An Ontology-Based Approach to Query Suggestion Diversification

Hai-Tao Zheng*, Jie Zhao, Yi-Chi Zhang, Yong Jiang, and Shu-Tao Xia

Tsinghua-Southampton Web Science Laboratory,
Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
{zheng.haitao,jiangy,xiast}@sz.tsinghua.edu.cn,
angelahai@126.com, zhangyichi0604@gmail.com

Abstract. Query suggestion is proposed to generate alternative queries and help users explore and express their information needs. Most existing query suggestion methods generate query suggestions based on document information or search logs without considering the semantic relationships between the original query and the suggestions. In addition, existing query suggestion diversifying methods generally use greedy algorithm, which has high complexity. To address these issues, we propose a novel query suggestion method to generate semantically relevant queries and diversify query suggestion results based on the WordNet ontology. First, we generate the query suggestion candidates based on Markov random walk. Second, we diversify the candidates according the different senses of original query in the WordNet. We evaluate our method on a large-scale search log dataset of a commercial search engine. The outstanding feature of our method is that our query suggestion results are semantically relevant belonging to different topics. The experimental results show that our method outperforms the two well-known query suggestion methods in terms of precision and diversity with lower time consumption.

Keywords: Query suggestion, search logs, semantic relationships, diversify.

1 Introduction

Since it is getting difficult for search engines to satisfy users' information needs directly, query suggestion is proposed to generate alternative queries and help users explore and express their information needs. Query suggestion plays an important role in improving the usability of search engines and it has been well studied in academia as well as industry. However, much efforts focus on how to suggest queries relevant to the original query without considering the diversity of query suggestion results. A study shows that queries submitted to search engines are usually short consisting of two or three words on average [15] . When a user knows little about the search topic, it is difficult to construct a good query.

* Corresponding author.

C.K. Loo et al. (Eds.): ICONIP 2014, Part II, LNCS 8835, pp. 437–444, 2014.

The two situations make search intend ambiguous, and query suggestion results diversity is necessary. Recently, some methods has been proposed to diversify query suggestion results. For example, Ma et al. [10] diversify the results by employing the Markov random walk and hitting time.

However, existing query suggestion diversification methods generally use greedy algorithm, which has high complexity, resulting relative low efficiency. In addition, many existing query suggestion methods only use search logs to obtain query suggestion without considering semantic relationships between queries, resulting in some query suggestion results unrelated to the original query semantically.

To address these issues, we propose a novel method, called Ontology-based Diversifying Query Suggestion (ODQS), to diversify query suggestion results using an ontology - WordNet [16]. First, we generate query suggestion candidates from search logs based on the forward random walk analysis. Second, we diversify the candidates based on the different senses of the original query in the WordNet. The paper has three main contributions as follows:

1) We exploit search logs and semantic relationships between queries based on WordNet to diversify query suggestion results and increase the relevance of query suggestion results.

2) Our method improves the efficiency of the query suggestion diversification by employing the WordNet ontology.

3) We evaluate the effectiveness and efficiency of the proposed method by comparing two existing query suggestions methods.

The rest of this is organized as follows. Section 2 is devoted to a detailed description of our method to diversify query suggestion results. We conduct experiment on AOL search logs to evaluate our proposal section 3. We discuss related work of query suggestion in Section 4. Finally, we conclude the paper with future work in Section 5.

2 Ontology-Based Diversifying Query Suggestion

In this section, we elaborate how to leverage search logs and semantic relationships to improve the quality and diversification of query suggestions. Fig.1 shows the ontology-based diversifying query suggestion model. We first generate suggestion candidates from search logs based on the forward random walk. A query-URL bipartite graph is constructed as $G = (V, E)$, in which vertexes are composed by two parts $V = V_1 \bigcup V_2$. V_1 represents the set of all unique queries and V_2 represents the set of all unique URLs. There exits an edge from node $x \in V_1$ to node $y \in V_2$ if y is a clicked URL of query x in the search logs. The weight $\omega(x, y)$ is the total number of times that y is clicked when query x is issued. Note that since the edge is an undirected edge, the weight $\omega(y, x)$ is the same as $\omega(x, y)$. The transition probabilities from node i to node j in this paper is defined as follows:

$$P_{t+1|t}(j \mid i) = \frac{\omega(i,j)}{\sum_k \omega(i,k)} \tag{1}$$

where k ranges over all nodes. $P_{t+1|t}(j \mid i)$ denotes the transition probability from node i at step t to node j at time step $t+1$.

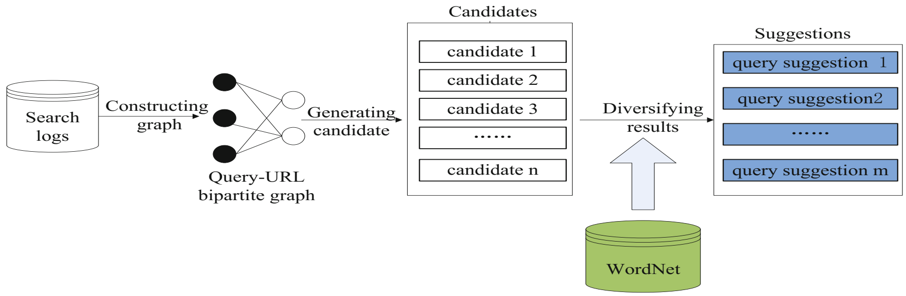

Fig. 1. Ontology-based diversifying query suggestion model

We represent the transitions as a sparse matrix A and perform the random walk using A. We calculate the probability of transition from node i to node j in t steps as $P_{t|0}(j \mid i) = [A^t]_{ij}$, right here A^t means t-step random walk transition matrix. It gives a measure of the volume of paths between these two nodes. If there are many paths the transition probability will be high. The larger the transition probability $P_{t|0}(j \mid i)$ is, the more the node j is similar to the node i. We select the top n largest $P_{t|0}(j \mid i)$ as candidates. In this study, we set the $n = 100$ empirically. In fact, $n = 100$ is large enough because we generally select 20 query suggestions at most.

To improve the quality of query suggestions and diversify query suggestions, we exploit semantic relationships between queries based on the WordNet ontology. In the study, we map the suggestion candidates based on the different senses of original queries. There are three steps for diversifying query suggestions based on WordNet as follows:

Step 1: For each query suggestion candidate c, we compute similarity scores between the different senses of an original query q and candidate c. In this way, we exploit the semantic relationships between query suggestion c and the orignal query q. We use the Lin similarity method [5], because it is a probabilistic model and is a universal similarity function, not tied to a particular application or a form of knowledge representation. For concept c_1 and c_2, their Lin similarity is defined as follows:

$$sim_L(c_1, c_2) = \frac{2 \times \lg p(lcs(c_1, c_2))}{\lg p(c_1) + \lg p(c_2)} \tag{2}$$

where $lcs(c_1, c_2)$ is the least common parent node of c_1 and c_2; $p(c)$ is probabilities of concept c and $p(c) = \frac{\sum_{w \in words(c)} count(w)}{N}$, where $words(c)$ is a set of words which concept c contains, $count(w)$ is the number of w in Brown corpus and N is the number of words in the Brown Corpus.

For two phrases consisted multiple terms, we calculate their similarity sim_P between the last noun of the two phrases as the final score. If the last noun of the two phrases are same, we calculate the similarity sim_L between the penultimate word of candidate and the last noun of the original query. The final score $sim_P = 0.5 + 0.5 \times sim_L$.

Step 2: For an original query having n senses in WordNet, we obtain a sense set $Q_s = \{sense[0], \ldots sense[n-1]\}$. After calculating the sim_L between c and $sense[i] \in Q_s$, we select the $sense[i]$ with highest sim_L and map the corresponding c to the $sense[i]$.

For example, Table 1 shows the similarity score of "apple" and "banana" in two different senses. In the table, the "n" represents "noun", the number "1" or "2" represents the different senses of the word. We find that the highest similarity is the sim_L score between "*apple#n#*1" and "*banana#n#*2", which means that the second sense of banana is most similarity to the first sense of apple. For an original query "apple", "banana" is a suggestion candidate, then we map "banana" to the first sense of "apple".

Table 1. The similarity results of "apple" and "banana" using Lin similarity

combinations of different senses	similarity score
$apple\#n\#1, banana\#n\#1$	0.11802556069890623
$apple\#n\#1, banana\#n\#2$	0.6867056880240358
$apple\#n\#2, banana\#n\#1$	0.0
$apple\#n\#2, banana\#n\#2$	0.0

Step 3: We rank suggestion candidates according to the similarity score for each $sense[i] \in Q_s$. Based on the ranking results, we select the queries which have the top k highest similarity score as suggestions. The value of k is determined on how many query suggestions we need. Since the candidates are selected from different groups based on $sense[i] \in Q_s$, our method ensures that the final query suggestions are semantically diversified and related to the Q_s.

3 Evaluation

3.1 Experimental Setup

In this section, we conduct empirical experiments to show the effectiveness of our proposed algorithm. We select AOL Search Data as our data set, which is a collection of real search log data based on real users. The data set consists of 20M web queries collected from 650k users over three months. We clean the data by keeping those frequent, well-formatted, English queries (queries which only contain characters 'a', 'b',. . . ,'z' and space, and appear more than 5 times). After cleaning, we obtain a total of 9,752,848 records, 604,982 unique queries and 785,012 unique URLs.

We construct a query-URL bipartite graph on our data set, and randomly sample a set of 50 queries from our data set as the testing queries. For each testing query, we obtain six query suggestions. In order to evaluate the quality of the results, three experts are requested to rate the query suggestion results with "0" or "1", where "0" means "irrelevant" and "1" means "relevant". We use the precision measurement in our experiment, i.e., precision at position n is defined as:

$$p@n = \frac{rn}{n} \tag{3}$$

where rn is the number of relevant queries in the first n results.

In order to evaluate the diversity of query suggestion results, we propose a diversity measurement method. Intuitively, if similarity score between two queries was high, then the diversity of the two queries are low. We select the first sense for each noun of q_i in WordNet, and then we use Bag of Words model to indicate q_i, denoting $\vec{qt_i} = (qt_{i1}, \ldots, qt_{in})$. We calculate the similarity (q_i, q_j) by Cosine similarity. Hence, we measure the diversity of two queries as follows:

$$D(q_i, q_j) = 1 - cos(\vec{qt_i}, \vec{qt_j}) = 1 - \frac{\sum_{k=1}^{n} qt_{ik} \times qt_{jk}}{\sqrt{\sum_{k=1}^{n} qt_{ik}^2} \times \sqrt{\sum_{k=1}^{n} qt_{jk}^2}} \tag{4}$$

We evaluate the diversity over a query set S_q as follows:

$$SD(S_q) = \frac{\sum_{i=1}^{K} \sum_{j=1, i \neq j}^{K} D(q_i, q_j)}{K \times (K-1)} \tag{5}$$

where K is the size of the query set S_q. We compare our method with Diversifying Query Suggestion (DQS) method [10] and Forward Random Walk (FRW) method [3]. We realize all the methods on a Core i5 computer, with CPU clock rate of 2.8 GHz, 4GB RAM, and running Windows 7.

3.2 Experimental Results

The comparison results are shown in Fig.2 and Fig.3. We found that our method is superior to FRW and DQS in precision as well as diversity. For example, when six query suggestions are returned, average precisions for the three methods are ODQS 0.59, DQS 0.54 and FRW 0.50. ODQS is 5% higher than DQS and 9% higher

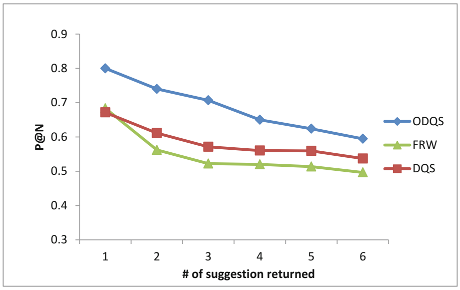

Fig. 2. The average precision

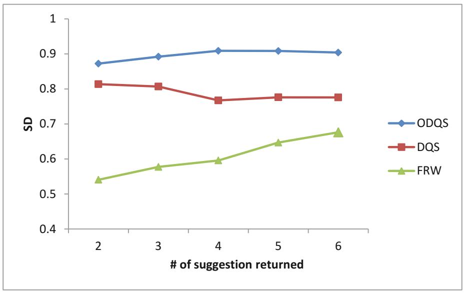

Fig. 3. The average of diversity

than FRW. The average diversities for the three methods are ODQS 0.9, DQS 0.78 and FRW 0.68. ODQS is 12% higher than DQS and 22% higher than FRW.

We list three cases for three queries in Table 2. Given a query, we discovered that our suggestion results cover much more latent topics intuitively. For example, given a query “birthday cakes”, our method suggested “kids birthday cakes” and “princess birthday cakes” which is the most relevant to the given query. It also suggests “birthday invitations” and “kids craft”, which is be closely related to birthday. “wedding cakes” as a kind of cake is also suggested, which means that the diversified suggestion are discovered by our method. DQS returns diversified but not so relevant query suggestions, such as “electric power washers”. The results of FRW are relevant, i.e., most suggestions are related to the “cakes”. However, most suggestions are redundant and represent the same topic “cakes”. The three cases in Table 2 show that suggestion results of our method are relevant as well as diversified.

Different from FRW and DQS, which only leverage search logs to generate suggestions, our method exploits semantic relationships between original query and suggestions based on WordNet. We re-ranking the suggestion candidates based on their relevance between different senses of the original queries. Therefore, the suggestions are more semantically related to the original queries. We also use WordNet to find latent topics - different senses of the original query, and map suggestion candidates to these latent topics. Our method is able to discover the query suggestions in different topics intrinsically. As a result, our method diversifies the query suggestions better than the FRW and DQS method.

Table 2. Query Suggestion Comparisons between ODQS and otherMethods

Query = travel		
ODQS	FRW	DQS
air fares	travelocity	travelocity
air travel	expedia	lonely planet
aarp passport	orbitz	travel texas
hotels	airline tickets	frommers california travel guide
Yahoo travel	airfare	frommers maui
travel channel	hotels	haunted travels
Query = world		
ODQS	FRW	DQS
world map	world map	world map
child labor	atlas	the world fact book
time zones	world atlas	cia united states
world bank	cnn	time in london
costa rica	child labor	world clock
education	time zones	time in spain
Query = birthday cakes		
ODQS	FRW	DQS
kids birthday cakes	birthday cakes	baby shower cakes
baby shower cakes	cakes	electric power washers
birthday invitations	baby shower cakes	wedding invitations
princess birthday cakes	princess birthday cakes	how to make icing roses
wedding cakes	wedding cakes	the cheese cake factory
kids crafts	birthday invitations	creative birthday gifts

In order to compare efficiency of the ODQS, FRW, and DQS, we implement three methods on the same computer and found the average times of three methods are DQS 721,469ms, FRW 220,277ms and ODQS 269,832ms. It is obvious that DQS consumes much more time than FRW and ODQS. It is chiefly because DQS algorithm has an dense matrix operations and iterative operation, which is time-consuming. We also found that our method has similar running time with FRW. It is because our method generate query suggestion candidates using the sparse transition matrix, and diversify query suggestions based on WordNet without using iterative operation. However, our method is able to generate diversified suggestions while FRW is not.

We analyze the time complexity of the three method here. To calculate a forward random walk, we encode the start distribution as a row vector ν_i with a unit entry at query node i, and obtain $P_{t|0}(j \mid i) = \nu_j [A]^t$. Then the time complexity of FRW is $O(n^2)$. The time complexity of DQS is $O(n^2) + (k - 1)O(n \times m)$, where k is the number of returned query suggestions and m is the number of neighbors of node i ($m << n$). The time complexity of our method is $O(n^2) + O(m)$ ($m << n$). To conclude, the experimental results show our method can suggest the diversified queries and improve the precision of suggestions in lower time consumption.

4 Related Work

Query suggestion techniques are used to improve search experience and help users to express these information needs. Most early query suggestion methods leverage document information or corpus [1, 2]. Recently, query suggestion using search logs has been widely studied.

Craswell et al. [3] proposed a backward random walk on query-URL bipartite. Mei et al. [4] introduced the concept of hitting time to iterative compute the hitting time. Beeferman et al. [6] used a hierarchical agglomerative on a query-URL bipartite to discover similar queries. Cao et al. [7] clustered similar queries into a concept and build the concept sequence suffix tree to realize query suggestion. Boldi et al. [8] proposed a concept of query-flow graph to obtain query suggestion results. There are some methods to improve the query-flow graph to suggest queries [9, 12–14]. In recent years, researchers presented the concept of diversifying query suggestions [10, 11]. However, most existing methods only use the search logs without considering the semantic relationships between queries, which may reduce the quality and the diversification of query suggestions.

5 Conclusion and Future Work

In this paper, we propose an ontology-based diversify query suggestion method. Query suggestion candidates are generated using Markov random walk. We rank the suggestion candidates and diversify the results based on an ontology - WordNet. Unlike existing query suggestions methods, we diversify query suggestions using not only search logs but also semantic relationships based on WordNet. The experimental results show that our method is able to suggest the diversified queries and outperform the FRW and DQS methods in terms of precision with lower time consumption.

In the future, we will focus on the latent topics mining and the optimization of the similarity algorithm. Since WordNet does not contain all words, especially spoken words and some names, we will try a comprehensive ontology to discover the latent topics related to the original query.

Acknowledgments. This research is supported by the 863 project of China (2013AA013300), National Natural Science Foundation of China (Grant No. 61375054) and Tsinghua University Initiative Scientific Research Program Grant No. 20131089256.

References

1. Meij, E., Bron, M., Hollink, L., Huurnink, B., de Rijke, M.: Learning Semantic Query Suggestions. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 424–440. Springer, Heidelberg (2009)
2. Gong, Z., Cheang, C.W., Hou U, L.: Web Query Expansion by WordNet. In: Andersen, K.V., Debenham, J., Wagner, R. (eds.) DEXA 2005. LNCS, vol. 3588, pp. 166–175. Springer, Heidelberg (2005)
3. Craswell, N., Szummer, M.: Random walks on the click graph. In: SIGIR 2007, pp. 239–246 (2007)
4. Mei, Q.Z., Zhou, D.Y., Church, K.: Query Suggestion Using Hitting Time. In: CIKM 2008, pp. 469–478 (2008)
5. Lin, D.: An Information-Theoretic Definition of Similarity. In: ICML 1998, pp. 296–304 (1988)
6. Beeferman, D., Berger, A.: Agglomerative clustering of a search engine query log. In: KDD 2000, pp. 407–416 (2000)
7. Cao, H.H., Jiang, D.X., Pei, J., Liao, Z., Chen, E., Li, H.: Context-Aware Query Suggestion by Mining Click-Through and Session Data. In: KDD 2008, pp. 875–884 (2008)
8. Boldi, P., Bonchi, F., Castillo, C., Donato, D., Gionis, A.: The Query-flow Graph: Model and Applications. In: CIKM 2008, pp. 56–63 (2008)
9. Anagnostopoulos, A., Becchetti, L., Castillo, C., Gionis, A.: An optimization framework for query recommendation. In: WSDM 2010, pp. 609–618 (2010)
10. Ma, H., Lyu, M.R., King, I.: Diversifying Query Suggestion Results. In: AAAI 2010, pp. 1399–1404 (2010)
11. Song, Y., Zhou, D., He, L.: Post-Ranking Query Suggestion by Diversifying Search Results. In: SIGIR 2011, pp. 815–824 (2011)
12. Baraglia, R., Nardini, F., Castillo, C., Perego, R., Donato, D., Silvestri, F.: The effects of time on query flow graph-based models for query suggestion. In: RIAO 2010, pp. 182–189 (2010)
13. Zhang, Z., Nasraoui, O.: Mining search engine query logs for query recommendation. In: WWW 2006, pp. 1039–1040 (2006)
14. Sadikov, E., Madhavan, J., Wang, L., Halevy, A.: Clustering Query Refinements by User Intent. In: WWW 2010, pp. 841–850 (2010)
15. Spink, A., Jansen, B.J.: A Study of Web Search Trends. Webology 1(2) (2004)
16. Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K.: WordNet: An online lexical database. Int. J. Lexicograph. 3(4), 235–244 (1990)