Novel Query Expansion Method based on User Interest Context and Ontology

Lizhou Feng[1, a], Wanli Zuo[1, b], Youwei Wang[1, c]

[1] College of Computer Science and Technology, Jilin University, Changchun 130012, China

[a]email:331902794@qq.com, [b]email:413326209@qq.com, [c]email:1036569449@qq.com

Keywords: contextual word, query expansion, user interest, ontology

Abstract. We proposed a novel query expansion method by combining user interest and ontology. Firstly, users' interests are described by contextual words which are generated based on ontology, and the user interest degree with respect to each contextual word is calculated. Secondly, the contextual words are organized according to ontology relevance and divided into different subsets, and each subset can be seen as a candidate suggestion set. By calculating the weight of each contextual word, we obtain the meaningful expansions for a query. Comparative experiments show that, the proposed method is superior to other methods when precision and recall measurement are used and gives personalized query suggestions to users efficiently.

Introduction

The effectiveness of information retrieval from the web largely depends on whether users can issue queries to search engines, which properly describe their information needs [1]. Writing queries is not very easy, because the queries are usually short and the words may be ambiguous [2, 3]. Most existing works on query expansion utilize query logs to suggest queries [4]. Generally, the web search engines have millions of users. When a user has some information needs, there always exist many users who have searched the same query before. Therefore, the search engine can use these large amounts of past usage data to offer possible query expansions [5]. Because the query submitted by the user is closely related to his interests and intents, different users who submit a same query may want to express different requirements.

Effective query expansion requires inferring user's query intent and then expanded queries that help retrieving webpages which contain the relevant information [6]. Inspired by this, we propose a method of query expansion based on user interest context and ontology. It does not depend on query logs of the whole web and utilizes only the terms occurring in the user browsed logs.

The proposed method

In this paper, the proposed query expansion method is executed in two steps: the user interest context mining and the query expansion. The details are given as follows.

(1) User interest context mining

Firstly, we execute webpage parsing to extract the main body of the webpage. Stop words are filtered out and the root of each word is extracted by using the Porter Stemming algorithm [7]. The webpage p_i is represented by the vector $W_i=(w_{i1},w_{i2},\ldots w_{im})$, where w_{im} is the term of p_i; Secondly, we use the natural language processing technology to implement word sense disambiguation[8]. Further, we obtain the hypernyms of the terms which are called contextual words and denoted as $C_i=(c_{i1},c_{i2},\ldots c_{im})$ in p_i through generic ontology, where, c_{im} is the contextual word of w_{im}.

To calculate the user interest degree of contextual word, the browsed webpages are organized by the day, and each day is seen as one session. The webpages user u browsed in j-th session is denoted as Day_j. The interest degree of contextual word c is formulated as follows, denoted as $I(c)$:

$$I(c)=\sum_{j=1}^{n}\left(\alpha\cdot f(c,Day_j)+\beta\cdot t(c,Day_j)\right)*e^{-\frac{\log 2}{l}(d-d_{init})} \quad (1)$$

Where $f(c,Day_j)$denotes the access frequencies of webpages which contain contextual word c in

Day_j, $t(c,Day_j)$denotes the continued access time of webpages which contain c in Day_j. User's attention on one interest will attenuate over time, so we added attenuation factor $e^{-\frac{\log 2}{l}(d-d_{init})}$ [9], where d_{init} is the time point of the first occurrence of c, d is the current time point, l is the span parameter, which indicates l sessions. α and β denote constant values which satisfy $0<\alpha, \beta\leq 1$.

(2) Query expansion

Given two terms c_x and c_y, they are said to have an ontology relevance, if c_x and c_y are the hypernym or hyponym for each other, or they have the same hypernym or hyponym, otherwise, they are synonyms. Contextual words are organized according to ontology relevance and are divided into different subsets. For each initial query q, we can find all the contextual words which have ontology relevance with q, and can be seen as candidate expansions. Once we get the candidate expansions, the weight for each expansion c_x is calculated by the following equation:

$$Weight(c_x) = A * Dis(q, c_x) * I(c_x) \tag{2}$$

Where, $Dis(q, c_x)$ denotes the distance between q and c_x, and it can be defined as following:

$$Dis(q, c_x) = \sum_{p_m \in P_x} \frac{f_{(ix, p_m)}}{\max\left\{ f_{(1x, p_m)}, f_{(2x, p_m)}, \cdots, f_{(rx, p_m)} \right\}} \cdot \log\left(\frac{count(q, c_x)}{count(c_x)} + 1\right) \tag{3}$$

Where, P_x is the webpage set whose contextual words contain c_x, $f_{(ix,pm)}$ is the times of q_i occurs in the webpage p_m, r is the number of terms contain in p_m, $count(q_i, c_x)$ is the times of q_i and c_x co-occurrence, that is the number of webpages which contain q_i and whose contextual words contain c_x, $count(c_x)$ is the times of c_x occurrence, that is the number of webpages whose contextual words contain c_x. And $I(c_x)$ is the interesting degree. We update the list of candidate expansion of q_i through the weight, and the meaningful expansions for a query can be selected by setting threshold.

Test results

The primary data in this paper is the browsing logs visited by 10 users, who opted in to provide data through a widely-distributed browser toolbar in three months. These log entries include a unique identifier for the user, a timestamp for each page view and the URL of the webpage visited. The initial queries which need to be expanded are all provided by the 10 users from the real user query records, and can reflect user interest completely.

(1) Evaluation of user interest mining

Table 1 shows three randomly selected user interests for a user, which are computer, travel and health. We only show a partial of contextual words with the high-frequency. It can be seen that these contextual words reflect individual user's interest clearly.

Table 1. A partial of high-frequency contextual words in three user interests

computer	travel	health
search engine	hotel	dieting
website	travel agencies	weight loss
social network	airline	low-fat diets
Twitter	ticket	running
community	hostel travel	nutrition
program language	tourist attractions	vitamin
java	southwest	caloric
linux	Travelocity	yoga
program	Seattle	exercise
software platform	local delicacies	dental hygienist

(2) Comparisons methods and analysis

Comparative experiments are carried out by using the following three methods.

Method 1: The method does none expansion for queries.

Method 2: A commonly used query expansion method [10] is to find similar queries in search logs and use those queries as expansions for each other.

Method 3: The proposed method.

Experiment one: The purpose is to test the relevance of the queries expanded. We mark the relevance of the expanded queries, where 0 indicates irrelevance, 0.5 points partial relevance and 1 indicates completely relevance. For query q, the relevance score RS_q is calculated using formula (4):

$$RS_q = \sum_{i=1}^{k} s_i \Big/ k \tag{4}$$

Where k is the number of queries expanded for q, s_i ($1<=i<=k$) is the score of the expanded query q_i of q. Allowing users marked relevance scores of the top k (k=1, 3, 5, 7, 10) expanded queries for q, 100 times query tests are executed, comparison of the average relevance score (denoted as RS_a) between Method 2 and the proposed method is shown in Fig 1. As can be seen, the more irrelevant queries are expanded as the increase in the number of expanded queries, the worse the results of the proposed method and Method 1 are. However, our method is significantly better than Method 2 because that it takes into account the impact of user interest and uses the ontology.

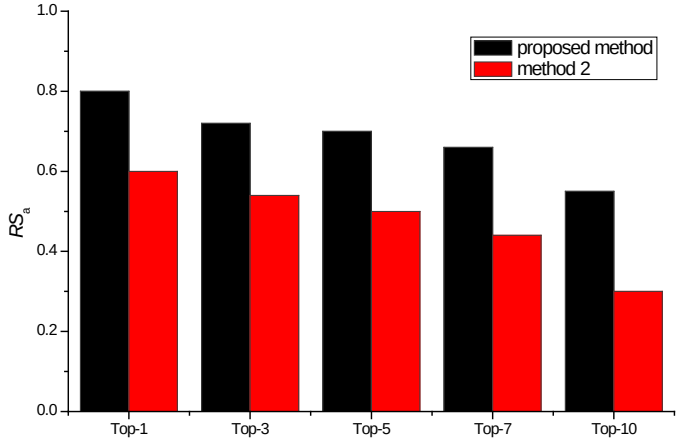

Fig. 1. RS_a values of 10 users when different methods are used

Experiment two: The purpose is to test whether the proposed method can improve the performance of existing search engine. For query q, the expansion queries which are generated by Method1, Method2 and the proposed method are retrieved in the same search engine. We calculated the precision and recall value to measure each method, formula is as follows：

$$precision = \frac{number\ of\ retrieved\ related\ webpages}{number\ of\ retrieved\ webpages}, \quad recall = \frac{number\ of\ retrieved\ related\ webpages}{number\ of\ all\ related\ webpages} \tag{5}$$

Most users always care about the top-k results in the retrieval process, therefore, we calculate the precision values corresponding to the ten given recall values. According to the user interest, 100 times query tests are carried out, and the averages precision values (denoted as p_a) are given in fig.2 when different methods are used.

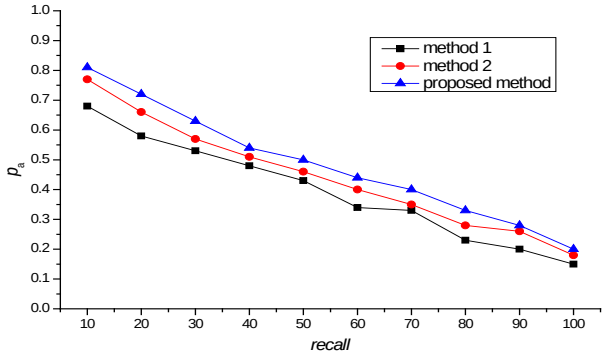

Fig. 2. p_a values of 10 users when different methods are used

Obviously, Method 1 performs the worst, deducing that it does not expand terms for queries. Moreover, the results of the proposed method is generally higher than those of Method 2, deducing that it takes into account the user interest is closer to the user's query intent. In addition, the using of ontology makes the suggestion terms more targeted.

Conclusion

We proposed a new query expansion method based on user interest and ontology. The traditional query expansion is not concerned with the needs of users; however, the proposed method can provide query expansion by using user interest. The hypernyms of the terms in the webpages are obtained by using ontology and different contextual words are added into the expansion set which is conducive to the expression of implicit user intent. The efficiency of the proposed method was examined by comparing it with two typical methods. The results show that the proposed method is superior to other methods when the relevance score, precision and recall measurements are used.

Acknowledgement

In this paper, the research was sponsored by the National Nature Science Foundation of China No. 60973040, the National Nature Science Foundation for Young Scientists of China No. 60903098, the Scientific and Technological Project of Jilin Province, China No. 20130206051GX.

References

[1] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He. Context-Aware Query Suggestion by Mining Click-Through. In KDD'08, 875-883, 2008.

[2] Ji rong Wen, Jian yun Nie, Hong jiang Zhang. Clustering user queries of a search engine. In WWW'01, 162-168, 2001.

[3] Hang Cui, Ji rong Wen, Jian yun Nie, et al. Probabilistic query expansion using querylogs. In WWW'02, 325-332,2002.

[4] P. Boldi, F. Bonchi, C. Castillo, et al. Query suggestions using query-flow graphs. In WSCD 2009, 56–63, 2009.

[5] B. Sumit, M. Debapriyo, M. Prasenjit. Query Suggestions in the Absence of Query Logs. In SIGIR'11, 795-804, 2011.

[6] S. Eldar, M.Jayant, W. Lu, et al, Clustering Query Refinements by User Intent. In WWW'10, 841-850, 2010.

[7] P. Willett, The porter stemming algorithm: then and now, Program: electronic library and information systems, 40 (3): 219-223, 2006

[8] Liu Xiulei, Liao Jianxin. Lexical Analysis Based on Combining Senses in Ontology Matching[J]. Acta Electronica Sinica, 40(8): 1624-1630, 2012.

[9] S. Kazunari, H. Kenji, Y. Masatoshi. Adaptive Web Search Based on User Profile Constructed without Any Effort from Users. In WWW'04, 675-684, 2004.

[10] Baeza-Yates, et al. Query recommendation using query logs in search engines. In EDBT 2004, 588-596, 2004.