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

within documents [48]. In [33,46,31], the semantic query expansion methods are adopted to extend query terms to their synonym and meronym sets defined in WordNet and/or Wikipedia. The vector-space model is adapted for the exploitation of ontology based knowledge bases to improve document ranking [20,34]. Approaches in [21,35,43,36], computed document relevance by considering the semantic relatedness between words or concepts defined in corresponding ontologies. By abstracting the free-text content into semantic graphs, references [23,32] reported on work that uses graph matching for document ranking. The work in [38,40,45] utilizes the semantic relations between concepts for query-page matching [53]. Since the semantic-based approaches consider the semantics of terms in queries and documents for their similarity calculation, they provide great potentials to tackle the complicate domain-specific terminologies and then enhance the quality of information retrieval.

Actually, the medical information retrieval can be seen as a domain-specific IR task [4,50,61,62], where the big challenge is to deal with the complexity and ambiguity within medical records and queries. 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), semantic-based IR approaches are widely utilized to tackle the ambiguity problem in medical search. A frequently used mechanism for improving the search quality is query expansion [56] and reformulation [58,62]. In [52,54], the semantic resources are exploited to represent queries in an expressive and meaningful context, through which to fill the semantic gap among the queries and documents and then to improve the medical search quality. Six existing domain-independent semantic similarity measures are adapted to the biomedical domain [52]. By defining the subsumptions as parent–child relationships between concepts, the medical hierarchy is exploited to address the granularity mismatch problem derived from the ambiguous queries [36]. However, all these proposed semantic-based medical IR approaches consider the relevance as the only measure for medical record ranking, which makes their capabilities to handle the query ambiguity limited. The proposed research in this paper can be classified into the category of semantic-based medical search. Our work is different from the existing researches since the proposed diversity-aware medical search approach borrows the novelty as an additional measure to tackle the inherent ambiguity of the user's information needs.

The novelty measure has been extensively used in the research on search result diversification. Current diversification approaches can be categorized as either implicit or explicit, based on how they account for the different query aspects underlying the user input query [10] (An aspect denotes a possible query intent. We use "query aspect" and "query intent" interchangeably in this paper.). Implicit methods [5–7] assume that similar documents cover similar query aspects. Based on this assumption, they promote diversity by directly comparing the documents retrieved for a given query to each other, and demote documents that are content-redundant with documents ranked already. Explicit methods [8–11] model the query aspects explicitly, and promote diversity by maximizing the coverage and minimizing the redundancy of the ranked documents with respect to the query aspects. Actually, the idea of explicitly modeling aspects represents a more promising direction, which is not only supported by theoretical analysis, but also validated by experimental study [8,10]. However, all the existing explicit diversification methods are developed for Web search. Due to inherent complexity of medical terminologies and without large amounts of query logs, they are not applicable for diversifying the search results in the medical IR setting.

In this paper, we use a combination of the technologies on semantic-based IR and Web search result diversification to tackle inherent ambiguity in the medical search process. Based on the given medical domain ontology, the underlying semantics (concepts and relations between them) of the query are discovered, and consumed for modeling various aspects underlying the given query; then based on the modeled query aspects, we propose a novel diversification strategy, which considers not only the aspect importance but also the aspect similarity, to perform medical record ranking.

3. The diversity-aware medical search approach

Our proposed approach on diversify-ware retrieval of medical records includes two steps, i.e., (1) query understanding to discover the implied aspects of the original query as multiple sub-queries; (2) diversity-aware medical retrieval to exploit multiple sub-queries to for diversifying the medical search results. The following of this section gives a detailed description on each of the two steps.

3.1. Query understanding

Since the keyword query is a simple and user-friendly search model, it is prevailing in many practical search systems. Our research assumes to use a keyword-based interface for the users to express their information needs and returns a list of relevant EMRs as the output.

The list of keywords in the query can be interpreted diversely, we need to handle the ambiguity problem, i.e., understand the meanings of the concepts specified in the user's queries and discover the potential aspects of the given query. More specifically, given a query q containing a list of keywords, the task of query understanding is to transforms it into a set of *derived queries* to model different aspects of q. As medical ontology contains rich and accurate professional knowledge that is shared by domain experts, we use it as background knowledge to uncover the underlying aspects of information needs. The detailed query understanding process contains three sub-steps as below.

3.1.1. Query transformation

This sub-step carries out two functions, i.e., keyword phrase identification and expansion. With the support of available semantic resources, e.g., WordNet and Consumer Health Vocabulary (CHV) [13], the former uses the maximum matching approach to scan the keywords in the query sequentially and find the longest matching subsequences defined in the semantic resources as the keyword phrases. For example, given a query "difficulty breathing headache", the longest maximum matching approach can find "difficulty breathing" as a keyword phrase and "headache" as the other keyword phrase.

For the latter, two types of expansions are conducted. On the one hand, the layman keywords input by lay persons should be mapped to professional medical terms, for examples, "difficulty breathing" is rewritten to "dyspnea". As previous researches demonstrated that professional terms were likely to achieve better search results than layman terms [2,13]. We employ the CHV [13], which provides a mapping between medical terms and layman terms, to perform this expansion. On the other hand, the input keywords (even the professional medical terms) may have synonyms. For instances, the distinction between "diagnosis" and "finding" is not clear, and "fever" is a synonym of "febrility". We use both the synonym knowledge in WordNet and the consumer health vocabulary to perform synonym expansion.

Formally, after the processing of this step, each keyword phrase $k_i \in q = \{k_1, k_2, \ldots, k_m\}$ is expanded to a keyword phrase set $S_i = \{k_{i1}, k_{i2}, \ldots, k_{ipi}\}$.

3.1.2. Candidate concept mapping

After m keyword expansion sets S_i ($i = 1, \ldots, m$) is obtained, this sub-step maps the user input keywords to a set of candidate concepts defined in the medical ontology. For each keyword phrase $k_i \in q = \{k_1, k_2, \ldots, k_m\}$, we can find a set of candidate concept nodes $C_i = \{c_{i1}, c_{i2}, \ldots, c_{ini}\}$ with ranking scores $R_i = \{r_{i1}, r_{i2}, \ldots, r_{ini}\}$. Assuming *match* (a, b) denotes the percentage of distinct token q-grams [22] of keyword phrase b is appeared in that of keyword phrase a, r_{ij} is set as $\max_{k \in S_i} \text{match}(k, c_{ij})$, which represents the confidence on mapping keyword k_i to concept c_{ij}. In the implementation, a concept c_{ij} is selected as a candidate one only if $r_{ij} > 0.75$, where 0.75 is chosen for balancing search accuracy and computing complexity for the following derived sub-queries generation.

3.1.3. Derived queries generation

This sub-step constructs a list of derived queries to model the various aspects of q. The basic idea is to extract a list of sub-graphs from the medical ontology, where each sub-graph covers at least one candidate concept of every keyword phrase $k_i \in q = \{k_1, k_2, \ldots, k_m\}$. Then each sub-graph is transformed as a sub-query representing one potential aspect of the interpretation of the user's information need.

Assuming that the user chooses each keyword phrase to represent a concept, then theoretically, there are $y = \prod_m^{i=1} n_i$ concept combinations for query q, where n_i is the number of concepts in C_i, and m denotes the number of keyword phrases in q. For each *combination* cb_j ($j = 1, \ldots, y$), we extract a set of sub-graphs (denoted as $GSet_j$) from the ontology. Each sub-graph in $GSet_j$ represents a possible interpretation of query q. It has a *graph weight* denoting the *importance score*, which can be explained as the possibility of this sub-graph representing the user's information need. After acquiring $GSet_j$ ($j = 1, \ldots, y$) for all y concept combinations, we perform a sub-graph ranking on $GSet_1 \cup \ldots \cup GSet_y$ based on their weights. Then the top ranked sub-graphs with weights larger than a threshold td are selected to represent the potential query interpretations of q. We have conducted experiments to test the different value settings of td, and the results show that our diversifying approach achieves the best performance on all the evaluation criteria when $td = 0.7$. Each selected sub-graph can then be transformed to a keyword-based sub-query, which includes the keyword phrases in the original query q and all the keyword phrases in concept nodes covered by the sub-graph. When the similarity of two sub-queries is bigger than a threshold value, the parents and children of the concepts in the sub-graph can also be used for the query expansion. The sub-graph weight is used to denote the *importance score* of the corresponding sub-query.

The sub-graph construction algorithm is illustrated as below, where the minimum number of edges linking all concepts in a combination together is taken into account for measuring the weight of a graph.

Algorithm 1 (Sub-graphs construction).
Inputs: a concept combination ($C_1, \ldots, C_m$), medical ontology O

Output: $GSet = \{(G_1, w_1), \ldots\}$, with $w \in [0,1]$ being the graph weight

Begin

1. $V = \Phi$, $E = \Phi$;//V is the node set, E is the edge set.
2. $S_C = \{C_1, \ldots, C_m\}$, $S_{initial} = \Phi$;
3. Choose a concept C_i from S_C randomly and set $S_{inital} = \{C_i\}$, $S_C = S_C - \{C_i\}$, $V = \{C_i\}$;
4. while ($S_{initial}$!= $\{C_1, \ldots, C_m\}$)
5. { $length = 1$;
6. do
7. { $S_{length} = \Phi$;
8. For each $C \in S_{initial}$
   $\{S_{length} = S_{length} \cup \{path(C,C') | path(C,C') \in O \&\&$
   $length(path(C,C')) == length \;\&\&\; C' \in S_C\};\}$
9. If ($S_{length} = \Phi$)
   $\{length = length + 1;\}$
10. } While ($S_{length} = \Phi$ && $length <$ Max)
11. For each $path(C, C') \in S_{length}$
12. $\{S_{inital} = S_{inital} \cup \{C'\}$; $S_C = S_C - \{C'\}$;
13. $V = V \cup S'$;//S' is the set of concepts lying on the path $path(C, C')$
14. $E = E \cup E'$;//E' is the edge set comprising of edges on $path(C, C')$
15. }
16. }
17. For pair (C_i,C_j)//$C_i,C_j \in \{C_1,\ldots, C_m\}$ and $i \neq j$.
    $\{PathSet_{i,j} = \{path(C_i,C_j) | path(C_i,C_j) \in G(V,E)$ && concepts in $\{C_1, \ldots, C_m\}$ should not appear on $path(C_i,C_j)\};\}$
18. $S_{graphs} = PathSet_{1,2} \times \ldots \times PathSet_{m-1,m}$;
19. $GSet = \{(G_i, weight_i) | G_i \in S_{graphs}$ && $weight_i = \sum_{i=1}^{m} r_i/m \times |E|\}$
    /*$|E|$ is the number of edges in G_i, r_i is a score measuring the match degree between keyword $k_i \in q$ and its corresponding concept in the input combination, which was computed in the concept mapping sub-step.*/

End

Firstly, Lines 1–16 extract a graph $G(V,E)$, which comprises all the candidate edges linking the candidate concepts of a combination in the tightest way. The graph extraction process is actually a concept search process. S_{inital} stores concepts that belong to the combination and have been found already. S_C stores concepts that have not been found yet. Lines 1–3 start the search from a concept randomly chosen from the combination. Lines 5–10 search unfound concepts using the breadth-first search strategy with the concepts in S_{inital} serving as start points. As Lines 11–15 show, when some concepts in S_C is found, the semantic paths linking it to some concepts in S_{inital} with the shortest length will be found. Then, all the concept nodes and edges on the paths will be added to the graph. After $G(V,E)$ is obtained, Lines 17–19 try to construct the corresponding sub-graph set $GSet$; Line 17 divides $G(V,E)$ into several semantic path sets (a semantic path set comprises all the paths in G that link two concepts together); Line 18 constructs the sub-graph set by combining paths from different semantic path sets, and a sub-graph is generated by choosing one path from each path set and combine them together; Line 19 assigns a graph weight to each sub-graph in $GSet$.

After the processing of this step, we get a list of derived queries $Q = \{(q_1,sc_1), \ldots, (q_t,sc_t)\}$, with sc_j ($j = 1, \ldots, t$) denoting the *importance score* of sub-query q_j. They are normalized across multiple query aspects, i.e., $\sum_j sc_j = 1$.

Given a query "difficulty breathing at night", a schematic example of the query understanding process is illustrated in Table 1, where the WordNet and CHV [13] are employed for the query transformation, and Medical Subject Headings (MeSH) [17] is used as the underlying medical ontology.

3.2. Diversity-aware medical search

The search result diversification can be stated as a tradeoff between relevance and novelty, i.e., given an initial ranking list R for a query q, find a re-ranking S that has the *maximum coverage* and the *minimum redundancy* with respect to different aspects implied by q [10,11].

Table 1
Diversification performance using different diversification strategies.

Original query	Query transformation	Candidate concept mapping	Sub-query generation
difficulty breathing at night	Two keyword phrases are identified: #1: *difficulty breathing*; #2: *at night*; Keyword: phrase expansion: #1 (*difficulty breathing*) sameAs: {*dyspnea*; *dyspnea; shortness of breath*;} #2 (*at night*) sameAs: {*night time*; *nocturnal*;}	Five concepts are identified for #1: c_{11}: *Dyspnea* [*ID*: C08.618.326] c_{12}: *Dyspnea* [*ID*: C23.888.852.371] c_{13}: *Dyspnea, Paroxysmal* [*ID*: C23.888.852.371.396] c_{14}: *Dyspnea, Paroxysmal* [*ID*: C08.618.326.396] c_{15}: *Dyspnea, Paroxysmal* [*ID*: C14.280.434.313] 12 concepts are identified for #2: c_{21}: *Somnambulism* [*ID*: C10.886.659.635.700] c_{22}: *Somnambulism* [*ID*: F03.870.664.635.700] c_{23}: *Nocturnal Enuresis* [*ID*: C12.777.934.284.500] C_{24}: *Nocturnal Enuresis* [*ID*: C13.351.968.934.252.500] C_{26}: *Sleep Bruxism* [*ID*: C07.793.099.500] c_{27}: *Sleep Bruxism* [*ID*: C10.886.659.637] ... Note that the entry term of a concept defined in the MeSH is considered as the synonym of the corresponding concept.	Three sub-queries are generated: q1: {(*respiration disorders*; *dyspnea*; *dyspnea, paroxysmal nocturnal*; *difficulty breathing*; *at night*;), 0.35} q2: {(*heart failure*; *dyspnea, paroxysmal nocturnal*; *difficulty breathing*; *at night*;), 0.3} q3: {(*signs and symptoms, respiratory*; *dyspnea, paroxysmal nocturnal*; *difficulty breathing*; *at night*;), 0.35}

This problem is an instance of the maximum coverage problem [12], which is NP-hard. Fortunately, there is a well-known greedy approximation to this problem [10,11]: Given a ranking R for an ambiguous query q, a re-ranking S can be produced by iteratively selecting a 'local-best' document from R/S (the 'local-best' document should provide the maximum coverage of the aspects underlying the initial query, and the minimum redundancy with respect to the aspects covered by the documents already in S). Based on this basic approximation idea, we propose a novel strategy for search result diversification.

Given the initial query q and its derived query list $Q = \{(q_1,sc_1), \ldots, (q_t,sc_t)\}$, we produce a document list for q and each of the derived queries, respectively. The ranking model adopted can be vector-space models, probabilistic rank models or learning-based rank models, which support relevance computing. For the sake of clarity, a document list for the initial query is denoted as *base-ranking*, while those generated for derived queries are denoted as *derived-rankings*.

Our proposed strategy performs search result diversification by producing a re-ranking S of the *base-ranking*, based on the derived query list Q and their document lists (*derived-rankings*). In practice, the re-ranking process is as follows:

3.2.1. Determining the first ranked document

The top-1 document in *base-ranking* is chosen to be put at the first position of the re-ranking list S, as it is the most relevant one to q.

3.2.2. Ranking the rest documents in base-ranking

After the selection of the first ranked document, we iteratively select the "local-best" document to be put in the next position. At each iteration, the document $d \in base\text{-}ranking \backslash S$ that scores the highest according to Eq. (1) is chosen as the "local-best" one:

$$LB(d,S) = \delta \cdot relevance(d,q) + (1-\delta) \cdot novelty(d,S) \quad (1)$$

where $relevance(d,q)$ models the relevance of d, whose value is actually the relevance score of d in the *base-ranking*; $novelty(d,S)$ models the *novelty* of d, given how well the user input query q is already satisfied by documents in S; δ ($\delta \in [0,1]$) controls the tradeoff between the two factors.

In order to derive $novelty(d,S)$, we compare a document d to each of the selected documents in S, and integrate these comparison results together according to Eq. (2):

$$novelty(d,S) = \prod_{d_i \in S} novel(d,d_i) \quad (2)$$

where $novel(d,d_i)$ measures the novelty of d, given how well the user query q is already satisfied by the document d_i. We explicitly use the various query aspects underlying q, in the form of derived queries $Q = \{(q_1,sc_1),\ldots,(q_t,sc_t)\}$, to compute $novel(d,d_i)$. The basic computing idea is as follows: (1) the more derived queries d covers, and the less derived queries d and d_i share, the more novel d is with respect to d_i; (2) the relative *importance* of a derived query, with respect to other derived queries associated with q, is a necessary factor for novelty computing. A query with a higher *importance score* should play a more significant role in identifying novelty than a less important query does. (3) The similarity between derived queries is another factor that should not be ignored. Similar queries should avoid redundant contribution in novelty measuring, in order to produce unbiased result. Eq. (3) shows the detailed computing method:

$$novel(d,d_i) = \sum_{q_j \in Q} rs(q_j) \cdot \prod_{m \in [1,j-1]} (1 - sim(q_j,q_m)) \cdot rel(d,q_j) \cdot (1 - rel(d,q_j) rel(d_i,q_j)) \quad (3)$$

where $rs(q_j)$ denotes the *relative importance* of q_j with respect to other queries in Q, which is derived by normalizing its importance score sc_j; $rs(q_j) = sc_j / \sum_{k=1}^{t} sc_k((q_j,sc_j) \in Q)$; sim (q_j,q_m) evaluates the similarity between two queries, which is computed by employing a graph matching algorithm [13] on the sub-graphs corresponding to q_j and q_m; rel (d,q) denotes the document relevance score with respect to a query, which is provided by *derived-rankings*; rel $(d,q_j)rel(d_i,q_j)$ measures how well d and d_i share the topic q_j.

4. A real-world pilot study

In this section, a pilot study is described for the implementation and deployment of the proposed diversity-aware medical search solution on a real-world Cloud of regional healthcare collaboration platform. The empirical experiments and the evaluation results are reported. As the reference for the development of similar application, we summaries the findings and learned lessons in the real-world trial for the application of the diversity-aware medical search.

4.1. The usage scenario and system implementation

As shown in Fig. 1, a real-world application of the diversity-aware medical search involves three types of organization, i.e., (A) community hospitals or senior people nursing centers with general medical practitioners, who have particular skills and practice a holistic approach for treating people with multiple health issues; (B) regional medical center or class A hospitals with domain specific medical doctors, who are expert of one or several diseases; (C) Cloud service providers with the healthcare collaboration platform enabling the provision of collaborative medical care to

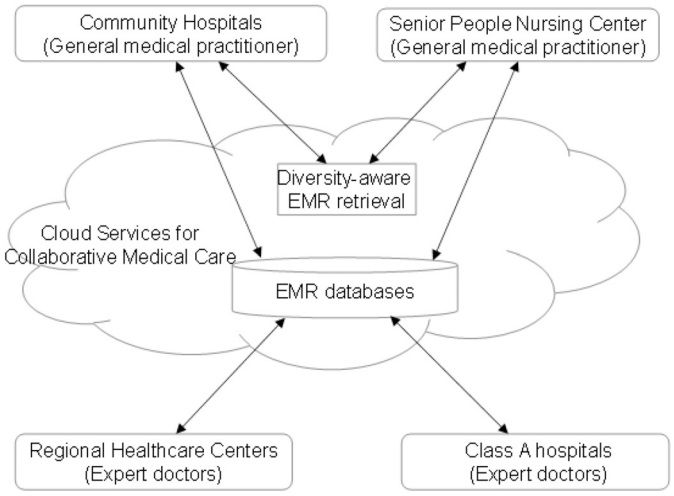


Fig. 1. The usage scenario of the real-world application of the diversity-aware medical search.

the outpatients or inpatients of organizations (A) and (B). The component of the diversity-aware medical search is deployed on the Cloud of healthcare collaboration platform, which supplies the medical data storage and management services under certain privacy protection mechanisms [37,59,60,64,65]. The goal of the real-world application is to use the medical records constructed by the expert doctors in organization (B) to train the general medical practitioners in organization (A) or provide references when they provides preventive care and health education to their patients.

The implementation and deployment of the diversity-aware medical search includes two phases, i.e., the build-time phase and run-time phase. In the former, the component for diversity-aware medical search is deployed in the Cloud to be accessed by the general medical practitioners in organization (A). More specifically, three tasks are involved:

(1) Full-text indexing of the EMR: The widely used Lucene library [67] is adopted here for the indexing implementation. According to the Health Level 7(HL7) CDA, the EMR conceptual model has three parts: (1) patient basic data; (2) patient history profile; (3) clinical data, including symptoms, diagnosis, and treatments of each hospital visit by a patient. The full-text indexing is conducted on all these three parts with a lot of fields. Since negative qualifiers, e.g., no or without, are commonly used in the patient profile and clinical data to describe the absence of a medical condition, the dedicated negation handling approach [69] is incorporated into the EMR indexing process.
(2) Semantic indexing of the medical ontology: To speed up the query transformation (see Section 3.1.1), concept mapping (see Section 3.1.2), and derived query generation (see Section 3.1.3), and then the query understanding, we index all vocabularies and concepts in the medical ontology.
(3) Software interface development: This component is used for connecting the health information systems of the organizations (A) and (B) to the storage service in the healthcare collaboration platform, through which the medical records generated from organization (B) are stored automatically in the medical Cloud platform. In this pilot study, the collected EMRs will be accessed by the general medical practitioners in organization (A).

In the run-time phase, whenever a medical record is constructed by the expert doctor in organization (B), one copy is stored remotely in the medical Cloud platform. At the same time, the full-text indexer is triggered to conduct the indexing task. The application supports the medical practitioners in organization (A) to conduct keyword-based query over the available EMR datasets. The implementation mainly covers two tasks:

(1) Query understanding: Both the Lucene [67] and Jena [68] libraries are utilized together to realize the query understanding. To achieve high efficient query transformation and concept mapping, their underlying matching process is implemented by using keyword-based search over the indexer of the involved vocabularies and medical ontology. Note that the keyword phrase matching is conducted in real-time when the user is typing the query in the search box, which can also speed up the matching process. For the sub-query generation, the Jena [68] is used for the semantic relation inference and subgraph identification.
(2) Diversity-aware EMR retrieval: The Lucene library [67] is employed for the implementation of the retrieval process. The original query and derived multiple sub-queries are sent to the indexer of EMR data. Their corresponding relevant EMR lists are retrieved, respectively. Then, the re-ranking-based diversification strategy are implemented to generate the EMR search results and present them to the medical practitioner for reference or online learning.

4.2. Experimental evaluation

This section reports the empirical experiments on the available EMR datasets and evaluation results.

4.2.1. Dataset

We use a real-world dataset for the empirical evaluation of the proposed approach for diversity-aware medical search, which comprises three parts described as below:

- *Document set*: over 100k medical records acquired from our cooperation hospitals are selected as target documents to perform ranking on. A medical record is a free-text description that keeps track of a patient's symptoms and how he (she) is diagnosed and medically treated.
- *Sample queries*: a total of 30 sample queries are generated, where the above mentioned "difficulty breathing at night" is a concrete example. For sample query generation, we first selected 100 most frequent symptom descriptions from the medical records as candidate queries (each description is dictated by patients stating their situations); then ambiguity of these candidate queries was estimated using an entropy-based measure [15]. Then, 30 queries with the highest ambiguity were chosen as target ones.
- *Semantic knowledge base*: A medical diagnosis ontology, MDO, was built based on the disease knowledge extracted from the MeSH ontology [17], as well as the diagnosis knowledge extracted from Collins's medical textbook [18] and our medical records. MDO contains 1526 classes, 9 types of relations linking symptoms and diseases, diseases and diseases, checking items and symptoms and so on, 5971 class instances and 33,627 relation instances.

4.2.2. Evaluation criteria

- *α-nDCG@10 (@20)*: α-nDCG [16] is an official metric adopted by the diversity task of TREC Web Track. It rewards both relevance and diversity of the results, and balances the two factors through a turning parameter α. α-nDCG is computed with $\alpha = 0.5$ in our implementation, in order to give equal weights to both relevance and diversity. α-nDCG@10 (@20) evaluates the performance of the top-10 (20) documents.

- *precision-IA@10(@20)*: The precision-IA metric extends the traditional notion of precision, which accounts for the possible aspects underlying a query and their relative importance [15]. Eq. (4) shows it in detail, where M denotes the number of sample queries, N_t denotes the number of derived queries for a sample query t ($1 \le t \le M$), and $j_t(i,j)$ is a binary score (0 or 1) denoting the relevance of the document returned for query t at depth j to the derived query i of sample query t.

$$precision - IA@k = \frac{1}{M}\sum_{t=1}^{M}\frac{1}{N_t}\sum_{i=1}^{N_t}\frac{1}{k}\sum_{j=1}^{k} j_t(i,j) \quad (4)$$

4.2.3. Experiment results

Previous researches have demonstrated that explicit diversification approaches outperforms the implicit ones. We compare our approach against two state-of-the-art work for explicit search result diversification:

- *IA-Select* [8]: IA-Select is a greedy algorithm focusing on maximizing the probability that the average user finds at least one useful result within the top-k results. Based on a predefined taxonomy to which both queries and documents are classified, it performs diversification by iteratively promoting documents that share a high number of classes with the query, while demoting those with classes already represented in the ranking.
- *xQuAD* [10]: xQuAD is a probability-based framework which explicitly uses the aspects underlying a query for search result diversifying. After setting the related queries (suggested queries) produced by some major web search engines as sub-queries of a user query, xQuAD performs search diversification based on the estimated relevance of documents to those sub-queries, as well as on the relative importance of each sub-query in light of the user query.
- *The Semantic-based Search Result Diversification (SSRD)*: It is the implementation of the proposed approach in this paper. The only parameter in SSRD that should be set in advance is δ, which balances the relevance score and novelty score for "local-best" document computing. In this research, we performed the α-nDCG@20 evaluation on the test dataset to select the optimal parameter of $\delta = \{0, 0.1, 0.2, \ldots, 1\}$. The results are shown in Fig. 2, where the vertical axis denotes the α-nDCG@20 measure, and the horizontal axis represents different δ values. Accordingly, δ was set to 0.8, which achieves a peak value of α-nDCG@20 on the real-world dataset. In addition, for the ranking model used to produce *base-ranking* and *derived-rankings*, we chose the DPH Divergence Form Randomness model [18], which was validated by [8] as the best among three widely used ones, BM25, DPH and LM.

Effectiveness evaluation of the query aspect discovery strategy

Several recent research findings [8–10] indicated that a high quality query aspect discovery result was crucial for a well-functioning search diversifying model. To be fair, we use the two existing search result diversification approaches, i.e., IA-select and xQuAD, to evaluate the contribution of the semantic based query aspect generation (SAG) approach as described in Section 3.1.

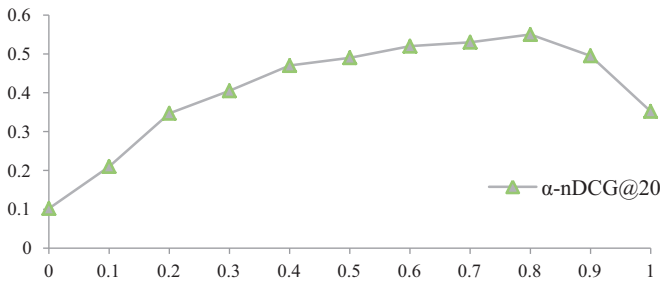


Fig. 2. The experiments about the tuning of parameter δ to balance relevance and novelty.

For performance comparison, the query log based technique (LAG) [10] is employed as the baseline. It relies on the query reformulation mechanism of Google for query aspect production, which is regarded as state-of-the-art query log mining solution for query aspect generation [10].

We hired 7 students in a medical school to help us judge the query-document relevance, which serves as the foundation for computing α-nDCG and precision-IA. For each of the top-10 (20) documents, each student makes a binary decision regarding whether the document is relevant to a derived query. And then if more than half of the 7 students give a positive vote to a document, it is evaluated as a matched one of the derived query.

The implementation LAG (SAG) + IA-select (xQuAD) uses LAG (SAG) for query aspect generation and IA-select (xQuAD) for search result diversification. Their performance comparison in terms of α-nDCG and precision-IA is illustrated in Table 2. By comparing the results of employing IA-select on LAG with that on SAG, we observed that SAG + IA-select outperforms LAG + IA-select on all settings, with a performance increase of up to 4.5% on α-nDCG and 0.6% on precision-IA. The same findings were also found when comparing SAG + xQuAD with LAG + xQuAD, and SAG improves LAG with a gain of up to 5.9% on α-nDCG and a gain of up to 0.8% on precision-IA. For the comparison of IA-select and xQuAD, for both LAG and SAG, xQuAD outperforms IA-select. This result is consistent with the experiment evaluation in [10].

The observations in Table 2 can provide indirect proof that, the quality of query aspects derived from the SAG (i.e., the query understanding described in Section 3.1), is better than that from LAG (mining query aspects from the Web query log). It demonstrates that the medical domain ontology can provide significant positive effect on the medical search quality. Furthermore, the comparison results in Table 2 also show that SAG approach is more appropriate than LAG to handle the complexity of domain terminologies.

Diversifying strategy evaluation

In this experiment, we test the effectiveness of the diversifying strategy of SSRD. To provide a fair comparison, both SSRD and baselines utilize the same query aspect generation method SAG. Two versions of SSRD were implemented for testing:

- $SSRD_1$: This SSRD variant disregard the derived queries similarity when performing diversification, which is realized by removing the $\prod_{m\in[1,j-1]}(1 - sim(q_j, q_m))$ item from Eq. (3).
- $SSRD_2$: This SSRD version employs our complete diversification strategy as introduced in Section 3.2 for diversity aware EMR ranking.

The experimental results are illustrated in Table 3, where the best results are highlighted in bold. The performance comparison between two baselines validates again that xQuAD outperforms IA-select in terms of both α-nDCG and precision-IA. Comparing the proposed SSRD to two baseline approaches we can find that, the

Table 2
Diversification performance using different query aspect generation methods.

	α-nDCG		Precision-IA	
	@10	@20	@10	@20
LAG + IA-select	0.339	0.36	0.165	0.153
SAG + IA-select	0.371	0.405	0.169	0.155
LAG + xQuAD	0.403	0.474	0.187	0.18
SAG + xQuAD	0.462	0.517	0.195	0.187

Table 3
Diversification performance using different diversification strategies.

	α-nDCG		Precision-IA	
	@10	@20	@10	@20
SAG + IA-select	0.371	0.405	0.169	0.155
SAG + xQuAD	0.462	0.517	0.195	0.187
SAG + $SSRD_1$	0.495	0.539	**0.221**	**0.206**
SAG + $SSRD_2$	**0.503**	**0.55**	0.217	0.204

two variants of SSRD outperform the baselines significantly on all settings, which demonstrates that the effectiveness of combining the proposed query aspect generation and diversification strategy for improving medical search quality. Among the four approaches, $SSRD_2$ ($SSRD_1$) has the best performance in terms of α-nDCG (precision-IA). This phenomenon shows that, the query aspect similarity is a valuable factor for improving diversification performance in terms of α-nDCG, and there is no clear indication that the query aspect similarity has positive effect on the EMR retrieval in terms of precision-IA.

The computing complexity is a crucial parameter to determine the feasibility of a document ranking algorithm. Since the concept matching is conducted in real-time when the user typing the query in the search box, two main time consuming tasks of SSRD are the derived query generation (described in Section 3.1.3) and diversity-aware medical search (described in Section 3.2). For the sub-graphs construction algorithm (algorithm1) for derived query generation, its time complexity is $O(\sum_{i=1}^{n} d_i e_i)$, where n denotes the number of concepts in the concept combination; d_i is the degree of concept c_i in the ontology graph; and e_i represents the maximum value of the path lengths that link concept c_i to other concepts belong to the combination. As there are y concept combinations derived from Section 3.1.2, the total time complexity of the derived query generation part is $O(\sum_{i=1}^{n} d_i e_i)$ which is linear to number of edges in the ontology graph. For the diversity-aware medical search, besides the time spent on the document relevance ranking cost by the existing relevance computing model (DPH model), $O(n_d^2 n_q)$ were consumed to combine the relevance-based ranking results and produce a final diversification-aware ranking list. n_d denotes the number of documents in base-ranking that is extracted by DPH from the document set; n_q represents the number of derived queries. Our experiment shows that, the execution time of the proposed SSRD approach is very similar to that of two baseline approaches.

4.3. Discussion

From the feedbacks of the general medical practitioner, they are basically satisfied with the search results of our pilot system, which demonstrates the effectiveness of our proposed approach to solve the problem of uncertainty in the information needs for ambiguous medical queries. The advantages of employing the diversity-aware approach for medical search can be summarized into three aspects:

(a) Enhancing the secondary use of medical records: The current dominant medical search approaches [56,60] use query expansion to handle the ambiguity problem. Since the relevance is the only employed measure for EMR ranking, their ability to retrieve accurate and complete EMRs about desired patient populations is limited [60]. Our proposed diversity-aware medical search exploits the medical domain ontology to discover the potential aspects of the original query as multiple sub-queries, through which to enable a way to answer ambiguous medical queries, i.e., the novelty is introduced as an additional measure for EMR ranking, which can increase the satisfaction of the average user with the medical search results. Note that to retain the privacy of citizens as much as possible, a generally accepted concept in healthcare domain is that only the "need to know" data should be presented to medical personnel. Thus a big challenge for the enhanced second use of EMRs is to balance the privacy protection and medical data utilization, which will an important task for our future research work.
(b) Improving the quality of medical care: The diversity-aware medical search approach enables the general medical practitioners to find their interested EMRs constructed by the expert doctors in a high effective and efficient way. Through the case-based learning, the general practitioners use the knowledge conveyed by the EMRs to guide and improve the quality of their provision of preventive care and health education to patients, which make them behave like expert doctors.
(c) Balancing the utilization of the medical resources: The healthcare resource utilization is extremely unbalanced in China. Due to the patients' preference to expert doctors, the class A hospitals are overcrowded with patients, thus resulting in the tension of relationship between doctors and patients. However, there is almost no patient in the community hospitals. The enhancement of general practitioners' medical care service has great potential to change the choice of the patients with chronic illnesses or the requirement of preventive care from class A hospitals to community hospitals, which make the expert doctor have more time to provide medical care to the patients with the intractable diseases.

We also report some of the learned lessons or important points of detailed design that can serve as references for the development of similar systems for enhanced healthcare services:

(A) Working closely with the user: The motivation barrier is an important issue for the adoption of information technologies to improve the healthcare service. To provide concrete incentives to the end users, we need to work closely with them and find the matching points of their real requirements and the proposed technologies.
(B) Medical domain knowledge playing the key role: The wide availability of medical domain ontology is the key to motivate our research and make our proposed approach be successfully applied in a real-world application. Our pilot study underpins the conclusion that lightweight semantic is more feasible to be produced and consumed in a real semantic application [66]. The vocabularies describing the diseases and symptoms in the domain ontology is widely used in the EMRs, which makes our solution have minimum disturbance to the existing infrastructure. According to the initial evaluation results of the query understanding, we find that many concepts were not defined in the ontology. In the future, we will adopt learning based approach [41,63] to recognize the unknown concepts from the text and link them with the nodes in the knowledge base to handling the incompleteness of the medical ontology.
(C) Redundant indexing to reduce the response time: To reduce the response time of the query processing, multiple copies of the free-text index are stored in the Cloud server. When given a query q, the constructed multiple sub-queries can be processed in a parallel way. Actually, this implementation is a tradeoff between the computing time and storage space.
(D) Addressing the privacy concerns: The current trial system has only been internal used by the general medical practitioners in organization (A). In the future, we will release it as an open website to be accessed by the registered patients for their self-education. For the implementation of such a public Internet

system, we plans to adopt the models of k-anonymity [59,60] or differential privacy [64] to privacy preserved medical data retrieval. In addition, the previous work [53] has illustrated that the synergy among multiple information sources can provide enhanced information utilization more than the sum of using them separately ($1 + 1 \geq 2$). The utilization of semantic linkages among multiple sources for enhanced medical data consumption and then for improved medical care needs to be conducted under certain privacy protection mechanisms.

5. Conclusion

This paper proposed a novel approach for diversity-aware IR in medical environment. With the support of domain ontology, the semantic information implied in a user query was extracted and consumed for query aspect modeling; then based on the modeled query aspects, a novel diversification strategy, which considers not only the aspect importance but also the aspect similarity, was proposed to perform document ranking. The empirical experiments on the real-world datasets illustrate the effectiveness of the proposed approaches. We report a pilot study on the application of the proposed medical search approach for the secondary use of medical records, through which to enhance the healthcare quality and balance the utilization of medical resources.

This paper gives the basic concept of the diversity-aware retrieval of medical records, which represents a beginning to combine the technologies of semantic IR and search result diversification for the secondary use of EMRs. This paper only considers the utilization of the semantics in the query for enhanced medical search. In future, we will extend current research to exploit the semantics in the free-text EMRs and the data fusion of multiple sources for improvement of the secondary use of available medical data and then enhancing the healthcare quality.

Acknowledgements

Dr. Rui Mao is the corresponding author. The main work mentioned in this paper was conducted when the first author worked at NEC Labs China. This research was supported by the following grants: China 863: 2012AA010239; NSF-China: 61170076.

References

[1] A.E.K. Sobel, The move toward electronic health records, Computer 45 (11) (2012) 22–23.
[2] G. Luo, Intelligent output interface for intelligent medical search engine, in: Proc. of AAAI, 2008, 1201–1206.
[3] D. Fridsma, Electronic health records: the hhs perspective, Computer 45 (November (11)) (2012) 24–26, http://dx.doi.org/10.1109/MC.2012.371.
[4] A. Betin, MedicoPort: a medical search engine for all, Computer Methods and Programs in Biomedicine 86 (April) (2007) 73–86.
[5] J. Carbonell, J. Goldstein, The use of MMR, diversity-based reranking for reordering documents and producing summaries, in: Proc. of SIGIR'98, 1998, 335–336.
[6] C. Zhai, W.W. Cohen, J. Lafferty, Beyond independent relevance: methods and evaluation metrics for subtopic retrieval, in: Proc. of SIGIR'03, 2003, 10–17.
[7] H. Chen, D.R. Karger, Less is more: probabilistic models for retrieving fewer relevant documents, in: Proc. of SIGIR'06, 2006, 429–436.
[8] R. Agrawal, S. Gollapudi, Diversifying search results, in: Proc. of WSDM, 2009, 5–14.
[9] F. Radlinski, S. Dumais, Improving personalized web search using result diversification, in: Proc. of SIGIR'06, 2006, 691–692.
[10] R.L.T. Santos, C. Macdonald, L. Ounis, Exploiting query reformulation for web search result diversification, in: Proc. of WWW'10, 2010, 881–890.
[11] R.L.T. Santos, C. Macdonald, L. Ounis, Intent-aware search result diversification, in: Proc. of SIGIR'11, 2011, 595–604.
[12] D.S. Hochbaum, Approximation Algorithms for NP-hard Problems, PWS Publishing Co, Boston, MA, 1997.
[13] Q.T. Zeng, T. Tse, Exploring and developing consumer health vocabularies, Journal of the American Medical Informatics Association 13 (2006) 24–29.
[14] L.A. Zager, G.C. Verghese, Graph similarity scoring and marching, Applied Mathematics Letters 21 (2008) 86–94.
[15] E. Demidova, et al., DivQ: diversification for keyword search over structured databases, in: Proc. of SIGIR'10, 2010, 331–338.
[16] C.L.A. Clarke, N. Craswell, I. Soboroff, Overview of the TREC 2009 web track, in: Proc. of TREC, 2009.
[17] MeSH Homepage, 2006 http://www.nlm.nih.gov/mesh/meshhome.html.
[18] R.D. Collins, Algorithmic Diagnosis of Symptoms and Signs: Cost-effective Approach, Lippincott Williams & Wilkins, Philadelphia, PA, 2002.
[20] P. Cstells, M. Fernandez, D. Vallet, An adaptation of the vector-space model for ontology-based IR, IEEE Transactions On Knowledge and Data Engineering 19 (2) (2007) 261–272.
[21] A.M. Rinaldi, An ontology-driven approach for semantic IR on the Web, ACM Transactions on Internet Technology 9 (3) (2009).
[22] J.-Q. Li, Y. Zhao, B. Liu, Exploiting external semantic resources for large scale text categorization, Journal of Intelligent Information Systems 39 (3) (2012) 763–788.
[23] F. Brauer, W. Barczynsi, RankIE: document retrieval on ranked entity graphs, Proceeding of VLDB Endowment 2 (2) (2009) 1578–1581.
[24] E. Demidova, P. Fankhauser, X. Zhou, W. Nejdl, DivQ: diversification for keyword search over structured databases, in: Proceedings of SIGIR'09, 2009.
[25] G. Salton, C. Buckley, Introduction to Modern IR, McGraw-Hill, New York, 1983.
[26] J. Li, Y. Zhao:, PathRank: web page retrieval with navigation path, in: Proc. ECIR09, 2009, 350–361.
[27] S. Deerwester, S. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the Society for Information Science 41 (1990) 391–407.
[28] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791.
[29] J.-Q. Li, Y. Zhao, H. Garcia-Molina, A path-based approach for web page retrieval, World Wide Web: Internet and Web Information System 15 (3) (2012) 257–283.
[30] S. Page, R.M. Brin, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford University, 1998.
[31] J. Li, Y. Zhao, B. Liu, Fully Automatic Text Categorization by Exploiting WordNet AIRS, 2009, 1–12.
[32] M. Fernández, I. Cantador, V. López, D. Vallet, P. Castells, E. Motta, Semantically enhanced information retrieval: an ontology-based approach, web semantics: science, Services and Agents on the World Wide Web 9 (4) (2011) 434–452.
[33] D.I. Moldovan, R. Mihalcea, Using wordnet and lexical operators to improve internet searches, IEEE Internet Computing 4 (2000) 34–43.
[34] L. Khan, D. McLeod, E.H. Hovy, Retrieval effectiveness of an ontology-based model for information selection, The VLDB Journal – The International Journal on Very Large Data Bases 13 (1) (2004) 71–85.
[35] O. Egozi, S. Markovitch, E. Gabrilovich, Concept-based information retrieval using explicit semantic analysis, ACM Transactions on Information Systems 29 (2) (2011) 1–34.
[36] G. Zuccon, B. Koopman, A. Nguyen, D. Vickers, L. Butt, Exploiting medical hierarchies for concept-based information retrieval, in: Proceedings of the Seventeenth Australasian Document Computing Symposium, 2012, pp. 111–114.
[37] B. Liu, J. Li, C. Liu, Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses, Journal of Biomedical Informatics (2014), http://dx.doi.org/10.1016/j.jbi.2014.01.005.
[38] F. Lamberti, A relation-based page rank algorithm for semantic web search engines, IEEE Transaction on Knowledge and Data Engineering 21 (2009) 123–136.
[39] Y. Li, Z. Bandar, D. McLean, An approach for measuring semantic similarity between words using multiple information sources, IEEE Transactions on Knowledge and Data Engineering 15 (4) (2003) 871–882.
[40] C. Liu, J. Li, Semantic-based composite document ranking, in: IEEE International Conference on Semantic Computing, 2012, 126–129.
[41] J. Li, C. Liu, Large scale sequential learning from partially labeled data, in: IEEE International Conference on Semantic Computing, 2013, 176–183.
[43] B.-C. Chien, C.-H. Hu, M.-Y. Ju, Ontology-based information retrieval using fuzzy concept documentation, Cybernetics and Systems 41 (1) (2010) 4–16.
[44] L. Kallipolitis, V. Karpis, I. Karali, Semantic search in the World News domain using automatically extracted metadata files, Knowledge-Based Systems 27 (2012) 38–50.
[45] S. Kara, O. Alan, O. Sabuncu, S. Akpınar, N.K. Cicekli, F.N. Alpaslan, An ontology-based retrieval system using semantic indexing, Information Systems 37 (4) (2012) 294–305.
[46] F. Zhao, F. Fang, F. Yan, H. Jin, Q. Zhang, Expanding approach to information retrieval using semantic similarity analysis based on Wordnet and Wikipedia, International Journal of Software Engineering and Knowledge Engineering 22 (2) (2012) 305–322.
[47] V. Jindal, S. Bawa, S. Batra, A review of ranking approaches for semantic search on Web, Information Processing & Management (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.004.
[48] B. Liu, J. Li, Y. Zhao, Repairing and reasoning with inconsistent and uncertain ontologies, Advances in Engineering Software 45 (1) (2012) 380–390.
[49] G. Leroy, J.E. Endicott, Combining nlp with evidence-based methods to and text metrics related to perceived and actual text di_culty, in: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, ACM, 2012, pp. 749–754.
[50] G. Luo, C. Tang, On iterative intelligent medical search, in: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2008, pp. 3–10.
[51] W.R. Hersh, Information Retrieval: A Health and Biomedical Perspective, 3rd ed., Springer, 2009.
[52] X. Yan, R.Y. Lau, D. Song, X. Li, J. Ma, Toward a semantic granularity model for domain-specifc information retrieval, ACM Transactions on Information Systems 29 (July) (2011), 15:1–15:46.

[53] J. Li, J.-J. Yang, C. Liu, Y. Zhao, B. Liu, Y. Shi, Exploiting semantic linkages among multiple sources for semantic information retrieval, Enterprise Information Systems (2014), http://dx.doi.org/10.1080/17517575.2013.879923.
[54] A. Babashzadeh, J. Huang, M. Daoud, Exploiting semantics for improving clinical information retrieval, in: SIGIR, 2013, 801–804.
[55] E. Voorhees, R. Tong, Overview of the TREC 2011 medical records track, in: Proc. of TREC, 2011.
[56] M. Ellen, E. Voorhees, The TREC Medical Records Track, in: Proc. of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, BCB, 2013.
[57] X. Yin, J.X. Huang, X. Zhou, Z. Li, A survival modeling approach to biomedical search result diversi_cation using wikipedia, in: SIGIR'10, 2010.
[58] H.K. Jain, C. Thao, H. Zhao, Enhancing electronic medical record retrieval through semantic query expansion, Information Systems and e-Business Management 10 (2) (2012) 165–181.
[59] J. Li, J.-J. Yang, Y. Zhao, B. Liu, A top-down approach for approximate data anonymisation, Enterprise Information Systems 7 (3) (2013) 272–302.
[60] C.M. Benjamin Fung, K. Wang, R. Chen, S. Yu, Privacy-preserving data publishing: a survey on recent developments, ACM Computing Surveys 42 (4) (2010) 1–53, Article No. 14.
[61] A. Hanbury, Medical information retrieval: an instance of domain-specific search, in: SIGIR, 2012, 1191–1192.
[62] T. Edinger, A. Cohen, et al., Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC Medical Records Track, in: AMIA 2012 Annual Symposium, Chicago, IL, 2012.
[63] J. Li, C. Liu, A co-learning approach for concept detection in documents, in: IEEE International Conference on Semantic Computing, 2012, 310–317.
[64] C. Dwork, Differential privacy: a survey of results, in: Proceedings of The 5th Annual Conference on Theory and Applications of Models of Computation, 2008, pp. 1–19.
[65] J.J. Yang, J. Li, Y. Niu, A hybrid solution for privacy preserving medical data sharing in the cloud environment, Future Generation Computer Systems (2014), http://dx.doi.org/10.1016/j.future.2014.04.004.
[66] C. Bizer, T. Heath, T. Berners-Lee, Linked data – the story so far, International Journal on Semantic Web and Information Systems 5 (3) (2009) 1–22.
[67] http://lucene.apache.org/.
[68] http://jena.apache.org/.
[69] N. Limsopatham, C. Macdonald, R. McCreadie, I. Ounis, Exploiting Term Dependence while Handling Negation in Medical Search, in: Proc. of SIGIR, 2012.

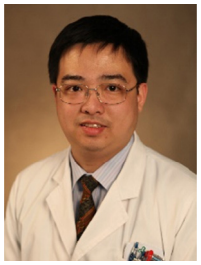
Jianqiang Li received his B.S. degree in Mechatronics from Beijing Institute of Technology, Beijing, China in 1996, M.S. and Ph.D degrees in Control Science and Engineering from Tsinghua University, Beijing, China in 2001 and 2004, respectively. He worked as a researcher in Digital Enterprise Research Institute, National University of Ireland, Galway in 2004–2005. From 2005 to 2013, he worked in NEC Labs China as a researcher, and Department of Computer Science, Stanford University, as a visiting scholar in 2009–2010. He joined Beijing University of Technology, Beijing, China, in 2013 as Beijing Distinguished Professor. His research interests are in Petri nets, enterprise information system, business process, data mining, information retrieval, semantic web, privacy protection, and big data. He has over 40 publications and 37 international patent applications (19 of them have been granted in China, US, or Japan). He served as PC members in multiple international conferences and organized the IEEE workshop on medical computing. He served as Guest Editor to organize a special issue on Telecommunication for Remote Medicine in China Communication.

Chunchen Liu received his Ph.D. degree in Computer Science from Jilin University, Jilin, China, in 2012, and is now working as an Associated Researcher in NEC Labs. China. Her research fields comprise Bayesian approximation inference, stochastic optimization, text analysis, semantic analysis.

Bo Liu received her B.S. degree in Department of Automation, Beijing Institute of Technology, China, in 2003, and M.S. and Ph.D. degree in System Integration Institute, Department of Automation, Tsinghua University, China, in 2008. She joined NEC Labs China after graduation and is currently a researcher in Spatio-temporal Data Analysis Research Department in NEC Labs China. She worked as a research professional in Computation Institute, the University of Chicago and Mathematics and Computer Science Division, Argonne National Laboratory in 2011–2012. Her research interests include Semantic Web, Ontology Reasoning, Big Data, Data Mining, Scientific Workflow and Cloud Computing. She is the author of more than 50 articles and inventions.

Rui Mao is an Associate Professor of the College of Computer Science and Software Engineering in Shenzhen University, China, and is the Executive Vice Director of the Guangdong Province Key Laboratory of Popular High Performance Computers and the Shenzhen City Key Laboratory of Service Computing and Applications. His research interest includes universal data management and analysis, distance-based indexing, high performance computing, and bioinformatics. Dr. Mao received BS (1997) and MS (2000) in computer science from the University of Science and Technology of China, and MS (2006) in Statistics and Ph.D. (2007) in Computer Science from the University of Texas at Austin. After three years work at the Oracle USA Corporation, he joined Shenzhen University in 2010. Dr. Mao has about 50 publications in internationally renowned journals and conferences. He proposed the pivot space model, a theoretical framework for distance-based indexing. This work was awarded the SISAP 2010 Best Paper award.

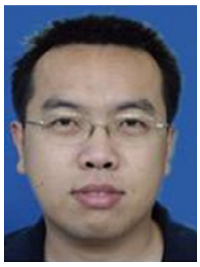
Yongcai Wang is an assistant researcher at Institute for Interdisciplinary Information Sciences (IIIS), Tsinghua University. He got his B.Sc. degree from Tsinghua University, China in 2006; Ph.D. degree from Tsinghua University China in 2011. He worked as an associate researcher in NEC Labs. China from 2007 to 2009. From 2009 to 2011, he was a post doctor at Institute for Interdisciplinary Information Sciences (IIIS). Since 2011, he is an assistant researcher at IIIS. He was a visiting scholar in Cornell University from February 2014 to August 2014. His major research interests lie in the broad area of sensor information fusion, network measurement, wireless locating algorithms, and systems. He co-authored several papers in major network science conferences and journals. He received the best paper awards at Ubicomm2009 and CWSN2011. He holds several Chinese and international patents on ultrasound locating systems. His research results "ultrasound locating system" were commercialized by NEC and exhibited in World Expo 2014.



Shi Chen, M.D. (born in 1980), was graduated from Peking Union Medical College and get the Doctor degree of Medicine in 2007. He is an attending doctor of Department of Endocrinology, Peking Union Medical College Hospital now. He is engaged in research of Medical Education and Endocrinology.

Ji-Jiang Yang got his B.S. and M.S. degree from Tsinghua University, and Ph.D. from National University of Ireland (Galway). His research areas involve in e-health, e-government/e-commerce, privacy preserving, Information resource management, data mining, cloud computing. Now he is an associate professor of Tsinghua University. Dr. Yang worked for CIMS/ERC (Computer Integrated Manufacturing System/Engineering Research Center) of Tsinghua University from 1995 to 1999. He had joined or been in charge of different projects funded by the State Hi-Tech program (863 program), NSF (China), and European Union. From 2009, Dr. Yang's main focus is e-Health and Medical Service. He has undertaken a few projects in the National Science & Technology Supporting Program about Digital Medical Service model and key technologies. He is also collaborating with a lot of medical institutions and hospitals. He is the member of Expert committee of IoT (Internet of Things) in Health at Chinese Electronic Association, Expert committee of Remote medicine and Cloud computing at Chinese Medicine Informatics Association. He has published more than 60 papers on professional Journals and Conferences.

Hui Pan, M.D. was graduated from Peking Union Medical College and get the Doctor degree of Medicine in 2004. He is an Endocrine professor of Department of Endocrinology, Peking Union Medical College Hospital, now. He is engaged in research of Medical Education and Endocrinology.



Qing Wang received his Ph.D. degree in Tsinghua University, and became a researcher in the Research Institute of Information Technology in Tsinghua University. His research interest covers the web service technology, data mining, machine learning, etc especially in the healthcare area. Recently, his research mainly focuses on the Big Data technology applied in medical service area.