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.