Computers in Biology and Medicine 101 (2018) 39–50

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Exploiting the semantic graph for the representation and retrieval of medical documents

Qing Zhao, Yangyang Kang[1], Jianqiang Li*, Dan Wang

Faculty of Information Technology, Beijing University of Technology, Beijing, China

ARTICLE INFO

Keywords:
Semantic information retrieval
Medical search
Document ranking
Electronic medical records

ABSTRACT

Objective: The objective of this study was to propose a graph-based semantic search approach by addressing the inherent complexity and ambiguity of medical terminology in queries and clinical text for enhanced medical information retrieval.

Methods: The supportive use of a medical domain ontology exploits the light-weight semantics discovered from queries and documents for enhanced document ranking. First, the implicit information regarding concepts and the relations between them is discovered in the documents and queries and is used to evaluate the relevance of the query-document; then, the semantic linkages between concepts distributed in target documents and reference documents are built and used to score the document's popularity; finally, the above two evaluations are integrated to produce the final ranking list for document ranking.

Results: Empirical experiments are conducted on two different datasets. The results demonstrate that the proposed graph-based approach significantly outperforms the baselines. For example, the average performance improvement on two datasets of the best variant of GSRM compared to the best baseline achieve 7.2% and 7.9% in terms of *P@20* and *NDCG@20*, respectively, which illustrates the effectiveness of the proposed approach.

1. Introduction

With the wide adoption of Electronic Medical Records (EMRs) systems and the continued increase in medical documentation being provided online, medical information retrieval is becoming a hot research topic since it is critical for enabling users to find useful patient information rapidly and effectively in large medical and clinical dataset [1].

Traditional information retrieval models mainly use two factors for document ranking, query-document relevance and document popularity. Query-document relevance measures how well a document matches the query submitted by the user, which is important for achieving a good search result since acquiring relevant results is the users' basic requirement. Many methods, such as vector space models [2–4], probabilistic rank models [7–9], and learning-based rank models [10–12,15], support relevance computing. Document popularity acts as a complementary factor that is also significant for ranking because there are usually so many documents that match the query that it becomes impossible for users to review them all; therefore, only the popular (important or authoritative) documents will be reviewed. The representative approaches comprise PageRank [5], HITS [6] and others.

However, until now, it has been quite clear that the traditional information retrieval technologies perform poorly when they are employed in healthcare. This finding is mainly caused by the complex and inherent ambiguity of the data or information in the medical domain. There are some characteristics in medical information retrieval that determine the following [1,16,18,19,24]: (1) A query's expression is fuzzy when the information inquirers are parents or non-medical professional users; (2) Query terms (such as anatomy and morphology) are ambiguous by themselves because most users have little medical knowledge. For example, a user feels pain in her abdomen, and as a result, she/he submits a query about 'pain in the abdomen'. In this case, the term 'pain' is ambiguous, which could mean 'stabbing pain', 'distending pain', 'labour pain', and so on. In another example, a user entered a query ‘eye infection’, which also has two meanings, ‘bacterial eye infections’ and ‘fungal eye infections’. Thus, finding an appropriate ranking approach to use in a medical search is a central problem.

In this paper, we propose a Graph-based Semantic Ranking Model (GSRM) for practical medical searches. Basically, our method uses domain knowledge as support to exploit the lightweight semantics

* Corresponding author.
E-mail addresses: zhaoqing1025@emails.bjut.edu.cn (Q. Zhao), kangyangemail@163.com (Y. Kang), lijianqiang@bjut.edu.cn (J. Li), wangdan@bjut.edu.cn (D. Wang).
[1] Co-first author.

https://doi.org/10.1016/j.compbiomed.2018.08.009
Received 10 May 2018; Received in revised form 1 August 2018; Accepted 5 August 2018
0010-4825/ © 2018 Elsevier Ltd. All rights reserved.

discovered from the query and documents, to enhance the document ranking. Our model is different from other ranking strategies because GSRM not only uses concepts defined in an ontology but also mines the implicit relations between them from the document queries, employing them for document understanding and relevance computing. In addition, without using hyperlinks, our model develops a novel method to compute the document's popularity, which will be used in conjunction with the document relevance to further improve the accuracy of the ranking results. We incorporate an open dataset and an internal dataset to demonstrate the performance of the GSRM. The experimental results clearly show that compared with the existing models, our proposed novel ranking approach GSRM performs better in terms of the ranking accuracy.

The remainder of this study is organized as follows: Section 2 summarizes the related works. Section 3 introduces details about our semantic-based synthetic rank model. The experiments are reported in Section 4, where the setup of the experiment is illustrated and the test results are discussed. Section 5 discusses the previous study and the similar wok in the literature, and the limitations of this study. Section 6 concludes with a summary of our research and directions for future work.

2. Related work

Traditional Information Retrieval (IR) strategies employ statistics for words in document text to calculate query-document relevance on which to base the rank. For example, the classic vector space model [2] constructs vectors based on term statistical information to represent documents and queries, and it calculates the vector, including the angle cosine, to evaluate the document relevance; probabilistic rank models [7–9] exploit the term distribution probability in a document set to estimate the document relevancy probability with which to decide the ranking list; the learning-based rank models [10–12] employ machine learning methods to construct ranking functions from training data, which help to compute the document relevance.

Later, as the web matured, an enormous number of medical documents of varying quality in traditional IR became available, and researchers found that the classic relevance-only retrieval strategies performed poorly in this environment [5,6,17]; thus, the document importance was introduced for the IR ranking. The earliest work was Google's PageRank algorithm, which computes the page importance by analysing the hyperlink structure of the web, with the assumption that the larger the number of recommendation hyperlinks a web page has, the more authoritative it is [5]. Another representative work is the Hyperlink-Induced Topic Search algorithm [6], which considers not only the authority of a page but also its role as a hub. Additionally, there are many similar methods that have been proposed by various researchers, such as topic-sensitive PageRank [20], Hilltop [22], and stochastic approach for link structure analysis (SALSA) [21], among others.

An ontology is a powerful knowledge representation and reasoning tool that has attracted researchers' attention in recent years [26]. The rich prior knowledge that it describes is considered to be a good candidate for directing and optimizing semantic similarity computing for IR [16,18,19,24]. With the support of standard terminologies or domain ontologies, such as the International Classification of Disease (ICD), Unified Medical Language System (UMLS), and Medical Subject Headings (MeSH) [32], semantic-based IR approaches are widely used for medical information retrieval. For example, the work in Ref. [42] proposes an approach for measuring the semantic similarity between words by using WordNet and other information sources. The vector-space model is suitable for the exploitation of ontological concepts that are recognized from both queries and documents to improve the document ranking [49]. References [25,31] propose approaches to compute the document relevance by determining the semantic relatedness between the words or concepts defined in the corresponding ontologies. By abstracting free-text content into semantic graphs, several papers [27,28,34,40] report work on using graph matching for document ranking. Considering queries as concepts and documents as instances, ontological reasoning is adopted to estimate the document relevance [33]. The work in Ref. [30] is similar to our work, which utilizes semantic relations between concepts for query-page matching, but it relies on the premise that a document has a complete semantic annotation graph that describes its content, which is only satisfied by annotated pages in Semantic Web, and not by documents or EMRs in the medical domain.

Complementary to the existing work, the contribution of this paper is as follows: (1) With the support of an ontology that describes the background knowledge on the medical domain, a novel approach to compute the relevance between queries and documents is proposed, where lightweight semantics (i.e. concepts and the relations between them) identified in queries and documents are exploited for calculating query-dependent scores; (2) Through building the semantic linkages between documents from two different sources, we propose an approach to calculate the document popularity for a medical search. (3) The evaluations derived from the query-document relevance and document popularity are combined for the document ranking, on which basis the experiments on the real-world datasets are reported. The present empirical study clearly demonstrates the effectiveness of the proposed approaches for medical information retrieval.

3. Graph-based semantic rank model

As shown in Fig. 1, the pipeline of the GSRM approach contains three modules, i.e., (1) query-document relevance evaluation; (2) document popularity calculation; and (3) synthetic document ranking. In the following, we will introduce each of those modules in detail.

3.1. Query-document relevance evaluation

Semantic relations among concepts matching the user's intention imply important information and are crucial for medical document ranking. It is intuitive that the more semantic relations between the keywords in a user's query and the concepts a document describes, the higher the probability the document satisfies the user's query. Taking the 'pain in the abdomen' as an example, if document A includes both 'stabbing pain' and 'distending pain' in the abdomen, whereas document B contains generalized 'pain in the abdomen', it is reasonable that we presume that document A is more relevant to the user's query than document B.

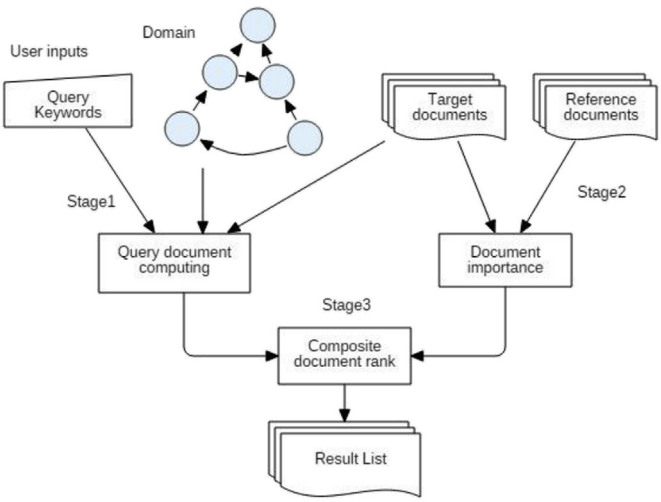

Fig. 1. GSRM architecture.

Download English Version:

https://daneshyari.com/en/article/6920376

Download Persian Version:

https://daneshyari.com/article/6920376

Daneshyari.com