4 grades: **0** for

it is very technical and difficult to read and understand, **1** for *it is somewhat technical and difficult to read and understand*, **2** for *it is somewhat easy to read and understand* and **3** for *it is very easy to read and understand*. In 2016, we decided to experiment with a larger scale for understandability allowing assessors to use a slider to determine how hard a document was. The scale mapped assessments to an integer value between 0 and 100, however now lower values were used for documents that were easy to understand. All assessments were collected through a purposely customized version of the *Relevation!* toolkit [112]. Figure A.3 shows an instance of the assessments made in 2015 (top) and 2016 (bottom).

A.4 CLEF eHealth 2015

In 2015, 12,092 documents were assessed with respect to their topical relevance and understandability. Out of those, 2,515 (21%) were assessed as *Somewhat Relevant* or *Highly Relevant* and 8,039 (67%) were assessed as *Easy* or *Somewhat Easy*. Figure A.4 shows, independently, the distribution of both topicality (left-hand side) and understandability (right-hand side) assessments. In Figure A.5, we depict how the understandability assessments are distributed for each different topical relevance level.

The full assessment distribution per topic of both topical and understandability relevance is shown in Figure A.6. No topically relevant documents were found for query 42, *patchy bleeding under skin*, and the largest number of topically relevant documents was found for query 46, *baby cough*. With respect to understandability, query 20, *parkinson's disease*, presented the lowest average understandability score, while query 30, *weird sounds when breathing*, presented the highest one.

A.5 CLEF eHealth 2016

In CLEF eHealth 2016, instead of considering each query variant as a unique query, we grouped the variants together and considered the assessments for each *topic*. Thus, each topic has 6 query variants. Five hundred documents were assessed per topic, in total 25,000 document assessments for the 50 topics.

Figure A.7 shows that most of the assessed documents were assessed as *Not Relevant* (85%). The overall distribution of understandability assessments is shown in Figure A.8: an understandability label close to 0 was assigned to easy-to-read documents and close to 100 was assigned to difficult-to-understand documents. It is important to mention that the default value for understandability assessments was 50, the mean value. Assessors had to compulsorily move the slider that controls the understandability assessment. Thus, assessors explicitly decided to assess 6,181 documents as average understandability (= label 50). Figure A.9 shows how the understandability assessments were distributed for each different topical relevance level.

Figure A.10 shows the distribution of topical relevance per topic. Topic 50 (with query variants such as *foley catheter bruising* and *painful erection following foley catheter worsening*) is the one with the lowest number of relevant documents: only one document assessed as *Somewhat Relevant*. The other extreme is Topic 2 (e.g., *anal skin tag removal pain* and *do I need general anaesthetic for anal skin tag removal*) with 336 relevant documents from which 126 were assessed as *Highly Relevant*.

Likewise, Figure A.11 shows the distribution of understandability assessments per topic. Topic 17 (e.g., *tylenol cold benylin extra strength combo risks* and *can I take Tylenol Cold and Benylin at same time?*) is the one with the easiest documents with an average understandability label of 21.71, while Topic 31 (e.g., *biopsy HSV benign penile skin mild infiltration explained* and *lymphocytic infiltration*) is the most difficult one with an average understandability score of 73.91.

The screenshot shows the Relevation interface for document ID 127. The document title is "127 - return from overseas with mean spots on legs". The document content is a clinical trial page titled "Subjective and Objective Methods of Assessing Walking Limitation Due to Claudication". The trial status is "This study has been completed." and it was first received on November 24, 2011, last updated on May 14, 2012. The sponsor is Sheffield Teaching Hospitals NHS Foundation Trust. The purpose of the study is to test the validity and reliability of a questionnaire to measure walking capacity in patients with leg artery disease. The interface includes a navigation bar, a document viewer, and a relevance criteria sidebar with options for Unjudged, Highly relevant, Somewhat relevant, and Not relevant. There are also sections for Readability and a comment field.

The screenshot shows the Relevation interface for document ID 129. The document title is "129 - Results should discuss the motor tics, fasciculations or cramping abdominal pain in the context of anxiety and/or depression...". The document content is an article titled "Tourette Syndrome: No Laughing Matter" from the ADD ADHD Information Library Home. The article discusses the defining criteria of Tourette syndrome, its prevalence, and its impact on children. The interface includes a navigation bar, a document viewer, and a judgement sidebar with a scale for Understandability and Trustworthiness. There are also sections for Comment (optional) and a save button.

Figure A.3: Screenshots of Relevation!, the system used to collect assessments at CLEF eHealth in 2015 (top) and 2016 (bottom).

A. CLEF eHEALTH DATA

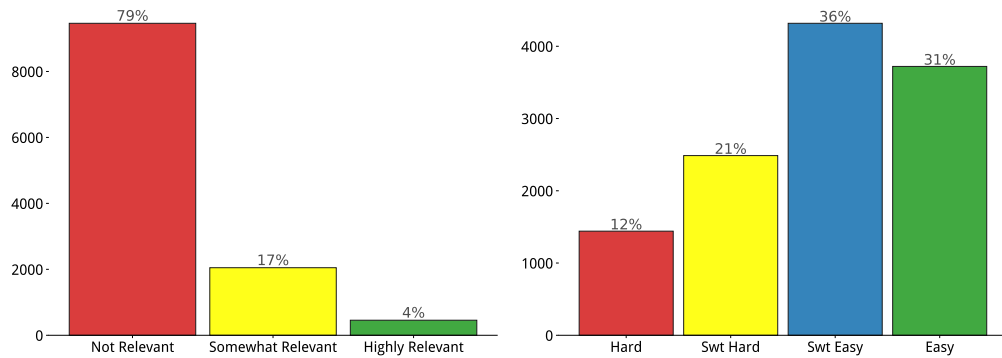


Figure A.4: The distribution of topicality (left) and understandability (right) assessments for CLEF eHealth 2015.

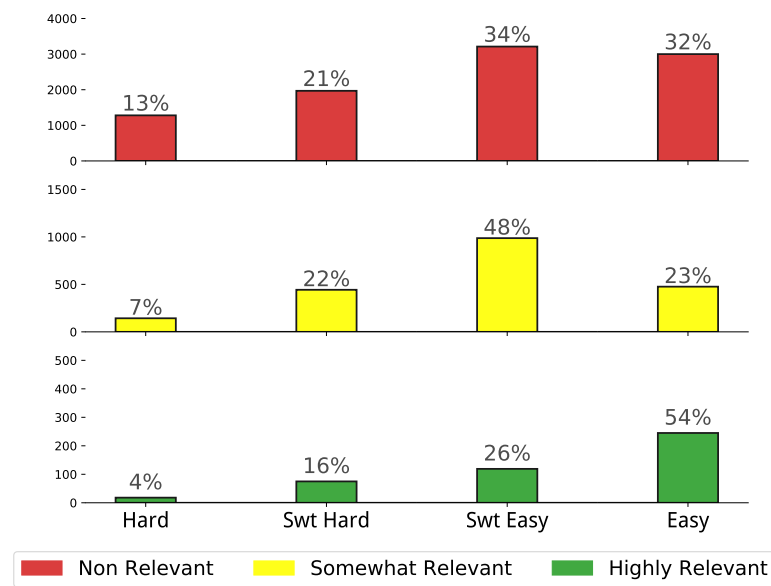


Figure A.5: The distribution of understandability assessments broken down by relevance level for CLEF eHealth 2015.

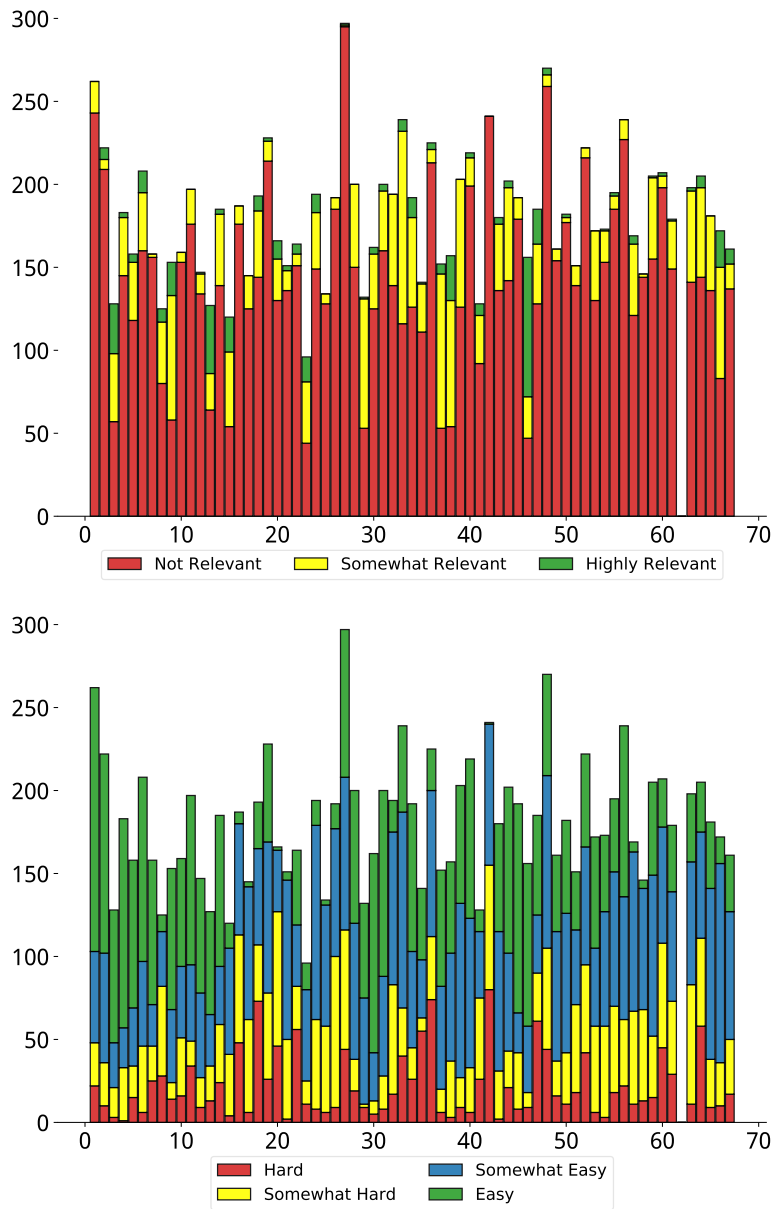


Figure A.6: Assessments for topical relevance (top) and understandability relevance (bottom) distributed per topic number in CLEF eHealth 2015. Note that topic 62 was removed from the evaluation and it is not considered here.

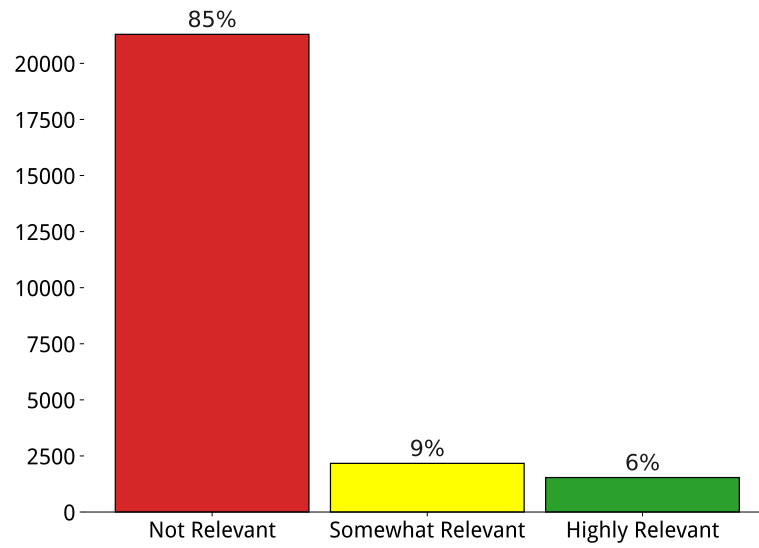


Figure A.7: Distribution of topical relevance assessment for the 25,000 assessed documents in CLEF eHealth 2016.

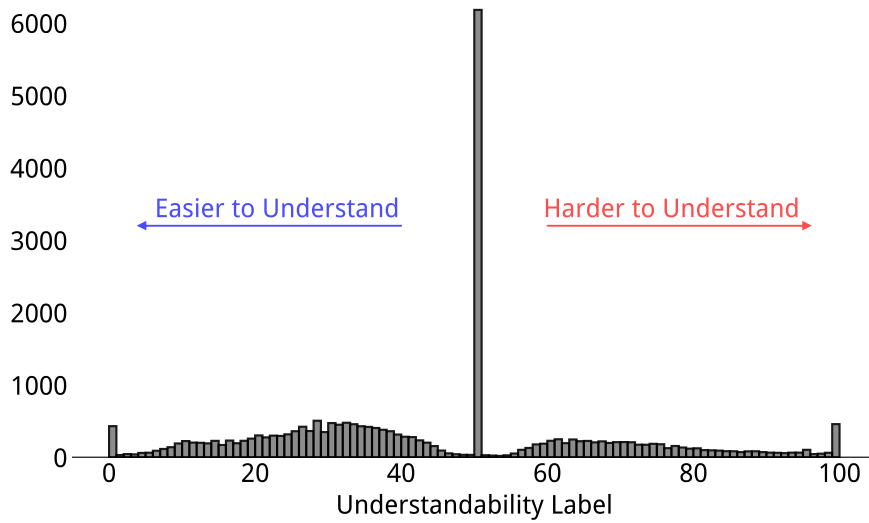


Figure A.8: Distribution of understandability assessments for CLEF eHealth 2016.

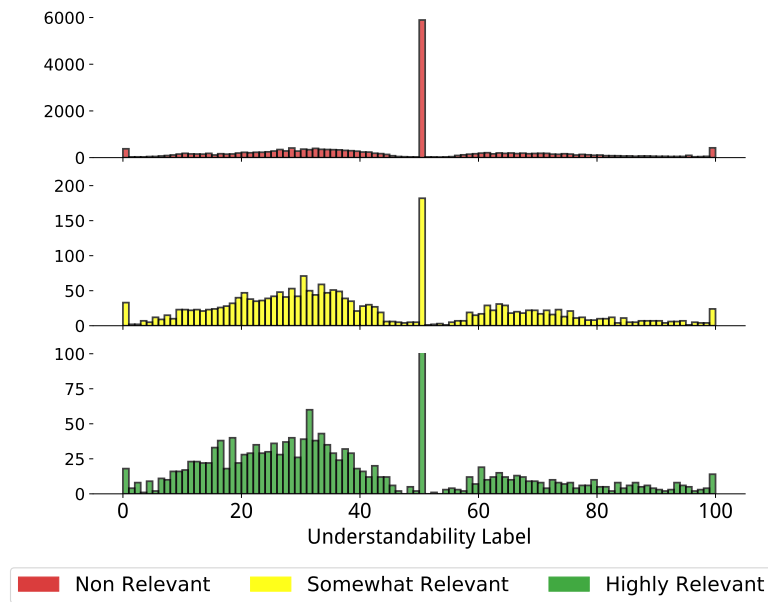


Figure A.9: The distribution of understandability assessments broken down by topical relevance level for CLEF eHealth 2016.

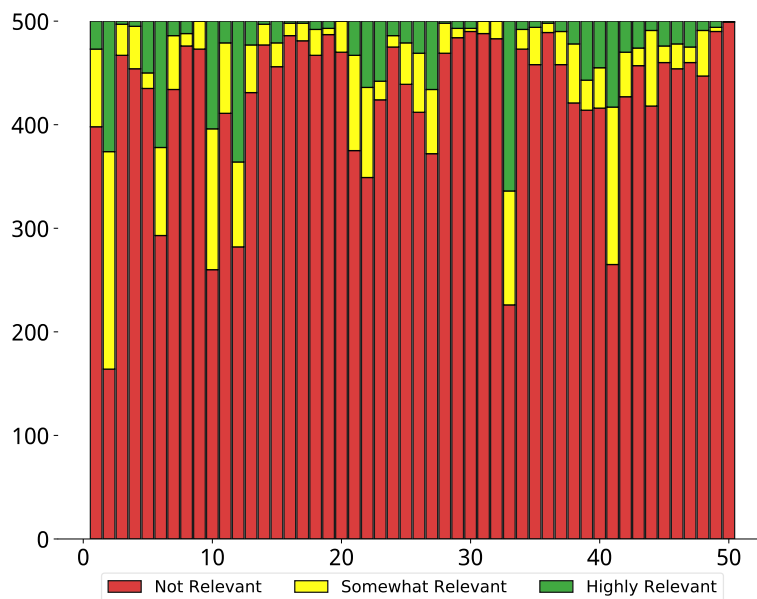


Figure A.10: Topical relevance assessments distributed for each of the 50 topics for CLEF eHealth 2016.

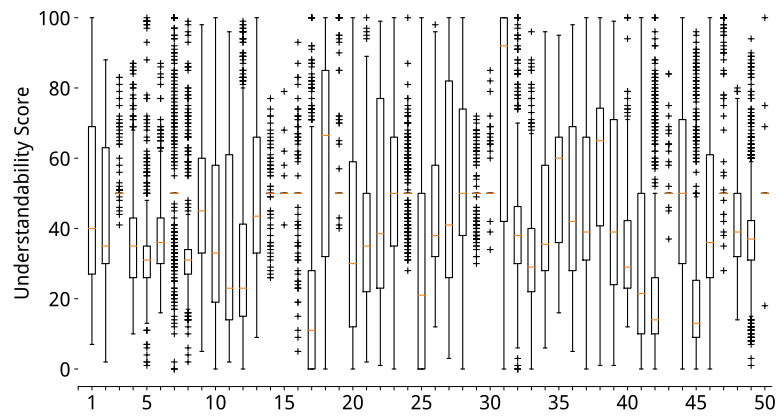


Figure A.11: Distribution per topic of understandability scores for all 50 topics of CLEF eHealth 2016 collection.

List of Figures

1.1	The cumulative distribution of the Dale-Chall Index (DCI) of search results. DCI measures the years of schooling required to understand a document. The average US resident reads at or below an 8th-grade level (dashed line) [43, 199, 50, 187], which is the level suggested by NIH for health information on the Web [196]. The distribution for HON is similar to that of the baseline used in this article (BM25). Our best method (XGB) re-ranks documents to provide more understandable results; its distribution is similar to that of an “Oracle” system.	7
3.1	Two real user sessions extracted from AOL logs, S_i is classified as a search for medical content, while S_j is not.	30
3.2	The datasets used here are placed on an expertise scale. The expertise level increases as a dataset is placed more to the right-hand side of the scale. . . .	32
3.3	MeSH hierarchy with the Disease branch expanded	33
3.4	Two different user queries are enriched with information extracted with MetaMap. In the top part, the same example used in Table 3.2 is processed by MetaMap. In the bottom part, the query “lung cancer treatment” is more ambiguous and results in different mappings, such as <i>Lung (Entire lung) / Cancer Treatment (Cancer Therapeutic Procedure)</i> and <i>Lung Cancer (Malignant neoplasm of lung) / Treatment (Therapeutic procedure)</i>	35
3.5	Popular categories according to MeSH mappings	42
3.6	The top most frequently used semantic types (frequency in percentage). Many of the most used types are aggregated to study the user focus described in Table 3.8	44
4.1	Feature importance according to the Gini importance score generated by the Random Forest classifier. The higher, the more important the feature. The error bars represent the standard deviation from the mean value for each feature.	55

5.1	Simplified Wikipedia entry for Readability (top) and the output of <i>Naïve</i> (bottom). In the bottom part, we show the result of the preprocessing approach termed <i>DoNotForcePeriod</i> (left), which does not modify the text extracted by the HTML parser, and that of the alternative preprocessing approach termed <i>ForcePeriod</i> (right), which adds a period as sentence boundary at the end of every line. The <i>DoNotForcePeriod</i> approach concatenates all the text till it reaches a sentence boundary, producing longer sentences than the <i>ForcePeriod</i> approach.	62
5.2	Readability scores for each measure based on various preprocessing and sentence boundary methods. Error bars indicate 95% confidence intervals around the mean.	65
5.3	Kendall τ correlation and 95% confidence intervals between the <i>ForcePeriod</i> and <i>DoNotForcePeriod</i> approaches for sentence boundary identification.	66
5.4	Kendall τ correlation and 95% confidence intervals between the approaches for HTML preprocessing, under different settings for sentence boundary identification.	67
6.1	Wikipedia document on hyperthermia. The rectangular red box identifies the Infobox on the right-hand side. This document is filtered as medical because it contains entries for <i>ICD-9</i> , <i>DiseasesDB</i> and <i>MeSH</i>	75
6.2	Correlation between the scores of various readability formulas and the understandability scores assessed by humans collected in CLEF eHealth 2015 considering different HTML preprocessing pipelines	78
6.3	Correlation between the scores of various readability formulas and the understandability scores assessed by humans collected in CLEF eHealth 2016 considering different HTML preprocessing pipelines	79
6.4	Box plots divided by feature groups. Correlations are calculated using understandability labels from relevant documents assessed in CLEF eHealth 2015	82
6.5	Box plots divided by feature groups. Correlations are calculated using understandability labels from relevant documents assessed in CLEF eHealth 2016	83
6.6	Average difference and confidence interval when comparing the Spearman correlation for each group. Groups whose confidence intervals overlap are not significantly different.	85
6.7	Average difference and confidence interval when comparing the Spearman correlation of the three preprocessing HTML tools we validated. The mean absolute Spearman correlation of the JusText preprocessing is significantly higher than the Naïve preprocessing. <i>Naïve</i> and <i>Justext</i> are the only groups that are significantly different, i.e., their confidence intervals do not overlap.	86

7.1	Gaussian distribution for different μ : higher μ generates higher understandability labels (more difficult documents were retrieved). In the experiments in this chapter, only documents with understandability lower than 40 are considered easy-to-understand (understandability threshold shown as dotted line).	102
A.1	Post from Reddit's Section AskDocs. It was used to generate queries ranging from 103001 to 103006 in CLEF eHealth 2016.	130
A.2	Extract from the query set released in CLEF eHealth 2016. The assessments for all variants were pooled together in the analysis of this Appendix.	130
A.3	Screenshots of Relevation!, the system used to collect assessments at CLEF eHealth in 2015 (top) and 2016 (bottom).	133
A.4	The distribution of topicality (left) and understandability (right) assessments for CLEF eHealth 2015.	134
A.5	The distribution of understandability assessments broken down by relevance level for CLEF eHealth 2015.	134
A.6	Assessments for topical relevance (top) and understandability relevance (bottom) distributed per topic number in CLEF eHealth 2015. Note that topic 62 was removed from the evaluation and it is not considered here.	135
A.7	Distribution of topical relevance assessment for the 25,000 assessed documents in CLEF eHealth 2016.	136
A.8	Distribution of understandability assessments for CLEF eHealth 2016.	136
A.9	The distribution of understandability assessments broken down by topical relevance level for CLEF eHealth 2016.	137
A.10	Topical relevance assessments distributed for each of the 50 topics for CLEF eHealth 2016.	137
A.11	Distribution per topic of understandability scores for all 50 topics of CLEF eHealth 2016 collection.	138

List of Tables

2.1	The most frequently used readability formulas. C is the number of characters, DCW is the number of words found in the Dale-Chall list of 3,000 common words, LW is the number of long words (words with 6 or more characters), PW is the number of polysyllables words (words with 3 or more syllables) S is the number of sentences, Sy is the number of syllables and W is the number of words in the text.	23
-----	---	----

3.1	ODP categories used to filter the AOL-Medical. These categories are the most relevant ones related to Medicine in ODP hierarchy (see http://www.dmoz.org/Health/Medicine/)	29
3.2	A concept is potentially linked to various AUI (atom), SUI (string), and LUI (term). We used MetaMap to map a user query, e.g. “Atrial Fibrillation” to the different existing concepts (C0004238, C1963067). Note that each concept is associated to one single semantic meaning.	34
3.3	General Statistics describing the query logs. We use * when the median value was used instead of the average, and <i>N/A</i> when the data was not available.	39
3.4	General Statistics – Stratified by expertise. C for consumers and E for experts	39
3.5	Top queries and terms and their relative frequency (%) among all queries	40
3.6	General MeSH Statistics	41
3.7	Aggregated percentages for query modifications along the sessions	43
3.8	User focus when searching for medical content in a single session.	46
3.9	Cycle Sequence along a single session	46
4.1	Groups and features used in this work. Two basic behavior features are used in the baseline models.	52
4.2	Results of the experiment on classifying users according to their expertise. Different combinations of classifiers and feature sets are explored. Underline is used to show values that are statistically different from the classifier with Behavior Features using a student’s t-test with Bonferroni correction, $p < .00069$	53
5.1	Number of words and sentences (mean and standard deviation) for documents in CLEF 2015 eHealth corpus (approximately 1 million documents) and CLEF 2016 pool of assessed documents (25.000 documents), as obtained by the three preprocessing tools and the two approaches to sentence boundaries (see Section 5.2).	64
6.1	Methods to estimate the understandability of Web documents. \star : raw values are used; \diamond : values normalized by the number of words in a document are used; \dagger : values normalized by the number of sentences in a document are used.	72
6.2	General statistics for the auxiliary corpora used in this work	75
6.3	Correlation results (Mean \pm Std) each correlation measure across different preprocessing pipelines for readability formula in Figures 6.2 (CLEF eHealth 2015) and 6.3 (CLEF eHealth 2016). The symbols <i>B</i> , <i>J</i> and <i>N</i> are used to show significant differences between the current value and the one obtained by using, respectively, <i>Boilerpipe</i> , <i>JusText</i> or the <i>Naïve</i> preprocessing pipeline.	80
6.4	Mean absolute difference between the average correlation coefficient when using <i>ForcePeriod</i> and <i>DoNotForcePeriod</i> for each correlation measure across different preprocessing pipelines for each readability formula in Figures 6.2 (CLEF eHealth 2015) and 6.3 (CLEF eHealth 2016). No significant differences were found when comparing <i>ForcePeriod</i> to <i>DoNotForcePeriod</i>	80

6.5	Correlation results (Absolute Mean \pm Std) for each readability formula across different preprocessing pipelines and heuristics. These results summarize those of Figures 6.2 (CLEF eHealth 2015) and 6.3 (CLEF eHealth 2016). Highest results for each collection are shown in bold.	81
6.6	Methods with highest correlation per group. In bold are the methods that achieved the highest correlation for a correlation measure.	86
6.7	Results for the document understandability classification task using the relevant documents from CLEF 2015 collection. Accuracy and Macro F_1 (both Mean \pm Std) were calculated across the ten folds in a 10-fold cross-validated experiment. Results of each fold are accumulated and correlation with human assessments is shown in the last three columns. Bold is used to show the best values per experiment.	90
6.8	Results for the document understandability classification task using the relevant documents from CLEF 2016 collection. R^2 and Mean Absolute Error (both Mean \pm Std) were calculated across the ten folds in a 10-fold cross-validated experiment. Results of each fold are accumulated and correlation with human assessments is shown in the last three columns. Bold is used to show the best values per experiment.	93
7.1	We varied T , the expected proportion of topically relevance documents (rows), and the mean μ of Gaussian distribution used to generate understandability labels (columns). A smaller μ means that easier to read documents are retrieved. We showed the average and standard deviation of each experiment. S1 (blue) and S2 (yellow) represent two systems with similar MM_{RBP} , but very different $RBP(=RBP_t)$ and RBP_u results.	103
7.2	Kendall- τ correlation for systems participating in CLEF eHealth 2015 and 2016.	104
8.1	Learning to rank settings.	111
8.2	Results obtained by integrating understandability estimations within retrieval methods on CLEF 2015. Baseline runs are reported at table indices 1–3 (the index column is labeled Idx). Re-ranking experiments are reported at indices 4–21. Fusion experiments are reported at indices 22–30. Learning to rank experiments are reported at indices 31–35. All measures were calculated up to rank $n = 10$. The highest result of each set of experiments is reported in boldface for each measure. Statistically significant differences compared to ECNU are indicated with \diamond , while differences between an original run (indices 1-3) and its modifications are indicated with \dagger (paired, two-tail t-test with Bonferroni correction, $p < .00017$ (.05 / 294)).	113

8.3	Results obtained by integrating understandability estimations within retrieval methods on CLEF 2016. Baseline runs are reported at table indices 1–3 (the index column is labeled Idx). Re-ranking experiments are reported at indices 4–21. Fusion experiments are reported at indices 22–30. Learning to rank experiments are reported at indices 31–35. All measures were calculated up to rank $n = 10$. The highest result of each set of experiments is reported in boldface for each measure. Statistically significant differences compared to GUIR are indicated with \diamond , while differences between an original run (indices 1-3) and its modifications are indicated with \dagger (paired, two-tail t-test with Bonferroni correction, $p < .00017$ (.05 / 294)).	114
A.1	Examples of queries from the CLEF eHealth 2015.	129

Bibliography

- [1] Eytan Adar. User 4XXXXX9: Anonymizing Query Logs. In *Query Logs Workshop at the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 1–8.
- [2] Linda Andersson, Mihai Lupu, João Palotti, Allan Hanbury, and Andreas Rauber. When is the Time Ripe for Natural Language Processing for Patent Passage Retrieval? In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 1453–1462, 2016.
- [3] Linda Andersson, Mihai Lupu, João Palotti, Florina Piroi, Allan Hanbury, and Andreas Rauber. Insight to Hyponymy Lexical Relation Extraction in the Patent Genre Versus Other Text Genres. In *Proceedings of the First International Workshop on Patent Mining and Its Applications (IPaMin 2014) co-located with Konvens 2014, Hildesheim, Germany, October 6-7, 2014*.
- [4] Alan R. Aronson. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: the MetaMap Program. *Proceedings of AMIA, Annual Symposium. Washington, DC. November 3-7, 2001*, pages 17–21, 2001.
- [5] Alan R. Aronson, Olivier Bodenreider, Florence Chang, Sussne M. Humphrey, James G. Mork, Stuart J. Nelson, Thomas C. Rindfleisch, and John Wilbur. The NLM Indexing Initiative. In *Proceedings of AMIA Annual Symposium. Los Angeles, CA. November 4-8, 2000*, pages 17–21, 2000.
- [6] Alan R. Aronson and François-Michel Lang. An Overview of MetaMap: Historical Perspective and Recent Advances. *JAMIA*, 17(3):229–236, 2010.
- [7] Alan R. Aronson and Thomas C. Rindfleisch. Query Expansion using the UMLS Metathesaurus. *Proceedings of the AMIA Annual Symposium. Nashville, TN. October 25-29, 1997*, pages 485–489, 1997.
- [8] Niraj Aswani, Thomas Beckers, Erich Birngruber, Célia Boyer, Andreas Burner, Jakub Bystron, Khalid Choukri, Sarah Cruchet, Hamish Cunningham, Jan Dedek, Ljiljana Dolamic, René Donner, Sebastian Dungs, Ivan Eggel, Antonio Foncubierta-Rodríguez, Norbert Fuhr, Adam Funk, Alba Garcia Seco de Herrera, Arnaud

- Gaudinat, Georgi Georgiev, Julien Gobeill, Lorraine Goeuriot, Paz Gomez, R. Mark Greenwood, Manfred Gschwandtner, Allan Hanbury, Jan Hajic, Jaroslava Hlaváčová, Markus Holzer, Gareth J. F. Jones, Blanca Jordan, Matthias Jordan, Klemens Kaderk, Franz Kainberger, Liadh Kelly, Sascha Kriewel, Marlene Kritz, Georg Langs, Nolan Lawson, Dimitrios Markonis, Iván Martínez, Vassil Momtchev, Alexandre Masselot, Hélène Mazo, Henning Müller, João R. M. Palotti, Pavel Pecina, Konstantin Pentchev, Deyan Peychev, Natalia Pletneva, Diana Pottecher, Angus Roberts, Patrick Ruch, Alexander Sachs, Matthias Samwald, Priscille Schneller, Veronika Stefanov, Miguel Angel Tinte, Zdenka Uresová, Alejandro Vargas, and Dina Vishnyakova. Khresmoi - Multilingual Semantic Search of Medical Text and Images. In *MEDINFO 2013 - Proceedings of the 14th World Congress on Medical and Health Informatics, 20-13 August 2013, Copenhagen, Denmark*, page 1266, 2013.
- [9] Samuel R. Atcherson, Ashley E. DeLaune, Kristie Hadden, Richard I. Zraick, Rebecca J. Kelly-Campbell, and Carlos P. Minaya. A Computer-Based Readability Analysis of Consumer Materials on the American Speech-Language-Hearing Association Website. *Contemporary Issues in Communication Science & Disorders*, 41, 2014.
- [10] Ricardo Baeza-Yates. Applications of web query mining. In *Advances in Information Retrieval, 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005, Proceedings*, pages 7–22. Springer-Verlag, 2005.
- [11] Ricardo Baeza-Yates and Berthier de Araújo Neto Ribeiro. *Modern Information Retrieval*. New York: ACM Press; Harlow, England: Addison-Wesley, 2011.
- [12] Alexandros Bampoulidis, Mihai Lupu, João Palotti, Sokratis Metallidis, Jon Brassey, and Allan Hanbury. Interactive Exploration of Healthcare Queries. In *14th International Workshop on Content-Based Multimedia Indexing, CBMI 2016, Bucharest, Romania, June 15-17, 2016*, pages 1–4, 2016.
- [13] Alexandros Bampoulidis, João Palotti, Mihai Lupu, Jon Brassey, and Allan Hanbury. Does Online Evaluation Correspond to Offline Evaluation in Query Auto Completion? In *Advances in Information Retrieval - 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings*, pages 713–719, 2017.
- [14] Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. Cleaneval: a Competition for Cleaning Web Pages. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association, 2008.
- [15] Regina Barzilay and Mirella Lapata. Modeling Local Coherence: An Entity-based Approach. *Computational Linguistics*, 34(1):1–34, March 2008.

- [16] Shirley Ann Becker. A Study of Web Usability for Older Adults Seeking Online Health Resources. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 11(4):387–406, 2004.
- [17] Mike Benigeri and Pierre Pluye. Shortcomings of health information on the Internet. *Health promotion international*, 18(4):381–386, 2003.
- [18] Suresh K. Bhavnani. Domain-specific Search Strategies for the Effective Retrieval of Healthcare and Shopping Information. In *Extended abstracts of the 2002 Conference on Human Factors in Computing Systems, CHI 2002, Minneapolis, Minnesota, USA, April 20-25, 2002*, pages 610–611. ACM, 2002.
- [19] Carl-Hugo Björnsson. Readability of Newspapers in 11 Languages. *Reading Research Quarterly*, 18(4):480–497, 1983.
- [20] Pia Borlund. The Concept of Relevance in IR. *Journal of the Association for Information Science and Technology*, 54(10):913–925, 2003.
- [21] Célia Boyer, Vincent Baujard, and Antoine Geissbuhler. Evolution of Health Web Certification Through the HONcode Experience. *Studies in Health Technology and Informatics*, 169:53–7, 2011.
- [22] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct 2001.
- [23] David J. Brenes and Daniel Gayo-Avello. Stratified Analysis of AOL Query Log. *Information Sciences*, 179(12):1844–1858, 2009.
- [24] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to Rank Using Gradient Descent. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, pages 89–96. ACM, 2005.
- [25] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to Rank: From Pairwise Approach to Listwise Approach. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, pages 129–136. ACM, 2007.
- [26] Ben Carterette. System Effectiveness, User Models, and User Utility: a Conceptual Framework for Investigation. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 903–912, 2011.
- [27] Marc-Allen Cartright, Ryen W. White, and Eric Horvitz. Intentions and Attention in Exploratory Health Search. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 65–74. ACM, 2011.

- [28] Miriam Cha, Youngjune Gwon, and H. T. Kung. Language Modeling by Clustering with Word Embeddings for Text Readability Assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 2003–2006. ACM.
- [29] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pages 621–630. ACM, 2009.
- [30] Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Editorial: Special Issue on Learning from Imbalanced Data Sets. *SIGKDD Explor. Newsl.*, 6(1):1–6, June 2004.
- [31] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 785–794. ACM, 2016.
- [32] Charles L Clarke, Nick Craswell, and Ian Soboroff. Overview of the TREC 2009 Web Track. In *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009*, 2009.
- [33] Cyril Cleverdon. Evaluation tests of information retrieval systems. *Journal of Documentation*, 26(1):55–67, 1970.
- [34] Cyril W. Cleverdon. The significance of the cranfield tests on index languages. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '91*, pages 3–12, New York, NY, USA, 1991. ACM.
- [35] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2 edition, July 1988.
- [36] Michael J. Cole, Jacek Gwizdka, Chang Liu, Nicholas J. Belkin, and Xiangmin Zhang. Inferring User Knowledge Level from Eye Movement Patterns. *Information Processing & Management*, 49(5):1075 – 1091, 2013.
- [37] Meri Coleman and T. L. Liau. A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology*, 1975.
- [38] Kevyn Collins-Thompson. Computational Assessment of Text Readability: A Survey of Current and Future Research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135, 2014.
- [39] Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Sebastian de la Chica, and David Sontag. Personalizing Web Search Results by Reading Level. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management*,

- CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 403–412. ACM, 2011.
- [40] Kevyn Collins-Thompson and Jamie Callan. Predicting Reading Difficulty with Statistical Language Models. *Journal of the Association for Information Science and Technology*, 56(13):1448–1462, 2005.
- [41] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 758–759. ACM, 2009.
- [42] Rachel Courtland. Bias Detectives: the Researchers Striving to Make Algorithms Fair. *Nature*, 558(7710):357, 2018.
- [43] Connie F. Cowan. Teaching Patients with Low Literacy Skills. *Fuszard’s Innovative Teaching Strategies in Nursing*, page 278, 2004.
- [44] Koby Crammer and Yoram Singer. Pranking with Ranking. In *Advances in Neural Information Processing Systems 14. Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada*, pages 641–647, 2002.
- [45] Kate Crawford and Ryan Calo. There is a blind spot in ai research. *Nature News*, 538(7625):311, 2016.
- [46] Fabio Crestani and Heather Du. Written Versus Spoken Queries: A Qualitative and Quantitative Comparative Analysis. *Journal of the American Society for Information Science and Technology*, 57(7):881–890, 2006.
- [47] W Bruce Croft, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*, volume 520. Addison-Wesley Reading, 2010.
- [48] Sarah L. Cutrona, Kathleen M. Mazor, Sana N. Vieux, Tana M. Luger, Julie E. Volkman, and Lila J. Finney Rutten. Health Information-Seeking on Behalf of Others: Characteristics of “Surrogate Seekers”. *Journal of cancer education*, 30(1):12–19, 2015.
- [49] Edgar Dale and Jeanne S. Chall. A Formula for Predicting Readability: Instructions. *Educational Research Bulletin*, 27(2):37–54, 1948.
- [50] Terry C. Davis and Michael S. Wolf. Health Literacy: Implications for Family Medicine. *Family Medicine*, 36(8):595–598, 2004.
- [51] Dina Demner-Fushman, Susanne M. Humphrey, Nicholas C. Ide, Russell F. Loane, James G. Mork, Patrick Ruch, Miguel E. Ruiz, Lawrence H. Smith, W. John

- Wilbur, and Alan R. Aronson. Combining Resources to Find Answers to Biomedical Questions. In *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 5-9, 2007*, 2007.
- [52] Joshua C. Denny, Jeffrey D. Smithers, Randolph A. Miller, and Anderson Spickard. "Understanding" Medical School Curriculum Content using Knowledge Map. *Journal of the American Medical Informatics Association*, 10(4):351–362, 2003.
- [53] Fernando Diaz. Worst Practices for Designing Production Information Access Systems. *ACM SIGIR Forum*, 50(1):2–11, 2016.
- [54] Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010*, pages 847–848. ACM, July 2010.
- [55] William H. Dubay. The Principles of Readability. *Costa Mesa, CA: Impact Information*, 2004.
- [56] Michael Eisenberg and Linda Schamber. Relevance: The Search for a Definition. In *Proceedings of the 51st Asis Annual Meeting, 1988/Information and Technology American Society For Information Science and Technology Meeting. Atlanta, Georgia, October 23-27, 1988*, pages 164–168, 1988.
- [57] Chandy Ellimoottil, Anthony Polcari, Adam Kadlec, and Gopal Gupta. Readability of Websites Containing Information about Prostate Cancer Treatment Options. *The Journal of urology*, 188(6):2171–2176, 2012.
- [58] John O. Elliott and Bassel F. Shneker. A Health Literacy Assessment of the epilepsy.com Website. *Seizure-European Journal of Epilepsy*, 18(6):434–439, 2009.
- [59] Eurobarometer. Flash Eurobarometer 404: European Citizens’ Digital Health Literacy. Technical Report 978-92-79-43607-9, European Commission, November 2014. Online at http://ec.europa.eu/public_opinion/flash/fl_404_en.pdf. Accessed on March, 2019.
- [60] Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. A Comparison of Features for Automatic Readability Assessment. In *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*, pages 276–284. Association for Computational Linguistics, 2010.
- [61] Ronald A Fisher. The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

- [62] Paul R. Fitzsimmons, Benedictg Michael, Joane L. Hulley, and G. Orville Scott. A Readability Assessment of Online Parkinson’s Disease Information. *The Journal of the Royal College of Physicians of Edinburgh*, 40(4):292–296, 2010.
- [63] Rudolf Flesch. How to Write Plain English. http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml, Accessed on March, 2019.
- [64] Edward A Fox and Joseph A Shaw. Combination of Multiple Searches. *NIST special publication SP*, 243, 1994.
- [65] Susannah Fox. Health Topics. Technical report, The Pew Internet & American Life Project, February 2011. Online at <http://www.pewinternet.org/2011/02/01/health-topics-2/>. Accessed on March, 2019.
- [66] Susannah Fox and Maeve Duggan. Health Online 2013. Technical report, The Pew Internet & American Life Project, January 2013. Online at <http://www.pewinternet.org/2013/01/15/health-online-2013/>. Accessed on March, 2019.
- [67] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An Efficient Boosting Algorithm for Combining Preferences. *Journal of Machine Learning Research*, 4:933–969, December 2003.
- [68] Jerome H Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, pages 1189–1232, 2001.
- [69] Norbert Fuhr. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum*, 51(3):32–41, 2017.
- [70] Norbert Fuhr, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Järvelin, Rosie Jones, Yiqun Liu, Josiane Mothe, Wolfgang Nejdl, Isabella Peters, and Benno Stein. An Information Nutritional Label for Online Documents. *SIGIR Forum*, 51(3):46–66, 2017.
- [71] Evgeniy Gabrilovich. Cura Te Ipsum: Answering Symptom Queries with Question Intent. In *Second WebQA workshop, SIGIR 2016 (invited talk), Pisa, Italy, July 17-21, 2016*, 2016. WebCitation: <http://www.webcitation.org/6yHTeM33k>.
- [72] Daniel Gayo-Avello. A Survey on Session Detection Methods in Query Logs and a Proposal for Future Evaluation. *Information Sciences*, 179(12):1822–1843, May 2009.
- [73] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely Randomized Trees. *Machine Learning*, 63(1):3–42, Apr 2006.
- [74] Lorraine Goeuriot, Gareth J. F. Jones, Liadh Kelly, Johannes Leveling, Mihai Lupu, João Palotti, and Guido Zuccon. An Analysis of Evaluation Campaigns in Ad-Hoc Medical Information Retrieval: CLEF eHealth 2013 and 2014. *Information Retrieval Journal*, pages 1–34, 2018.

- [75] Lorraine Goeuriot, Gareth JF Jones, Liadh Kelly, Johannes Leveling, Allan Hanbury, Henning Müller, Sanna Salanterä, Hanna Suominen, and Guido Zuccon. ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. In *Working Notes for CLEF 2013 Conference, Valencia, Spain, September 23-26, 2013*, volume 8138. Springer, 2013.
- [76] Lorraine Goeuriot, Liadh Kelly, Wei Li, João Palotti, Pavel Pecina, Guido Zuccon, Allan Hanbury, Gareth J. F. Jones, and Henning Müller. ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred Health Information Retrieval. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014*, pages 43–61, 2014.
- [77] Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Leif Hanlen, Aurélie Névél, Cyril Grouin, João Palotti, and Guido Zuccon. Overview of the CLEF eHealth Evaluation Lab 2015. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings*, pages 429–443, 2015.
- [78] Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Aurélie Névél, Aude Robert, Evangelos Kanoulas, Rene Spijker, João Palotti, and Guido Zuccon. CLEF 2017 eHealth Evaluation Lab Overview. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings*. Lecture Notes in Computer Science (LNCS), Springer, 2017.
- [79] Lorraine Goeuriot, Liadh Kelly, Guido Zuccon, and João Palotti. Building Evaluation Datasets for Consumer-Oriented Information Retrieval. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA), 2016.
- [80] Yoav Goldberg. Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017.
- [81] Mark A Graber, Cathy M Roller, and Betsy Kaeble. Readability Levels of Patient Education Material on the World Wide Web. *Journal of Family Practice*, 48(1):58–59, 1999.
- [82] Arthur Graesser, Danielle McNamara, Max Louwerse, and Zhiqiang Cai. Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavior Research Methods, Instruments, & Computers*, 2004.
- [83] Robert Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.
- [84] Manish Gupta, Michael Bendersky, et al. Information Retrieval with Verbose Queries. *Foundations and Trends® in Information Retrieval*, 9(3-4):209–354, 2015.

- [85] Allan Hanbury, Célia Boyer, Manfred Gschwandtner, and Henning Müller. Khresmoi: Towards a Multi-lingual Search and Access System for Biomedical Information. *Med-e-Tel, Luxembourg*, 2011:412–416, 2011.
- [86] Stephen P. Harter. Psychological Relevance and Information Science. *Journal of the American Society for information Science*, 43(9):602, 1992.
- [87] Daqing He and Ayse Göker. Detecting Session Boundaries from Web User Logs. In *Proceedings of the BCS-IRSG 22nd annual colloquium on information retrieval research*, pages 57–66, 2000.
- [88] Daqing He, Ayşe Göker, and David J Harper. Combining Evidence for Automatic Web Session Identification. *Information Processing & Management*, 38(5):727–742, 2002.
- [89] Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467, 2007.
- [90] William Hersh. *Information Retrieval: a Health and Biomedical Perspective*. Springer Science & Business Media, 2008.
- [91] Jorge Herskovic, Len Tanaka, William Hersh, and Elmer Bernstam. A Day in the Life of PubMed: Analysis of a Typical Day’s Query Log. *Journal of the American Medical Informatics Association*, 14(2):212–220, 2007.
- [92] Vera Hollink, Theodora Tsirikika, and Arjen P. de Vries. Semantic Search Log Analysis: A Method and a Study on Professional Image Search. *Journal of the American Society for Information Science and Technology*, 62(4):691–713, 2011.
- [93] Ingrid Hsieh-Yee. Effects of Search Experience and Subject Knowledge on the Search Tactics of Novice and Experienced Searchers. *Journal of the Association for Information Science and Technology*, 44(3):161–174, 1993.
- [94] Jeff Huang and Efthimis N Efthimiadis. Analyzing and Evaluating Query Reformulation Strategies in Web Search Logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 77–86. ACM, 2009.
- [95] Matthew Hutson et al. Even Artificial Intelligence Can Acquire Biases Against Race and Gender. *Science Magazine*, 10, 2017.
- [96] Rezarta Islamaj Dogan, G. Craig Murray, Aurélie Névéol, and Zhiyong Lu. Understanding PubMed® User Search Behavior through Log Analysis. *Database*, 2009:bap018, January 2009.

- [97] Ashutosh Sopan Jadhav, Amit P. Sheth, and Jyotishman Pathak. Online Information Searching for Cardiovascular Diseases: An Analysis of Mayo Clinic Search Query Logs. *Studies in Health Technology and Informatics*, pages 702–706, 2014.
- [98] Bernard J. Jansen and Amanda Spink. How are We Searching the World Wide Web?: a Comparison of Nine Search Engine Transaction Logs. *Information Processing & Management*, 42(1):248–263, January 2006.
- [99] Bernard J. Jansen, Amanda Spink, Judy Bateman, and Tefko Saracevic. Real Life Information Retrieval: A Study of User Queries on the Web. *SIGIR Forum*, 32(1):5–17, April 1998.
- [100] Bernard J Jansen, Amanda Spink, Chris Blakely, and Sherry Koshman. Defining a Session on Web Search Engines. *Journal of the American Society for Information Science and Technology*, 58(6):862–871, 2007.
- [101] Bernard J. Jansen, Amanda Spink, and Isak Taksai. *Handbook of Research on Web Log Analysis*. Information Science Reference - IGI Global Publishing, Hershey, PA, 2008.
- [102] Kalervo Järvelin and Jaana Kekäläinen. Cumulated Gain-based Evaluation of IR Techniques. *ACM Transactions on Information Systems*, 20(4):422–446, October 2002.
- [103] Jimmy, Guido Zuccon, João Palotti, Lorraine Goëuriot, and Liadh Kelly. Overview of the CLEF 2018 Consumer Health Search Task. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, 2018.
- [104] Rosie Jones and Kristina Lisa Klinkner. Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*, pages 699–708. ACM, 2008.
- [105] Liadh Kelly, Sebastian Dungs, Sascha Kriewel, Allan Hanbury, Lorraine Goëuriot, Gareth J. F. Jones, Georg Langs, and Henning Müller. Khresmoi Professional: Multilingual, Multimodal Professional Medical Search. In *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, pages 754–758, 2014.
- [106] Liadh Kelly, Lorraine Goëuriot, Hanna Suominen, Aurélie Névéol, João Palotti, and Guido Zuccon. Overview of the CLEF eHealth Evaluation Lab 2016. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, pages 255–266, 2016.

- [107] Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Tobias Schreck, Gondy Leroy, Danielle L. Mowery, Sumithra Velupillai, Wendy W. Chapman, David Martínez, Guido Zuccon, and João Palotti. Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings*, pages 172–191, 2014.
- [108] Hyeoneui Kim, Sergey Goryachev, Graciela Rosemblat, Allen Browne, Alla Kesselman, and Qing Zeng-Treitler. Beyond Surface Characteristics: a new Health Text-specific Readability Measurement. In *AMIA Annual Symposium Proceedings*, volume 2007, page 418. American Medical Informatics Association, 2007.
- [109] Jin Young Kim, Kevyn Collins-Thompson, Paul N. Bennett, and Susan T. Dumais. Characterizing Web Content, User Interests, and Search Behavior by Reading Level and Topic. In *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012, Seattle, WA, USA, February 8-12, 2012*, pages 213–222. ACM, 2012.
- [110] Peter Kincaid, Robert Fishburne, Richard Rogers, and Brad Chissom. Derivation of New Readability Formulas for Navy Enlisted Personnel. Technical report, 1975. Online at <https://stars.library.ucf.edu/cgi/viewcontent.cgi?&article=1055&context=istlibrary>. Accessed on March, 2019.
- [111] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate Detection Using Shallow Text Features. In *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*, pages 441–450, 2010.
- [112] Bevan Koopman and Guido Zuccon. Relevation!: An Open Source System for Information Retrieval Relevance Assessment. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 1243–1244. ACM, 2014.
- [113] Marlene Kritz, Manfred Gschwandtner, Veronika Stefanov, Allan Hanbury, and Matthias Samwald. Utilization and Perceived Problems of Online Medical Resources and Search Tools Among Different Groups of European Physicians. *Journal of Medical Internet Research*, 15(6), Jun 2013.
- [114] Udo. Kruschwitz and Charlie Hull. Searching the Enterprise. *Foundations and Trends in Information Retrieval*, 11(1):1–142, 2017.
- [115] Udo Kruschwitz, Deirdre Lungley, M-Dyaa Albakour, and Dawei Song. Deriving Query Suggestions for Site Search. *Journal of the American Society for Information Science and Technology*, 64(10):1975–1994, 2013.

- [116] Udo Kruschwitz, Nick Webb, and Richard Sutcliffe. Query Log Analysis for Adaptive Dialogue-driven Search. In *Handbook of Research on Web Log Analysis*, pages 389–414. IGI Global, 2009.
- [117] Eve-Marie Lacroix and Robert Mehnert. The US National Library of Medicine in the 21st Century: Expanding Collections, Nontraditional Formats, New Audiences. *Health Information and Libraries Journal*, 19(3):126–132, 2002.
- [118] David Lane. Online Statistics Education: A Multimedia Course of Study. In *EdMedia: World Conference on Educational Media and Technology*, pages 1317–1320. Association for the Advancement of Computing in Education (AACE), 2003.
- [119] Tessa Lau and Eric Horvitz. Patterns of Search: Analyzing and Modeling Web Query Refinement. In *UM99 User Modeling*, pages 119–128. Springer, 1999.
- [120] Gondy Leroy, Stephen Helmreich, James R Cowie, Trudi Miller, and Wei Zheng. Evaluating Online Health Information: Beyond Readability Formulas. In *AMIA Annual Symposium Proceedings*, volume 2008, page 394. American Medical Informatics Association, 2008.
- [121] Philip Ley and Tony Florio. The use of Readability Formulas in Health Care. *Psychology, Health & Medicine*, 1(1):7–28, 1996.
- [122] Aldo Lipani, Mihai Lupu, João Palotti, Guido Zuccon, and Allan Hanbury. Fixed Budget Pooling Strategies based on Fusion Methods. In *Proceedings of the Symposium on Applied Computing, SAC 2017, Marrakech, Morocco, April 3-7, 2017*, pages 919–924, 2017.
- [123] Aldo Lipani, João Palotti, Mihai Lupu, Florina Piroi, Guido Zuccon, and Allan Hanbury. Fixed-Cost Pooling Strategies Based on IR Evaluation Measures. In *Advances in Information Retrieval - 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings*, pages 357–368, 2017.
- [124] Tie-Yan Liu. Learning to Rank for Information Retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.
- [125] Xiaoyong Liu, W. Bruce Croft, Paul Oh, and David Hart. Automatic Recognition of Reading Levels from User Queries. In *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, pages 548–549. ACM, 2004.
- [126] Tian Lu, Yunjie Calvin Xu, and Scott Wallace. Internet Usage and Patient’s Trust in Physician during Diagnoses: A Knowledge Power Perspective. *Journal of the Association for Information Science and Technology*, 69(1):110–120, 2018.
- [127] Marco Lui and Timothy Baldwin. Langid.Py: An Off-the-shelf Language Identification Tool. In *The 50th Annual Meeting of the Association for Computational*

- Linguistics, Proceedings of the System Demonstrations, July 10, 2012, Jeju Island, Korea*, pages 25–30. Association for Computational Linguistics, 2012.
- [128] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.
- [129] G Harry Mc Laughlin. SMOG Grading - A New Readability Formula. *Journal of Reading*, 12(8):639–646, 1969.
- [130] David McDaid and A-La Park. Online Health: Untangling the web. Evidence from the BUPA Health Pulse 2010 International Healthcare Survey. Technical report, Bupa Health, 2011. Online at https://www.bupa.com.au/staticfiles/Bupa/HealthAndWellness/MediaFiles/PDF/LSE_Report_Online_Health.pdf. Accessed on March, 2019.
- [131] Emma Meats, Jon Brassey, Carl Heneghan, and Paul Glasziou. Using the Turning Research Into Practice (TRIP) database: how do clinicians really search? *Journal of the Medical Library Association*, 95(2):156–63, 2007.
- [132] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. Auditing Search Engines for Differential Satisfaction Across Demographics. In *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, WWW, pages 626–633, 2017.
- [133] Andrew Meillier and Shyam Patel. Readability of Healthcare Literature for Gastro-paresis and Evaluation of Medical Terminology in Reading Difficulty. *Gastroenterology Research*, 10(1):1–5, 2017.
- [134] Alistair Moffat and Justin Zobel. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems*, 27(1):2:1–2:27, 2008.
- [135] National Library of Medicine. UMLS Reference Manual. Technical report, 2009. Online at <https://www.ncbi.nlm.nih.gov/books/NBK9676/>. Accessed on March, 2019.
- [136] Aurélie Névéol, Rezarta Islamaj Dogan, and Zhiyong Lu. Semi-automatic Semantic Annotation of PubMed Queries: A Study on Quality, Efficiency, Satisfaction. *Journal of Biomedical Informatics*, 44(2):310–318, 2011.
- [137] Aurélie Névéol, Won Kim, W. John Wilbur, and Zhiyong Lu. Exploring Two Biomedical Text Genres for Disease Recognition. In *Proceedings of the BioNLP Workshop, BioNLP@HLT-NAACL 2009, Boulder, Colorado, USA, June 4-5, 2009*, pages 144–152. Association for Computational Linguistics, 2009.

- [138] Heung-Seon Oh, Yuchul Jung, and Kwang-Young Kim. KISTI at CLEF eHealth 2015 Task 2. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*, 2015.
- [139] Higgins Orla, Sixsmith Jane, Barry Margaret, and Domegan Christine. A Literature Review on Health Information-seeking Behaviour on the Web: A Health Consumer and Health Consumer and Health Professional Perspective. Technical report, European Centre for Disease Prevention and Control (ECDC), Stockholm, 2011. Online at <https://ecdc.europa.eu/sites/portal/files/media/en/publications/Publications/Literature%20review%20on%20health%20information-seeking%20behaviour%20on%20the%20web.pdf>. Accessed on March, 2019.
- [140] João Palotti, Veronika Stefanov, and Allan Hanbury. User Intent Behind Medical Queries: An Evaluation of Entity Mapping Approaches with Metamap and Freebase. In *Fifth Information Interaction in Context Symposium, IIX '14, Regensburg, Germany, August 26-29, 2014*, pages 283–286. ACM, 2014.
- [141] João Palotti, Guido Zuccon, and Allan Hanbury. The Influence of Pre-processing on the Estimation of Readability of Web Documents. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 1763–1766. ACM, 2015.
- [142] João Palotti. Beyond Topical Relevance: Studying Understandability and Reliability in Consumer Health Search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, page 1167, 2016.
- [143] João Palotti. Learning to Rank for Personalized E-Commerce Search at CIKM Cup 2016. Technical report, CIKM CUP 2016, Indianapolis, IN, USA, October 24-28, 2016, 2016.
- [144] João Palotti. Leveraging Wikipedia’s article structure to build search agents. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*.
- [145] João Palotti, Lorraine Goeuriot, Guido Zuccon, and Allan Hanbury. Ranking Health Web Pages with Relevance and Understandability. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 965–968. ACM, 2016.
- [146] João Palotti and Allan Hanbury. TUW @ TREC Clinical Decision Support Track 2015. In *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*, 2015.

- [147] João Palotti, Allan Hanbury, and Henning Muller. Exploiting Health Related Features to Infer User Expertise in the Medical Domain. In *Proceedings of the Workshop on Web Search Click Data (WSCD) co-located with the 7th ACM International Conference on Web Search and Data Mining (WSDM), New York, NY, USA, February 24-28, 2014*.
- [148] João Palotti, Allan Hanbury, Henning Müller, and Charles E. Kahn Jr. How Users Search and What They Search for in the Medical Domain - Understanding Laypeople and Experts Through Query Logs. *Information Retrieval Journal*, 19(1-2):189–224, 2016.
- [149] João Palotti and Navid Rekabsaz. Exploring Understandability Features to Personalize Consumer Health Search: TUV at CLEF 2017 eHealth. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*, 2017.
- [150] João Palotti, Navid Rekabsaz, Linda Andersson, and Allan Hanbury. TUV @ TREC Clinical Decision Support Track. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*, 2014.
- [151] João Palotti, Navid Rekabsaz, Mihai Lupu, and Allan Hanbury. TUV @ Retrieving Diverse Social Images Task 2014. In *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, 2014*.
- [152] Joao Palotti, Harris Scells, and Guido Zuccon. TrecTools: an Open-source Python Library for Information Retrieval Practitioners Involved in TREC-like Campaigns. In *Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Paris, France, July 21-25, 2009*. ACM, 2009.
- [153] João Palotti, Guido Zuccon, Johannes Bernhardt-Melischnig, Allan Hanbury, and Lorraine Goeuriot. Assessors Agreement: A Case Study Across Assessor Type, Payment Levels, Query Variations and Relevance Dimensions. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, pages 40–53, 2016.
- [154] João Palotti, Guido Zuccon, Lorraine Goeuriot, Liadh Kelly, Allan Hanbury, Gareth J. F. Jones, Mihai Lupu, and Pavel Pecina. ShARe/CLEF eHealth Evaluation Lab 2015, Task 2: User-centred Health Information Retrieval. In *Working Notes for CLEF 2015 Conference, Toulouse, France, September 8-11, 2015*, 2015.
- [155] João Palotti, Guido Zuccon, Jimmy, Pavel Pecina, Mihai Lupu, Lorraine Goeuriot, Liadh Kelly, and Allan Hanbury. CLEF 2017 Task Overview: The IR Task at the eHealth Evaluation Lab - Evaluating Retrieval Methods for Consumer Health

- Search. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*, 2017.
- [156] João Palotti, Guido Zuccon, and Allan Hanbury. MM: A new Framework for Multidimensional Evaluation of Search Engines. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*. ACM, 2018.
- [157] João Palotti, Guido Zuccon, and Allan Hanbury. Consumer Health Search on the Web: Study of Web Page Understandability and Its Integration in Ranking Algorithms. *Journal of Medical Internet Research*, 21(1):e10986, 2019.
- [158] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2(1-2):1–135, 2008.
- [159] Taemin Kim Park. The Nature of Relevance in Information Retrieval: An Empirical Study. *The Library Quarterly: Information, Community, Policy*, 63(3):318–351, 1993.
- [160] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of Search. In *Proceedings of the 1st International Conference on Scalable Information Systems, Infoscale 2006, Hong Kong, May 30-June 1, 2006*. ACM, 2006.
- [161] Chirag R. Patel, Deepa V. Cherla, Saurin Sanghvi, Soly Baredes, and Jean Anderson Eloy. Readability Assessment of Online Thyroid Surgery Patient Education Materials. *Head & Neck*, 35(10):1421–1425, 2013.
- [162] Emily Pitler and Ani Nenkova. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 186–195. Association for Computational Linguistics, 2008.
- [163] Frances A Pogacar, Amira Ghenai, Mark D Smucker, and Charles LA Clarke. The Positive and Negative Influence of Search Results on People’s Decisions about the Efficacy of Medical Treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2017, Amsterdam, The Netherlands, October 1-4, 2017*, pages 209–216. ACM, 2017.
- [164] Jan Pomikálek. *Removing Boilerplate and Duplicate Content from Web Corpora*. PhD thesis, Masaryk University, Czech Republic, 2011.
- [165] Wanda Pratt and Meliha Yetisgen-Yildiz. A Study of Biomedical Concept Identification: MetaMap vs. People. In *AMIA Annual Symposium Proceedings*, volume 2003, pages 529–533. American Medical Informatics Association, 2003.
- [166] Leonard Richardson. BeautifulSoup V4.4. <https://www.crummy.com/software/BeautifulSoup/>, 2017. Accessed on March, 2019.

- [167] Kirk Roberts, Matthew Simpson, Dina Demner-Fushman, Ellen Voorhees, and William Hersh. State-of-the-art in Biomedical Literature Retrieval for Clinical Cases: a Survey of the TREC 2014 CDS Track. *Information Retrieval Journal*, 19(1):113–148, 2016.
- [168] Kirk Roberts, Matthew S. Simpson, Ellen M. Voorhees, and William R. Hersh. Overview of the TREC 2015 Clinical Decision Support Track. In *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*, 2015.
- [169] Serwah Sabetghadam, João Palotti, Navid Rekabsaz, Mihai Lupu, and Allan Hanbury. TUW @ MediaEval 2015 Retrieving Diverse Social Images Task. In *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15, 2015*, 2015.
- [170] Tetsuya Sakai. Alternatives to Bpref. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 71–78. ACM, 2007.
- [171] Mark Sanderson. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends® in Information Retrieval*, 4(4):247–375, 2010.
- [172] Tefko Saracevic. Relevance: A Review of and a Framework for the Thinking on the Notion in Information Science. *Journal of the Association for Information Science and Technology*, 26(6):321–343, 1975.
- [173] Linda Schamber. Relevance and Information Behavior. *Annual review of information science and technology (ARIST)*, 29:3–48, 1994.
- [174] Julia Schwarz and Meredith Morris. Augmenting Web Pages and Search Results to Support Credibility Assessment. In *Proceedings of the International Conference on Human Factors in Computing Systems, CHI 2011, Vancouver, BC, Canada, May 7-12, 2011*, pages 1245–1254. ACM, 2011.
- [175] Sarah J. Shoemaker, Michael S. Wolf, and Cindy Brach. Development of the Patient Education Materials Assessment Tool (PEMAT): a new Measure of Understandability and Actionability for Print and Audiovisual Patient Information. *Patient Education and Counseling*, 96(3):395–403, 2014.
- [176] Luo Si and Jamie Callan. A Statistical Model for Scientific Readability. In *Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management, Atlanta, Georgia, USA, November 5-10, 2001*, pages 574–576. ACM, 2001.
- [177] Ricardo Silva and Carla Lopes. The Effectiveness of Query Expansion when searching for Health related Content: InfoLab at CLEF eHealth 2016. In *Working*

Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016, pages 130–142, 2016.

- [178] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum*, 33(1):6–12, September 1999.
- [179] Fabrizio Silvestri. Mining Query Logs: Turning Search Usage Data into Knowledge. *Foundations and Trends in Information Retrieval*, 4(1:2):1–174, January 2010.
- [180] E. A. Smith and R. J. Senter. *Automated Readability Index*. AMRL-TR-66-220. Aerospace Medical Research Laboratories, 1967.
- [181] Luca Soldaini, Arman Cohan, Andrew Yates, Nazli Goharian, and Ophir Frieder. *Retrieving Medical Literature for Clinical Decision Support*, pages 538–549. Springer International Publishing, 2015.
- [182] Luca Soldaini, Will Edman, and Nazli Goharian. Team GU-IRLAB at CLEF eHealth 2016: Task 3. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016*, pages 143–146, 2016.
- [183] Yang Song, Yun He, Qinmin Hu, Liang He, and E. Mark Haacke. ECNU at 2015 eHealth Task 2: User-centred Health Information Retrieval. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*, 2015.
- [184] Yang Song, Yun He, Hongyu Liu, Yueyao Wang, Qinmin Hu, and Liang He. ECNU at 2016 eHealth Task 3: Patient-centred Information Retrieval. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016*, pages 157–161, 2016.
- [185] Amanda Spink, Yin Yang, Jim Jansen, Pirrko Nykanen, Daniel P. Lorence, Seda Ozmutlu, and Cenk Ozmutlu. A Study of Medical and Health Queries to Web Search Engines. *Health Information & Libraries Journal*, 21(1):44–51, March 2004.
- [186] Isabelle Stanton, Samuel Ieong, and Nina Mishra. Circumlocution in Diagnostic Medical Queries. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast, QLD, Australia - July 06 - 11, 2014*, pages 133–142. ACM, 2014.
- [187] Lauren M. Stossel, Nora Segar, Peter Gliatto, Robert Fallar, and Reena Karani. Readability of Patient Education Materials Available at the Point of Care. *Journal of general internal medicine*, 27(9):1165–1170, 2012.
- [188] Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, Aurélie Névéol, Lionel Ramadier, Aude Robert, Evangelos Kanoulas, Rene Spijker, Leif Azzopardi, Dan Li, Jimmy Jimmy, João Palotti, and Guido Zuccon. Overview of the CLEF eHealth Evaluation

- Lab 2018. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings*, pages 286–301. Springer, 2018.
- [189] Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy Webber Chapman, Guergana K. Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martínez, and Guido Zuccon. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings*, pages 212–231, 2013.
- [190] Don R. Swanson. Subjective Versus Objective Relevance in Bibliographic Retrieval Systems. *The Library Quarterly*, 56(4):389–398, 1986.
- [191] Chenhao Tan, Evgeniy Gabrilovich, and Bo Pang. To Each His Own: Personalized Content Selection Based on Text Comprehensibility. In *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012, Seattle, WA, USA, February 8-12, 2012*, pages 233–242. ACM, 2012.
- [192] Amalia Todirascu, Thomas François, Nuria Gala, Cédric Fairon, Anne-Laure Ligozat, and Delphine Bernhard. Coherence and Cohesion for the Assessment of Text Readability. In *International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2013)*, page 9p, Marseille, France, 2013.
- [193] Elaine G. Toms and Celeste Latter. How Consumers Search for Health Information. *Health Informatics Journal*, 13(3):223–235, 2007.
- [194] Theodora Tsirikika, Henning Müller, and Charles Kahn Jr. Log Analysis to Understand Medical Professionals’ Image Searching Behaviour. *Studies in health technology and informatics*, 180:1020—1024, 2012.
- [195] John W Tukey. Comparing Individual Means in the Analysis of Variance. *Biometrics*, 5(2):99–114, 1949.
- [196] National Cancer Institute (U.S.). Clear & Simple: Developing Effective Print Materials for Low-literate Readers. <https://www.nih.gov/institutes-nih/nih-office-director/office-communications-public-liaison/clear-communication/clear-simple>, Accessed on March, 2019.
- [197] Joost van Doorn, Daan Odijk, Diederik M Roijers, and Maarten de Rijke. Balancing Relevance Criteria Through Multi-objective Optimization. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 769–772. ACM, 2016.
- [198] Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*, volume 1. The MIT Press, 2005.

- [199] Lorraine Silver Wallace and Elizabeth S Lennon. American academy of family physicians patient education materials: Can patients read them? *Family medicine*, 36(8):571–574, 2004.
- [200] Liupu Wang, Juexin Wang, Michael Wang, Yong Li, Yanchun Liang, and Dong Xu. Using Internet Search Engines to Obtain Medical Information: A Comparative Study. *Journal of Medical Internet Research*, 14(3):e74, May 2012.
- [201] Ruixue Wang, Wei Lu, and Ke Ren. WHUIRGroup at the CLEF 2016 eHealth Lab Task 3. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016*, pages 193–197, 2016.
- [202] Marc Weeber, Henny Klein, Alan R. Aronson, James G. Mork, Lolkje T. W. de Jong-van den Berg, and Rein Vos. Text-based Discovery in Biomedicine: the Architecture of the DAD-system. In *Proceedings of the AMIA Symposium*, pages 903–907, 2000.
- [203] Barry D. Weiss. Communicating with Patients who have Limited Literacy Skills. *Journal of Family Practice*, 46(2):168–176, 1998.
- [204] Gary M Weiss and Foster Provost. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. *Journal of artificial intelligence research*, 19:315–354, 2003.
- [205] Ryen White. Beliefs and Biases in Web Search. In *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013*, pages 3–12. ACM, 2013.
- [206] Ryen W. White. *Interacting with Search Systems*. Cambridge University Press, 2016.
- [207] Ryen W. White, Susan T. Dumais, and Jaime Teevan. Characterizing the Influence of Domain Expertise on Web Search Behavior. In *Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009*, pages 132–141. ACM, 2009.
- [208] Ryen W. White and Eric Horvitz. Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. *ACM Transactions on Information Systems*, 27(4):23:1–23:37, November 2009.
- [209] Ryen W. White and Eric Horvitz. Studies of the Onset and Persistence of Medical Concerns in Search Logs. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*, pages 265–274, 2012.
- [210] Constance Wiener and Regina Wiener-Pla. Literacy, Pregnancy and Potential Oral Health Changes: The Internet and Readability Levels. *Maternal and child health journal*, 18(3):657–662, 2014.

- [211] Barbara M. Wildemuth. The Effects of Domain Knowledge on Search Tactic Formulation. *Journal of the Association for Information Science and Technology*, 55(3):246–258, February 2004.
- [212] Jun Xu and Hang Li. AdaRank: A Boosting Algorithm for Information Retrieval. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 391–398. ACM, 2007.
- [213] Xin Yan, Raymond Y. Lau, Dawei Song, Xue Li, and Jian Ma. Toward a Semantic Granularity Model for Domain-specific Information Retrieval. *ACM Transactions on Information Systems*, 29(3):15:1–15:46, July 2011.
- [214] Xin Yan, Dawei Song, and Xue Li. Concept-based Document Readability in Domain Specific Information Retrieval. In *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management, Arlington, Virginia, USA, November 6-11, 2006*, pages 540–549. ACM, 2006.
- [215] Michele L. Ybarra and Michael Suman. Help Seeking Behavior and the Internet: a National Survey. *International journal of medical informatics*, 75(1):29–41, 2006.
- [216] Qing Zeng-Treitler, Sergey Goryachev, Tony Tse, Alla Keselman, and Aziz Boxwala. Estimating Consumer Familiarity with Health Terminology: a Context-based Approach. *Journal of the American Medical Informatics Association*, 15(3):349–356, 2008.
- [217] Qing Zeng-Treitler, Eunjung Kim, Jon Crowell, and Tony Tse. A Text Corpora-based Estimation of the Familiarity of Health Terminology. *Biological and Medical Data Analysis*, pages 184–192, 2005.
- [218] Qing Zeng-Treitler and Tony Tse. Exploring and Developing Consumer Health Vocabularies. *Journal of the American Medical Informatics Association*, 13(1):24–29, 2006.
- [219] Xiangmin Zhang, Michael Cole, and Nicholas Belkin. Predicting Users’ Domain Knowledge from Search Behaviors. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 1225–1226. ACM, 2011.
- [220] Yan Zhang. Searching for Specific Health-related Information in MedlinePlus: Behavioral Patterns and User Experience. *Journal of the Association for Information Science and Technology*, 65(1):53–68, 2014.
- [221] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. Multidimensional Relevance Modeling via Psychometrics and Crowdsourcing. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast, QLD, Australia - July 06 - 11, 2014*, pages 435–444. ACM, 2014.

- [222] Wei Zhou, Vetle Torvik, and Neil Smalheiser. ADAM: Another Database of Abbreviations in MEDLINE. *Bioinformatics*, 22(22):2813–2818, 2006.
- [223] Guido Zuccon. Understandability Biased Evaluation for Information Retrieval. In *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, pages 280–292. Springer, 2016.
- [224] Guido Zuccon and Bevan Koopman. Integrating Understandability in the Evaluation of Consumer Health Search Engines. In *Proceedings of the Medical Information Retrieval Workshop at SIGIR co-located with the 37th annual international ACM SIGIR conference (ACM SIGIR 2014), Gold Coast, Australia, July 11, 2014*, pages 32–35, 2014.
- [225] Guido Zuccon, Bevan Koopman, and João Palotti. Diagnose This If You Can. In *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*, pages 562–567, 2015.
- [226] Guido Zuccon, João Palotti, Lorraine Goeuriot, Liadh Kelly, Mihai Lupu, Pavel Pecina, Henning Mueller, Julie Budaher, and Anthony Deacon. The IR Task at the CLEF eHealth Evaluation Lab 2016: User-centred Health Information Retrieval. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016*, volume 1609, pages 15–27, 2016.
- [227] Guido Zuccon, João Palotti, and Allan Hanbury. Query Variations and their Effect on Comparing Information Retrieval Systems. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 691–700, 2016.