Expert Systems with Applications 41 (2014) 366–372

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

An effective query recommendation approach using semantic strategies for intelligent information retrieval

Wei Song [a,b], Jiu Zhen Liang [a], Xiao Long Cao [b], Soon Cheol Park [b,*]

[a] School of IOT Engineering, Jiangnan University, Wuxi 214122, China
[b] Department of Electronics and Information Engineering, Chonbuk National University, Jeonju, Jeonbuk 561756, Republic of Korea

ARTICLE INFO

Keywords:
Query recommendation
Genetic algorithm
Knowledge discovery
Clustering
Information retrieval

ABSTRACT

With the explosive growth of web information, search engines have become the mainstream tools of information retrieval (IR). However, a notable problem emerged in the current IR systems is that the input queries are usually too short and too ambiguous to express their actual idea which largely affects the performance of IR systems. In this study, a novel query recommendation technology which suggests a list of related queries is proposed to resolve these problems. The query concepts can be firstly extracted from the web-snippets of the search result returned by the input query. A bipartite graph is subsequently built to identify the related queries, and the query similarity can be calculated by such bipartite graph. Moreover, by analyzing the URLs clicked by users, we find that some tokens appeared in URLs are very meaningful, especial for some typical topic-based pages. Therefore, these potential tokens which can provide a brief description from the subject of the URL are also considered. In order to reveal the real semantics between queries, the approach TF-IQF model is further discussed, and three features of a query, i.e. clicked documents, associated query and reversed query, are utilized in our approach in depth. Such a method could hopefully acquire the comprehensive idea of a query. To investigate how these three features could be used effectively for query recommendation in search engine, we adopt the benchmark evaluation criterions in our experiments, and the experimental results show its promising results in comparison with state of the art methods.

1. Introduction

As the web keeps expanding, the number of pages indexed in search engines increases correspondingly (Gholam & Ali, 2011; Zhang & Alexandra, 2005a, 2005b). So far we have obtained large amounts of available information and a high rate of new information have been updated, but contradictions in the available information, a low signal-to-noise ratio (proportion of useful information found to all information found), and inefficient methods for comparing and processing different kinds of information characterize the situation. The issue is no longer having enough information or not; it is just the opposite – too much information, in various formats and not all of similar value (Mohamed, 2011). The result is the information overload of the user, which is a well-recognized problem to the World Wide Web (Angela & Anne, 2000).

With such a large volume of data, it is increasingly difficult to find relevant information which can satisfy users' needs based on simple search queries (Azzah & Mark, 2011; Rahmatollah, Concepción, & Fletcher, 2008; Shen, Cheng, Chen, & Meng, 2008; Wang & Wu, 2008). Queries submitted to search engine by users tend to be short and ambiguous (Liao, 2008). It has found that the average length of queries submitted to search engines was only 2.35 terms (Jansen, Spink, & Bateman, 1998). Through the analysis of users' log from Chinese Sogou search engine (Wang, Chen, & Peng, 2004; Yu, Liu, Zhang, Ru, & Ma, 2007), researchers found that the average length of the queries was only 1.8 terms, while the proportion of query length less than 3 is nearly 93.15%. These short queries are not likely to precisely express what the user really needs. As a result, lots of pages retrieved may be irrelevant to the users' needs because of the ambiguous queries (Vaughan & Thelwall, 2004). On the other hand, users may not want to reformulate their queries by using more search terms, since it imposes additional burden on them during searching (Beg, 2005; Seik, Jonathan, & Janet, 2007).

In this paper, we propose a hybrid semantic strategy to evaluate query similarity based on user click-through data which is exploited to identify what the user is interested in. A user clicks on a search result mainly because the web-snippet contains the relevant topic that the user is interested in (Kumar & Kang, 2010). In this paper three approaches to compute query similarities, i.e. term similarities in vector space model (VSM) (Jean, Yves, & Philippe, 2011; Jing, Michael, & Huang, 2010; Rajan, Ramalingam, Ganesan, Palanivel, & Palaniappan, 2009), concept extraction and

[*] Corresponding author. Tel.: +82 63 270 2467; fax: +82 63 270 2461.
E-mail address: goldenview@foxmail.com (S.C. Park).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.052

TF-IQF models, are firstly and extensively discussed, and a hybrid semantic similarity strategy is subsequently proposed. Such a method consists of the following three major steps: (1) when a user submits a query, concepts (i.e. important terms or phrases in web-snippets) or tokens (i.e. important terms in the clicked URL) and their relations are mined from web-snippets to build a bipartite graph; (2) the query similarity is then calculated based on such constructed bipartite graph, and a hybrid similarity calculation method is proposed; (3) the most similar queries are suggested to the user for searching refinement. In experiments, some independent colleagues from our lab were firstly invited to implement test with the queries selected from a spectrum of topical categories for testing. We evaluate the performance of our approach using the standard measures, i.e. precision, recall and F-Measure, which are extensively adopted methods in research literature of IR and natural language processing (NLP).

The following parts of this paper are organized as below: Section 2 describes the related work. Section 3 depicts the details of concept extraction, TF-IQF model, and the construction of bipartite graph. In Section 4, a hybrid strategy for semantic similarity calculation is proposed. Experiment results and analysis are given in Section 5. Conclusions are given in Section 6.

2. Related work

On Web search engines, click-through data (Alptekin, 2012; Ma, Yang, & King, 2008) is a kind of implicit feedback from users. Clearly, it is a valuable resource for query recommendation (Bordogna, Alessandro, Giuseppe, & Stefania, 2012; Broccolo, Marcon, Nardini, Perego, & Silvestri, 2011; Liu, Miao, Zhang, Ma, & Ru, 2011; Yates, Hurtado, & Mendoza, 2005; Zhang & Nasraoui, 2008). Beeferman and Berger proposed an agglomerative clustering algorithm for exploiting user query log by clustering URLs and queries to find related queries (Beeferman & Bergge, 2000). They made use of bipartite graph shown in Fig. 1. The left nodes in Fig. 1 represent queries while the right nodes represent URLs clicked by a user. If a user clicks an URL, a link between the corresponding query and the URL is created on the bipartite graph. After the bipartite graph is obtained, an iterative algorithm is used to cluster two queries and two URLs successively. The disadvantage of this algorithm is that it cannot effectively deal with some so-called noisy data, that is, if the user clicked an URL wrongly, two non-related queries will be assigned together forever (Chan, Leung, & Lee, 2004).

In order to calculate the similarity between queries, Wen et al. also considered the similarity between clicked documents (Wen, Nie, & Zhang, 2002). They suggested that the two queries should be clustered together, if they contain the same or similar terms, and lead to the selection of the same URLs. However, since the queries are usually short and merely very few URLs that point to the same document, their method may not be effective for disambiguating Web queries. And it requires a pre-built document classification system with relatively high classification accuracy.

In summary, these two methods have a common major problem, that is to say, the number of common clicks on URLs for different queries is rare, only a few of popular queries have sufficient information for mining their common clicked URLs while distance matrices between most queries are very sparse. As a result, many queries with semantic similarity might appear orthogonal in such matrices. Thus, the chance for the users to see the same results would be small, let alone clicking on them. To alleviate this problem, Leung et al. introduced the notion of concept-based graphs by considering the concepts extracted from web-snippets and adopted Beeferman and Berger's method to this new context (Leung & Lee, 2010; Leung, Ng, & Lee, 2008). The use of concepts helps reduce the size of the resulted profiles while retaining the accuracy and capability to capture users' interests. Liu et al. proposed a method for query recommendation with TF-IQF model which is different from others (Liu, Jiang, & Chen, 2008). Such a model divided the URLs into tokens by splitting the URL string with some separators and measured the weight of these tokens to calculate their similarity which will be thoroughly discussed in next section.

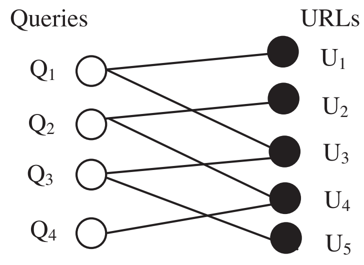

Fig. 1. Query-URL bipartite graph.

3. Similarity calculation

3.1. Concept extraction

The concept extraction method is inspired by the well-known problem of finding frequent item sets in data mining (Goethals & Zaki, 2003; Lahiri & Tirthapura, 2010; Manerikar & Palpanas, 2009). When a user submits a query to a search engine, a series of web-snippets will be returned to the user. The assumption is as follows: if a keyword or phrase t_i appears frequently in the web-snippets of a particular query q, then we can regard t_i as a concept related to q, because it coexists in close proximity with q. We use the following cutoff formula to measure the interest of t_i with respect to the returned web-snippets arising from q.

$$cutoff(t_i) = \frac{sf(t_i)}{n} \times |t_i| \tag{1}$$

where n is the total number of web-snippets returned, $sf(t_i)$ is the number of web-snippets containing t_i, and $|t_i|$ is the number of terms in t_i. In order to extract candidate concepts for q, first we need to get the set of keywords or phrases T from the web-snippets returned by q. After T is obtained, we can use formula (1) to compute the cutoff for each t_i ($t_i \in T$). If the cutoff of t_i is bigger than the appropriately assigned threshold s ($cutoff(t_i) > s$), then t_i is treated as a candidate concept for q. Suppose we input a query q_{pp}, i.e. Potala Palace, to Sogou search engine, Table 1 illustrates the extracted candidate concepts for query q_{pp}.

The extracted candidate concepts usually distribute in all of the web-snippets, however, it is natural that users are used to browsing and searching results from the first page (usually the top 10 web-snippets) returned by search engine. Thus, from each of the top 10 web-snippets, the extracted target concepts will be seen as the right nodes of the Query-Concept bipartite graph.

Table 1
The concepts extracted from query "Potala Palace".

Concepts	Support
Tibet	0.2
Lhasa	0.2
Potala Palace	0.2
Songtsen Gampo	0.14
Princess Wencheng	0.12
Buildings	0.12
Capital	0.1
Journey	0.1
Tang Dynasty	0.06
Mabu	0.06
Putuo	0.06