Computers in Industry 69 (2015) 81–91

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

Diversity-aware retrieval of medical records

Jianqiang Li [a,b,1], Chunchen Liu [c,1], Bo Liu [c], Rui Mao [a,*], Yongcai Wang [d], Shi Chen [f], Ji-Jiang Yang [e], Hui Pan [f], Qing Wang [e]

[a] GDPHPC Labs, Shenzhen University, Guangdong, China
[b] School of Software Engineering, Beijing University of Technology, Beijing, China
[c] NEC Labs, China
[d] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[e] Research Institute of Information and Technology, Tsinghua University, Beijing, China
[f] Department of Endocrinology, Peking Union Medical College Hospital, Chinese Academe of Medical Sciences & Peking Union Medical College, Beijing, China

ARTICLE INFO

Article history:
Received 30 January 2014
Received in revised form 27 June 2014
Accepted 25 September 2014
Available online 25 October 2014

Keywords:
Medical search
Query understanding
Search result diversification
Medical information retrieval

ABSTRACT

The widely adoption of Electronic Medical Records (EMRs) causes an explosive growth of the medical and clinical data. It makes the medical search technologies become critical to find useful patient information in the large medical dataset. However, the high quality medical search is a challenging task, in particular due to the inherent complexity and ambiguity of medical terminology. In this paper, by exploiting the uncertainty in ambiguous medical queries, we propose a novel semantic-based approach to achieve the diversity-aware retrieval of EMRs, i.e., both the relevance and novelty are considered for EMR ranking. With the support of medical domain ontologies, we first mine all the potential semantics (concepts and relations between them) from a user query and consume them to model the multiple query aspects. Then, we propose a novel diversification strategy, which considers not only the aspect importance but also the aspect similarity, to perform the diversity-aware EMR ranking. A real-world pilot study, which utilizes the proposed medical search approach to improve the second use of the EMRs, is reported. We believe that our experience can serve as an important reference for the development of similar applications in a medical data utilization and sharing environment.

1. Introduction

With the notion that EMRs are the bedrock of modern healthcare, EMR systems are widely deployed for the exchange of medical information among various healthcare related parties [1]. According to the recent surveys, 82 percent of physicians indicated that they are currently using an EMR system or plan to do so [3].

This EMR movement causes an explosive growth of the medical and clinical datasets [51]. Secondary use of EMR data relies on the ability to retrieve accurate and complete information about desired patient populations. This fact makes the medical search is becoming a critical technique for the rapid and effective access of patient information [51], which also provide great potential for facilitating research and improving quality in medical practice [49]. Along this trend, the medical records track is established in Text REtrieval Conference (TREC) [55,62].

Medical information retrieval is challenging because of the inherent ambiguity within the posed queries [62]. Such ambiguity is manifested in different ways: (1) A query expresses a clearly defined sense, but the genuine needs under this sense may cover a broad range. Taking a common scenario where an ordinary user performs medical search for example, he feels uncomfortable (he has a high fever and rash erupts on his body) but is uncertain about his exact medical problems, so he inputs "fever" and "rash" as keywords into a search engine. In this case, as many diseases may cause these symptoms, the user may prefer to learn knowledge about all these diseases, so as to have a preliminary understanding about his situation and better prepare for the interview with doctors. (2) Query terms themselves are ambiguous, as most users have little medical knowledge. For instance, a pregnant woman feels pain in her abdomen, so she submits a query composed of "pain in the abdomen" and "pregnant". In this case, the term "pain" is ambiguous, which may mean "stabbing pain", "distending pain", "labor pain", etc. The user-cared reasons causing these different kinds of pain, however, may be totally different. Considering the ambiguous queries, as users provide no more information for disambiguating their intents, a medical search engine should

* Corresponding author. Tel.: +86 755 2653 4207x81.
E-mail address: mao@szu.edu.cn (R. Mao).
[1] These authors contributed equally to this paper.

http://dx.doi.org/10.1016/j.compind.2014.09.004
0166-3615/

produce a set of diversified results that cover all possible intents implied by the given query, in order to enable users to find their interested medical information.

From technical point of view, traditional IR technologies can be classified into two categories, i.e., content-based [25,27,28] and semantic-based [14,20,21,38,40,47] approaches. The former predicts the relevance of a document to the given query by considering only the document-inside content. Due to the inherent complexity and ambiguity of the medical terminologies, it is inappropriate for applying it directly for medical search [61]. The later exploits the external semantic resources to improve IR quality by taking into account the meaning of terms as they appear in the query and documents. Since the semantic-based approaches provide great potentials to tackle the ambiguous medical queries and the complexity of medical terminologies, current researches on medical information retrieval mainly fall into this category, such as query expansion [51,56,58], semantic similarity calculation [36,57,54], and granularity match [36,52,61]. However, since these IR models rank each document independently, the resulting top-ranked documents often contain excessively redundant information. The fact that they consider the relevance as the only measure for medical record ranking makes their capability to handle the query ambiguity is limited [62].

Recently, search result diversification is becoming a hot research topic with the aim to minimize the risk of dissatisfaction of the average user [8,10,24]. Since the novelty is introduced as an additional measure for document ranking, it has shown as a promising way to tackle ambiguous queries [6–10]. Based on how the different query aspects underlying the user input query are accounted for, existing approaches can be categorized as either implicit or explicit ones [9,10]. The theoretical analysis and experimental study [8,10] has illustrated that the explicit approaches, i.e., to explicitly model the possible aspects underlying a query, are more effective than the implicit ones. Along this trend, this paper will focus on the diversity-aware retrieval of EMRs, where both relevance and novelty are taken into account for the medical document ranking.

Since almost all the existing explicit diversification methods are developed for the scenario of Web search, it is not appropriate to deploy them in the medical search setting: (1) Query log, which is adopted for query aspect modeling by existing methods, is not a reliable resource for medical data retrieval. On the one hand, for some medical search environments such as enterprise search, query logs are not available or their scale is not large enough for supporting query reformulation. On the other hand, most users have no background knowledge and thus input queries at will, which leads the query aspects derived from query logs to be inaccurate. (2) Unlike the Web search that covers a wide range of application domains, medical search focuses on a concrete domain in which the rich medical knowledge is available. This domain knowledge is essential for a well-functioning diversifying model, which contributes to improve query aspect identifying accuracy and comprehensiveness [8]. However, none of the existing approaches use the domain knowledge. (3) Similarity between query aspects is an important factor for search diversification, but it is ignored by existing methods. Considering an instance in which four aspects are identified for a given query, with the first three express similar topics but the fourth one expresses a new idea. Using the existing aspect similarity measuring methods, a ranking, where documents about the fourth aspect are excluded from top positions, will be produced. Nevertheless, the documents related to the fourth aspect are more novel and deserve upper positions.

In this paper, we propose a novel approach to achieve the diversity-aware retrieval of medical records, where the semantic-based IR and search result diversification are combined together to tackle inherent ambiguity of the medical search. Different from existing diversifying strategies relying heavily on large amounts of query logs, the proposed approach employs a medical ontology that comprises rich medical knowledge to disambiguate the original query into multiple sub-queries (or query aspects). Each sub-query represents one aspect of the implied intents of the original query. Based on the modeled aspects of the sub-queries, we gives a novel strategy that exploiting the query disambiguation results for the diversity-aware medical search. The performance of the proposed approach is demonstrated on a real-world medical dataset. Experiment results show that the proposed approach fits well for the medical search environments and outperforms existing methods on both diversity and accuracy.

The contribution of this paper can be summarized as follows: (1) A novel approach for exploiting the ambiguity in a medical query for diversity-aware medical search is proposed, which first employs the medical domain knowledge for query understanding to construct multiple sub-queries from the original query and then the medical record relevance and novelty are combined together to handle the uncertainty in the information needs; (2) The empirical experiments on the real-world dataset are reported, which demonstrate the effectiveness of the proposed approaches; (3) A pilot study is described for the application of the proposed medical search approach in a real-world usage scenario.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 describes the technical details of the proposed approach for diversity-aware retrieval of medical records. In Section 4, the evaluation results and the real-world application are reported. Section 5 concludes the paper.

2. Related work

Information Retrieval (IR) is the process of searching within a document collection for the information most relevant to a user's query. It mainly uses keyword-based query as an input and returns a list of relevant documents as the output [25,47].

Most searching systems running for traditional document collections use content-based approaches, e.g., the vector space model [25], Latent Semantic Indexing [27], or Nonnegative Matrix Factorizations [28]. Since only the internal information of a document is employed to measure the similarity between queries and documents, they are not applicable to handle the complexity of the medical terminologies [55].

While in traditional IR only the document content is of concern for the query-dependent ranking, in web page retrieval, the link structure of the Web also plays an important role for the query-independent ranking. Current popular models for web page retrieval are mainly combinations of content-based and hyperlink-based approaches [29,30], where the location of a Web page in the Web's graph structure to determine its importance. Based on the assumption that hyperlinks in the global Web have the semantics of recommendations, hyperlink-based approaches utilize the location of a Web page in the Web's graph structure to determine its importance [30]. By observing that the majority of the links in a website are used to organize information and convey no recommendations, a path-based method for Web page ranking is described in [26,29] by distinguishing the hyperlinks for recommendation and information organization, respectively. The existing of inter-hyperlink relationship between documents is the pre-condition for the utilization of the hyperlink-based approaches. This fact limits their application for IR on general document collections.

Different from content-based IR approaches that focus on the frequency of word appearance, semantic-based IR methods more likely tend to understand the meaning hidden in retrieved documents and users' queries, by means of adding semantic tags into texts [22,39,44], structuring and conceptualizing the objects