Journal of Biomedical Informatics xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Semantic concept-enriched dependence model for medical information retrieval

Q1 Sungbin Choi [a], Jinwook Choi [a], Sooyoung Yoo [b], Heechun Kim [a], Youngho Lee [c,*]

[a] *Department of Biomedical Engineering, Seoul National University, Seoul, Republic of Korea*
[b] *Center for Medical Informatics, Seoul National University Bundang Hospital, Gyeonggi-do, Republic of Korea*
[c] *Department of Information Technology, Gachon University, Incheon, Republic of Korea*

ARTICLE INFO

Article history:
Received 21 May 2013
Accepted 29 August 2013
Available online xxxx

Keywords:
Search engine
Term dependence model
Ranking
Semantic concept
Unified medical language system

ABSTRACT

Objective: In medical information retrieval research, semantic resources have been mostly used by expanding the original query terms or estimating the concept importance weight. However, implicit term-dependency information contained in semantic concept terms has been overlooked or at least underused in most previous studies. In this study, we incorporate a semantic concept-based term-dependence feature into a formal retrieval model to improve its ranking performance.

Design: Standardized medical concept terms used by medical professionals were assumed to have implicit dependency within the same concept. We hypothesized that, by elaborately revising the ranking algorithms to favor documents that preserve those implicit dependencies, the ranking performance could be improved. The implicit dependence features are harvested from the original query using MetaMap. These semantic concept-based dependence features were incorporated into a semantic concept-enriched dependence model (SCDM). We designed four different variants of the model, with each variant having distinct characteristics in the feature formulation method.

Measurements: We performed leave-one-out cross validations on both a clinical document corpus (TREC Medical records track) and a medical literature corpus (OHSUMED), which are representative test collections in medical information retrieval research.

Results: Our semantic concept-enriched dependence model consistently outperformed other state-of-the-art retrieval methods. Analysis shows that the performance gain has occurred independently of the concept's explicit importance in the query.

Conclusion: By capturing implicit knowledge with regard to the query term relationships and incorporating them into a ranking model, we could build a more robust and effective retrieval model, independent of the concept importance.

Q3 1. Introduction

With the growing availability of medical literature and clinical records in digital form, as well as predominant evidence-based medicine (EBM) philosophy [1], having an effective information search technique has become increasingly important in the medical domain. Semantic knowledge has been actively used for medical information retrieval (IR) research, by virtue of extensive resources such as UMLS® [2]. In previous studies, semantic concept information has been used mostly in two different directions. The first direction is to recognize important concepts that belong to semantic types such as Problem, Intervention, Comparison and Outcome (PICO) and to strengthen their term weights appropriately in a ranking formula [3,4]. The second direction is to expand the original query terms or document terms with lexical variants or synonymous terms to resolve the term mismatch problem [5–7].

In a separate line of work, recent studies in general information retrieval research attempted to combine query term dependencies into a ranking formula to utilize the mutually interdependent characteristics of term occurrences [8,9]. In most of these studies, sequential query terms are blindly assumed to be dependent, which is a fairly reasonable conjecture, but this approach still leaves a substantial amount of opportunity for improvement.

In a specialized domain such as the medical field, people strongly tend to use standardized terms when they communicate, whether in the spoken or written form. For example, 'Disseminated intravascular coagulation' is a medical concept; this concept will be less-often described as 'Coagulations are spreading out through the entire blood vessel' or 'Coagulations are observed, and they are intravascularly disseminated' by medical professionals, although

Q2 * Corresponding author. Address: 534-2 Yeonsu3-dong, Yeonsu-gu, Incheon, Republic of Korea. Fax: +82 32 820 4504.
E-mail address: lyh@gachon.ac.kr (Y. Lee).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.08.013

the alternatives could convey similar meanings. Those preferred terms are combined to represent a single semantic concept and would more often occur in a static pattern rather than appearing separately.

In this study, we assumed that the implicit term dependencies provide information on a search user's behavior, on the way that the users perform their work, and on the way that they choose words and arrange them in sentences. Utilizing domain-specific semantic tools such as MetaMap, we can capture implicit term dependencies that are contained in a query automatically. We hypothesized that, by elaborately revising the ranking algorithms to favor documents that preserve those implicit user requests, we can improve the ranking performance further.

In this paper, our main contributions can be summarized as follows. First, we propose a semantic concept-enriched dependence model, which utilizes semantic knowledge to refine query term dependency elements in a ranking formula. Second, we conduct extensive experiments on both a clinical document corpus (TREC Medical records track) and a medical literature corpus (OHSUMED). Experimental results show that our semantic concept-enriched dependence model is more effective compared to the state-of-the-art baseline. Third, we show that these performance gains are achieved independently of the concept importance, which is distinct from previous approaches.

In the following section, we provide a detailed summary of related studies. In Section 3, we describe other state-of-the-art methods and introduce our semantic concept-enriched dependence model. In Sections 5 and 6, experimental results are presented, and in-depth analysis is performed. In Section 7, we summarize our entire work and introduce future research directions.

2. Background

2.1. Utilizing semantic concepts in medical IR

2.1.1. MetaMap

MetaMap [10] is a computer program that was developed by the U.S. National Library of Medicine (NLM®) for mapping concepts in the UMLS Metathesaurus® from biomedical texts. MetaMap is freely available and is the most popular semantic tool in medical text processing.

Input texts go through lexical and syntactic analysis, variant generation, candidate identification and mapping processes. For a detailed explanation of MetaMap's inner mechanisms, please see [11].

In this study, we did not use our own special methods to recognize medical concepts. Instead, we used default MetaMap settings to improve the ranking performance, which means that our work is generalizable and can be easily reproduced by others.

2.1.2. Concept importance weighting

Many recent studies in IR attempted to identify important concepts and strengthen their term weights to improve the retrieval effectiveness, especially with regard to verbose queries [12–14]. In medical IR, PICO elements are considered to be key information from the perspective of EBM, and it is recommended to formulate clinical questions by the PICO framework [15]. Demner-Fushman and Lin [16] developed a series of knowledge extractors to automatically identify various types of knowledge, including the PICO elements, clinical task type and strength of evidence. To score relevance from the EBM perspective, they designed a direct relevance scoring function that is based on those knowledge components. Boudin et al. [3] identified PICO elements in documents and queries automatically using a machine learning classifier. They incorporated PICO elements into a baseline KL divergence ranking formula by increasing their weights to improve the retrieval performance. Both studies made use of semantic resources to identify the PICO elements.

2.1.3. Lexical expansion

To resolve the term-mismatch problem in IR, a large number of studies used semantic resources, finding relevant terms to allow matching between word pairs that are semantically related to each other. For a comprehensive survey with regard to ontology-based query expansion, please see [17]. In the TREC Genomics track [18], Zhong and Huang [19] attempted to apply a concept-oriented retrieval technique by utilizing the full name of the biomedical entity, its abbreviations and morphological variants. Zhou et al. [6] expanded query terms utilizing several vocabulary sources, including the Medical Subject Headings (MeSH®) [20]. They utilized not only the synonym relationships but also the hypernyms, hyponyms, lexical variants or implicitly related concepts, and they proposed a unified conceptual IR model. Ide et al. [5] used the ESSIE concept-based search engine. Utilizing UMLS, they experimented with five different levels of expansion when they handled synonyms, word variants or missing fragments. In the ImageCLEF medical retrieval task [21] and the TREC medical records track [22], many studies utilized the MeSH in query expansion [23–27]. To name a few, Crespo Azcárate et al. [28] utilized the MeSH tree hierarchy structure when identifying expansion terms. Sondhi et al. [29] used both a medical thesaurus (MeSH) and manual physician feedback in query expansion, comparing different combinations of methodology.

This study is targeted in an orthogonal direction to the two prevailing approaches of concept utilization because it incorporates implicit term dependency that is veiled in semantic concept knowledge.

2.1.4. Incorporating query term dependency in IR

Bag-of-words retrieval models [30,31] represent queries and documents as unordered sets of terms; this strategy is based on an independence assumption. Although bag-of-words models have been shown to be simple and effective, a richer method of representation has been sought to enable further performance gain. Several models had been proposed, such as the n-gram language model [9] or the inference network model [32], but these models are typically more complex yet less effective [33].

Metzler and Croft [8] proposed the Markov random field model as a formal framework to include various term dependencies as ranking features. (A detailed explanation of [8] is presented in Section 3.) Others investigated proximity-based retrieval models [34–39], but they are roughly in the same line of work with Metzler, utilizing term position information.

Recently, many studies have attempted to improve the dependence model. Concept importance was estimated by utilizing collection statistics or external resources and was applied to weight both individual terms and sequential term dependencies [13,14,40]. Park et al. [41] incorporated syntactic relationships that were acquired from syntactic dependency parsing into the retrieval model. Inspired by a quasi-synchronous stochastic process in machine translation, four different types of syntactic dependency were specified to allow inexact matching between them. Lioma et al. [42] analyzed discourse structure in texts and utilized rhetorical relations for retrieval.

Bendersky and Croft [43] introduced the query hypergraph to represent higher-order term dependencies. By using a hypergraph structure, they laid abstract theoretical background to flexibly model dependencies between different query concepts, each one corresponding to the arbitrary number of query terms. However, their work utilized only three types of structures (query term, exact phrase and proximity) as a local factor, which had been com-

Please cite this article in press as: Choi S et al. Semantic concept-enriched dependence model for medical information retrieval. J Biomed Inform (2013), http://dx.doi.org/10.1016/j.jbi.2013.08.013