A Framework for Enhancing the Query and Medical Record Representations for Patient Search

Nut Limsopatham

School of Computing Science, University of Glasgow, Glasgow, UK

nut.limsopatham@glasgow.ac.uk

Abstract

This thesis focuses on enhancing the search of electronic medical records (EMRs), with the aim of identifying patients with medical histories relevant to the medical conditions stated in a text query. During retrieval, a healthcare practitioner indicates a number of inclusion criteria describing the medical conditions of the patients of interest. However, finding patients with particular medical conditions is challenging, due to the implicit knowledge inherent within the patients' medical records and queries - such knowledge may be known by medical practitioners, but may be hidden from an information retrieval (IR) system. For instance, the mention of a treatment such as a drug may indicate to a practitioner that a particular diagnosis has been made for the patient, but this diagnosis may not be explicitly mentioned in the patient's medical records. Moreover, the use of negated language (e.g. 'without', 'no') to describe a medical condition of a patient (e.g. the patient has no fever) may cause a search system to erroneously retrieve that patient for a query when searching for patients with that medical condition (e.g. find patients with fever). To attain effective retrieval performance, we hypothesise that, in a patient search system, both the information needs and patients' histories should be represented based upon *the medical decision process*. In particular, this thesis argues that since the medical decision process typically encompasses four aspects (symptom, diagnostic test, diagnosis and treatment), a patient search system should take into account these aspects and apply inferences to recover the possible implicit knowledge. We postulate that considering these aspects and their derived implicit knowledge at three different levels of the retrieval process (namely, sentence, medical record and inter-record levels) enhances the retrieval performance. Indeed, we propose a novel framework that can gain insights from EMRs and queries, by modelling and reasoning upon information during retrieval in terms of the four aforementioned aspects at the three levels of the retrieval process, and can use these insights to enhance patient search.

Firstly, at the sentence level, we extract the medical conditions in the medical records and queries. In particular, we propose to represent only the medical conditions related to the four medical aspects in order to improve the accuracy of our search system. In addition, we identify the context (negative/positive) of terms, which leads to an accurate representation of the medical conditions both in the EMRs and queries. In particular, we aim to prevent patients whose EMRs state the medical conditions in the contexts different from the query

from being ranked highly. For example, preventing patients whose EMRs state "no history of dementia" from being retrieved for a query searching for patients with dementia.

Secondly, at the medical record level, using external knowledge-based resources (e.g. ontologies and health-related websites), we leverage the relationships between medical terms to infer the wider medical history of the patient in terms of the four medical aspects. In particular, we estimate the relevance of a patient to the query by exploiting association rules that we extract from the semantic relationships between medical terms using the four aspects of the medical process. For example, patients with a medical history involving a *CABG surgery* (treatment) can be inferred as relevant to a query searching for a patient suffering from *heart disease* (diagnosis), since a CABG surgery is a treatment of heart disease.

Thirdly, at the inter-record level, we enhance the retrieval of patients in two different manners. First, we exploit knowledge about how the four medical aspects are handled by different hospital departments to gain a better understanding about the appropriateness of EMRs created by different departments for a given query. We propose to aggregate EMRs at the department level (i.e. inter-record level) to extract implicit knowledge (i.e. the expertise of each department) and model this department's expertise, while ranking patients. For instance, patients having EMRs from the cardiology department are likely to be relevant to a query searching for patients who suffered from a heart attack. Second, as a medical query typically contains several medical conditions that the relevant patients should satisfy, we propose to explicitly model the relevance towards multiple query medical conditions in the EMRs related to a particular patient during retrieval. In particular, we rank highly those patients that match all the stated medical conditions in the query by adapting coverage-based diversification approaches originally proposed for the web search domain.

Finally, we examine the combination of our aforementioned approaches that exploit the implicit knowledge at the three levels of the retrieval process to further improve the retrieval performance by adapting techniques from the fields of data fusion and machine learning. In particular, data fusion techniques, such as CombSUM and CombMNZ, are used to combine the relevance scores computed by the different approaches of the proposed framework. On the other hand, we deploy state-of-the-art learning to rank approaches (e.g. LambdaMART and AdaRank) to learn from a set of training data an effective combination of the relevance scores computed by the approaches of the framework. In addition, we introduce a novel selective ranking approach that uses a classifier to effectively apply one of the approaches of the framework on a per-query basis.

This thesis draws insights from a thorough evaluation and analysis of the proposed framework using a standard test collection provided by the TREC Medical Records track. The experimental results show the effectiveness of the framework. In particular, the results demonstrate the importance of dealing with the implicit knowledge in patient search by focusing on the medical decision criteria aspects at the three levels of the retrieval process.

Supervisors: Iadh Ounis and Craig Macdonald

Available: `http://theses.gla.ac.uk/5983/1/2014LimsopathamPhD.pdf`