Query Expansion Based on Query Log and Small World Characteristic

Yujuan Cao[1,2], Xueping Peng[1], Zhao Kun[1], Zhendong Niu[1,3], Gx Xu[1], and Weiqiang Wang

[1] The School of Computer Science and Technology, Beijing Institute of Technology, Beijing, 100081, China
[2] Beijing Command & Control Center, Beijing, 100094, China
{qiushuichangtian,pengxp,kzhao,zdniu,xuguixian2000, princeWang}@bit.edu.cn

Abstract. Automatic query expansion is an effective way to solve the word mismatching and short query problems. This paper presents a novel approach to Expand Queries Based on User log and Small world characteristic of the document (QEBUS). When the query is submitted, the synonymic concept of the query is gotten by searching a synonymic concept dictionary. Then the query log is explored and the key words are extracted from the user clicked documents based on small world network (SWN) characteristic. By analyzing the semantic network of the document based on SWN and exploring the correlations between the key words and the queries based on mutual information, high-quality expansion terms can be gotten. The experiment results show that our technique outperforms some traditional query expansion methods significantly.

Keywords: Query expansion, query log analysis, small world characteristic, mutual information.

1 Introduction

Search engines have become the main tool for information retrieval. Most of search engines still rely on the key words in the query to search and rank Web Pages. However it is general consensus that the authors and the users always use different terms to describe the same concept, and web users always submit short term queries. These are the key reasons that affect the precision of the search engine. Query expansion is an efficient approach to solve the problem by automatically adding additional terms to the query.

Although web users usually input short queries with little or no context information associated with them, they click the URL which they consider relevant. These clicks associate a set of query terms with a set of WebPages and can provide high level suggestions for expanding the original user query with additional context.

This paper investigate a new approach (QEBUS) Expand Queries Based on User log and Small world characteristic of the document. Our work is different from the traditional approach in three aspects.

G. Vossen, D.D.E. Long, and J.X. Yu (Eds.): WISE 2009, LNCS 5802, pp. 573–580, 2009.

(1) For the queries not in the user logs, we maintain a synonymic concept dictionary. By searching the dictionary, both the query and the synonymic concept are submitted to the search engine. As the user log becomes larger, the dependence on the concept dictionary will get weaker.

(2)For the queries in the user logs, the WebPages clicked by the users with same preference are selected. The keywords are extracted based on small world characteristic of the document. This step can improve the quality of extracted terms dramatically.

(3) Correlations between the key words and the queries are explored not only based on mutual information but also on key words distribution in related documents.

Our experiments show that our query expansion approach can improve the precision of the search result significantly.

The remainder of this paper is organized as follows. In Section 2, we discuss the related work. QEBUS are described in Section 3. Experiment results and evaluations are presented in Section 4. Conclusion and future work are introduced in Section 5.

2 Related Work

Traditional works on query expansion can be divided into three categories: global analysis, local analysis (pseudo-relevant) and local context-sensitive analysis.

Global analysis examines the relationship of words in the whole collection, includeing Deerwester's Latent Semantic Index (LSI)[4], Y. Jing's PhraseFinder[15] and the approach presented by Fabienne[6]. The disadvantage of these approaches is that corpus-wide statistic is required, so it is expensive in terms of space and time.

Local analysis is the approach extracting terms from top-N documents retrieved by the initial queries instead of global concept database. The most frequent non-stop words among the top ranked passages are counted and added to the original queries [14,2,10]. Local analysis highly depends on the quality of the documents retrieved in the initial retrieval.

One of the best ways to determine search needs is through user's observation. Cui [8],Fonseca and Golgher [1] utilize user click through data to extract semantic similarity. Radlinski[5], Rosie[12] use query sessions as the source information for query expansion. By studying the web search behavior, White[13] found that experts are more successful in finding what they are looking for than non-experts. Instead of teaching a non-expert to be an expert, our QEBUS helps users with query expansion.

3 Key Words Extraction and Query Expansion Model

3.1 Small World Phenomenon

The small-world experiment was conducted by Milgram who examined the average path length for social networks of people in the United States. The research was groundbreaking in that it suggested that human society is a small world type network characterized by short path lengths and high cluster, which are often associated with the phrase "six degrees of separation".

Recent research on networks that occur in a number of biological, social and man-made systems showed that they share a common feature. To formalize the notion of a small world, Watts and Strogatz [3] define the clustering coefficient *C* and the characteristic path length *d* of the graph.

The clustering coefficient *C* is a measure of the clique of the local neighborhoods. The characteristic path length *d* is the average path length over all pairs of nodes. The graphs that have SW property are often neither completely regular nor completely random. Graphs with SW structure are highly clustered, but the path length between nodes is small (It can be seen that $C \gg C_{random}$ and $d \approx d_{random}$).

Yutaka Matsu [16] showed that the graph derived from a document also has the small world characteristics. It has been proved by Ramon Ferrer [11] that the small world phenomenon also exists in human language.

3.2 Building Term Co-occurrence Graph

A term co-occurrence graph can be constructed from a document as follows.

(1) After word segmentation (for Chinese document) and stop words removing, we select frequent terms $\{t_i\}$ which appear frequency $f>f_{thr}$. With this method, the nodes of the Term Co-occurrence Graph were selected.
(2) For a given term (node), which nodes should be selected as its neighbors? Jaccard coefficient is appropriate for our feature extraction task and has previously shown excellent empirical performance in natural language processing. For every pair of $\{t_i, t_j\}$, Jaccard coefficient $J_{ti,tj} = \frac{n_{ti,tj}}{n_{ti}+n_{tj}-n_{ti,tj}}$ is calculated. Where $n_{ti,tj}$ is the number of sentences that contain both t_i and t_j, $n_{ti} + n_{tj} - n_{ti,tj}$ is the number of sentences that contain either of t_i or t_j. If $J_{t_i,t_j} > J_{thr}$(J_{thr} is the user-given threshold), an edge is added between t_i and t_j.
(3) The term co-occurrence graph of the document can be defined as $G_L = (T_L, E_L)$, where $T_L = \{t_i\}$, t_i is the terms selected after step (1), $E_L = \left\{\{t_i, t_j\}\right\}$ is the set of edges or connections between terms $t_i$and t_j, *L* is the number of nodes in the graph.

3.3 SW Properties of Term Co-occurrence Graph

After constructing the term co-occurrence graph, we will calculate its two basic statistical properties: the clustering coefficient *C* and the path length *d*. For a term $t_i \in T_L$ with k neighbors, we use $\Gamma_i = \left\{j \middle| \xi_{i,j} = 1\right\}$ to indicate the set of nearest neighbors. $\xi_{i,j} = 1$ indicates there is an edge between t_j and t_i otherwise $\xi_{i,j} = 0$. The clustering coefficient of the graph can be defined as $C = \frac{1}{L}\sum_{i=1}^{L} C_i$, where C_i is the node $t_i's$ clustering coefficient. $C_i = \frac{\varphi_i}{k\times(k-1)\div 2}$, k is the number of t_i 's neighbors; φ_i is the number of edges which actually exist between the neighbors of t_i, $\varphi_i = \sum_{m,n=1}^{k} \xi_{m,n} | t_m, t_n \in \Gamma_i, m \neq n$.The second property is the path length. For given two terms $t_i, t_j, \in T_L$, let $d_{min}(i,j)$ represent the minimum path length between

them. For term t_i, the path length can be calculated by $d_i = \frac{1}{L-1}(\sum_{j=1,j\neq i}^{L} d_{min}(i,j))$. The average path length of the term co-occurrence graph is $d = \frac{1}{L}(\sum_{i=1}^{L} d_i)$.

3.4 User Log Description

User access log are derived from the database in one of our laboratory research project IICSS (Internet Information Crawl and Services System), including user id, query terms, user clicked URL, user category and the visiting time. Table1 describes the log information. Where userCatagory is the Category ID of the user (User selected the second level categories he interested in from the Open Directory Project manually).

Table1. Information included in web access log

#	Query terms	visitTime	userID	userCata	User clicked URL
1	Topol (Bai Yang)	08-09-18 10:33:48	Adm 001	A13,A15	http://news.xinhuanet.com/mil/2008-02/28/content_7685125.htm
2	Topol (Bai Yang)	08-09-18 10:34:40	Adm 001	A13,A15	http://www.space.cetin.net.cn/docs/ht9903/ht990316.htm
	…	…	...		…

3.5 Key-Words Extracting

The small world property of the document gives us some inspiration. The distribution of terms in the document is not equal. Each term has different contribution to the content and structure of the document. When expressing his idea, the author may repeat some concepts and then extend document basing on them. The key words represent the main topic and the fundamental concepts of the document.

Key words are the key nodes in the term co-occurrence graph. For the convenience of description, we cite the definitions adopted by Zhu [9]:

Definition 1: *CN* is the original terms co-occurrence graph which was constructed after word segmentation and stop word removing, and *d* is the average path length of the *CN*.

Definition 2: CN_i is the terms co-occurrence graph where the i^{th} node is absent, and d_i is the average path length of the CN_i.

Definition 3: $CB_i=d_i-d$ is the contribution of the i^{th} node. The nodes with larger CB_i are more important to keep the graph well connected.

In the case of term co-occurrence graph, the terms with high CB_i are the 'short cuts' connect vertices. If the node with large CB_i is absent in the graph, the average length of the graph will get very large. In the context of documents, if the terms with large CB_i is absent, the topics are divided and the basic concepts are lost connections. So by finding the terms with high CB_i, the key-words can be extracted accordingly.

3.6 Query Expansion

Our statistic shows that more than 90% query terms are extracted as the key words from the user clicked document. By exploring the relationship between the query terms and

the key words in users' access documents, terms which have close relationships to the original query can be gotten.

Users in the same category have same preference. But the users in different category have different interests. For a certain query, the WebPages clicked by users in category 1 may far from the WebPages clicked by users in category 2. For example, when a user inputs a Chinese query 'Chang'E', many WebPages related to 'Chang'E' will present to him. Some of the WebPages are the classical stories about a Chinese goddess live in the moon; others are news about the 'Chang'E' moon satellite of China. The literary lovers will click the classical stories with high property but an astronomic enthusiast will look through the news about 'Chang'E' satellite instead.

For a certain query, from the documents clicked by users in same category, we extract 20 key words using previous method. $t_{1,1},t_{1,2},t_{1,3}\ldots t_{1,20}$ are the key words extracted from the document1, $t_{2,1},t_{2,2},t_{2,3}\ldots t_{2,10}$ are the key words extracted from document2, After removing overlap terms, A new list $Vec=(T_1,T_2,T_3\ldots T_k)$ is gotten.

The relativity between query terms and key words can be evaluated on the basis of mutual information and key words distribution in related documents. The correlation between query terms and key words can be expressed as:

$$Rel(q,T_i)=\frac{N_{userclick\_T_i}}{N_{userclick}}\times\frac{N_{q,T_i}}{N}\times log\frac{N_{q,T_i}}{N_qN_{T_i}} \quad (1)$$

Let's define $D_{userclick}$ as the users with similar interest clicked documents when submit query q. $N_{userclick}$ is the total number of $D_{userclick}$, $N_{userclick}$ is the number of documents which contain T_i in the $D_{userclick}$, $N_{q,Ti}$ is the number of documents where query q and term T_i are co-occurrence in the corpus, N_q is the number of documents which contain q in the corpus, N_{Ti} is the number of documents which contain the term T_i in the corpus, and N is the total number of documents in the corpus. By search $\{q,T_i\},\{q\}$ and $\{T_i\}$ in our BIT-Search-Engine system (one of our laboratory research project), $N_{q,Ti}$, N_q and N_{Ti} are easy to be gotten.

4 Experiment

4.1 Experiment Data Set

Experiment data set comes from one of our laboratory research project CICS (Internet Information Crawl and Services System), which contains about 2,768,763 WebPages and more than 10 thousand query requests are recorded in the user log. There are 19 undergraduate students as volunteers who give their feedbacks to this experiment. The length of queries in our experiments is very close to those employed by the real web users and the average length of all queries is 2.1 words. 16 queries used in the experiments are listed in Table 2.

Relevant documents are judged according to the volunteers' manual selections and standard relevant document sets are prepared for all of the 16 queries.

Table 2. 16 queries used in the experiment

Topol(Bai Yang)	Nuclear submarine
unmanned aerial vehicle	Space Shuttle Endeavor
Chang'E	Chandrayaan-1 spacecraft
Shenzhou VII	Bulava
F-35	ARJ21-700
Somali pirate	Phoenix Mars probe
Large Aircraft Company	Airshow
Apollo program	BeiDou Satellite

4.2 Quality of Expansion Terms

In Chinese, Topol (Bai Yang) is a word has multi-meanings. It can be the Topol-M missile; the name of a table tennis player; or a famous prose written by Mao Dun. For the users who are interested in military topics, query "**TOPOL**" is mainly related with a missile of Russia. Some very good terms, such as **"TOPOL-M", "Russia", "missile silo"**, **"intercontinental ballistic missile"**, even **"GLONASS", "DF-31"** can be obtained by our techniques.

We chose TF/IDF as the base line extracting key words from the document. Relevant terms are judged according to the volunteers' manual selections. Table 3 shows the percentage of the relevant terms in the top 36 suggested by Local Context Analysis and our method based on small world network and user logs. As we can see, the terms expanded by our method have better quality.

Table 3. Percentage of relevant terms

Query	TF/IDF (base line)	QEBUS	Query	TF/IDF (base line)	QEBUS
Topol(Bai Yang)	47.2%	69.4%	Nuclear submarine	55.6%	72.2%
unmanned aerial vehicle	63.9%	75.0%	Space Shuttle	52.8%	61.1%
Chang'e	50.0%	66.7%	Chandrayaan-1	55.6%	69.4%
Shenzhou VII	52.8%	72.2%	Bulava	63.9%	80.6%
F-35	52.8%	58.3%	ARJ21-700	66.7%	83.3%
somali pirate	50.0%	63.9%	Phoenix Mars	52.8%	75.0%
Large Aircraft Company	52.8%	61.1%	Airshow	55.6%	72.2%
Apollo	47.2%	77.8%	BeiDou	50.0%	66.7%

4.3 The Effectiveness of Query Expansion

We use the popular precision score in IR and authority score following Xue[7] to evaluate the results before and after query expanding. For a given query Q, let $|D|$ be the size of relevant WebPages to the query. Let TOP be the top N documents retrieved by our system. *Precision* can be defined as:

$$precision = \frac{|D \cap TOP|}{|TOP|} \tag{2}$$

For a given query Q, we ask our volunteers to identify top 10 authoritative pages according to their own judgments. Let A be the set of 10 authority WebPages to the query Q, and N be the set of top 10 documents retrieved by our system. *Authority* can be defined as:

$$authority = \frac{|A \cap N|}{|A|} \tag{3}$$

Precision measures the degree of accuracy of the algorithm, while *authority* measurement is more relevant to users' degree of satisfactory on the performance of the search engine.

We chose no query expansion as the base lines, compare the local context analysis extracting keywords by TF/IDF (LTF/IDF); the local context analysis extracting keywords by SWM (LSWM); Query Expansion Based on User Log extracting keywords by TF/IDF(QETF/IDF) and QEBUS, Overall *precision* and *authority* is presented in Fig.1 and Fig.2.

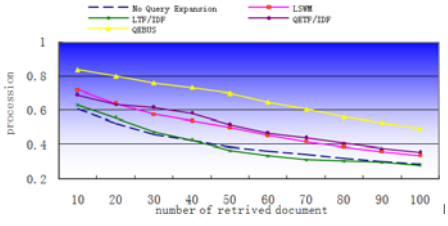

Fig. 1. Precision for no query expansion, LTF/IDF, LSWM, QETF/IDF and QEBUS

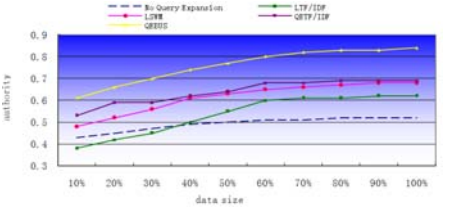

Fig. 2. Authority for no query expansion, LTF/IDF, LSWM, QETF/IDF and QEBUS

QEBUS shows best precision and authority scores over other approaches. One reason is local context analysis searches expansion terms in the top-ranked retrieved documents and is more likely to add some irrelevant terms into the original query, but QEBUS selects expansion terms in a relatively narrower but more concentrated area based on user log. The other reason that expansion terms extracted based on SWN are more relevant to the original queries. But for the TF/IDF method, the noise metadata is introduced with high probability.

The experiment results show that our query expansion approach based on user log and small world characteristic of the document achieves best performance. It brings 69.6% and 54.1% improvement in both precision and authority over the base line.

5 Conclusion and Future Work

In this paper, we presented a query expansion method (QEBUS) based on user log and small world characteristic of the document.. Experiments show that QEBUS can achieve substantial performance improvements.

Here we expand queries in the document which the users in same category clicked, that is simple but somewhat coarseness. Future work includes personalized query

expansion by accurately finding the document match the user's interest in the user log when he input a query.

Small World Characteristic of document is not only helpful to query expansion but also significant to text category, topic analysis, texts digest etc. We believe this is a very promising research direction.

References

1. Fonseca, B.M., Golghe, P.: Concept Based Interactive Query Expansion. In: Proceedings of CIKM 2005, pp. 696–703 (2005)
2. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New Retrieval Approaches Using SMART. In: Proceedings of the TREC 4 Conference, pp. 25–48 (1995)
3. Watts, D., Strogatz, S.: Collective dynamics of small-world networks. Nature, 440–442 (1998)
4. Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the Society for Information Science 41(6), 391–407 (1990)
5. Radlinski, F., Joachims, T.: Query chains: learning to rank from implict feedback. In: Proceedings of the Eleventh ACM SIGKDD, pp. 239–248 (2005)
6. Moreau, F., Claveau, V., Seillot, P.: Automatic Morphological Query Expansion Using Analogy-Based Machine Learning. In: Advances in information Retrieval, pp. 222–233 (2007)
7. Xue, G.-R., Zeng, H.-J., et al.: Optimizing Web Search Using Web Click-through Data. In: Proceeding of CIKM 2004, pp. 118–126 (2004)
8. Cui, H., Wen, J.-R., Nie, J.-Y., Ma, W.-Y.: Probabilistic query expansion using query logs. In: Proceedings of the Eleventh International Conference on WWW, pp. 325–332 (2002)
9. Zhu, M.X., Cai, Z., Cai, Q.S.: Automatic Keywords Extraction of Chinese Document Using Small World Structure. In: Proceeding of Natural Language Processing and Knowledge International Conference, pp. 26–29 (2003)
10. Theobald, M.: Efficient and Self-tuning. Incremental Query Expansion for Top-k Query Processing. In: Proceeding of the SIGIR 2005, pp. 242–249 (2005)
11. Cancho, R.F., Sole, R.V.: The small world of human language. In: Proceedings of the Royal Society of London, pp. 2261–2265 (2001)
12. Jone, R., Rey, B.: Generating Query Substitutions. In: Proceedings of International Conference on WWW, pp. 387–396 (2006)
13. White, R.W., Dumais, S.T., et al.: Characterizing the Influence of Domain Expertise on Web Search Behavior. In: Proceeding of WSDM 2009, pp. 132–141 (2009)
14. Xu, J.X., Croft, W.B.: Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems 18(1), 79–112 (2000)
15. Jing, Y., Croft, W.B.: An association thesaurus for information retrieval. Proceedings of RIAO 1994, 146–160 (1994)
16. suo, Y.M., Ohsawa, Y.: KeyWorld: Extracting Keywords in a Documents as a Small World. In: Proceedings of DS-2001, pp. 271–281 (2001)