Query suggestions with global consistency on user click graph

Md. Zia Ullah[a] Masaki Aono[†, b]

Toyohashi University of Technology †Toyohashi University of Technology

a) *arif@kde.cs.tut.ac.jp* b) *aono@tut.jp*

概要 Users express their information needs in terms of queries in search engines to find some relevant documents on the Internet. However, users' search queries are usually short, ambiguous and/or underspecified. Sometimes, users have been found to struggle formulating queries based on keywords given their limited vocabulary. To help users in formulating query, query suggestion by mining query logs plays an important role and has been attracted attention in the recent years. A query log is generally represented as a bipartite graph on a query set and a URL set. Most traditional approaches used the raw click frequency to weigh the link between a query and a URL on the click graph. In order to alleviate the spurious effects of raw click frequency, some entropy-biased model by incorporating raw click frequency with the inverse query frequency or inverse URL frequency was proposed as the weighting scheme for query representation. In this paper, we observe that popular query and URLs are very diverse in nature, and user click frequency can be considered as local property of the URL, and link structures of query and URLs from both sides of the bipartite graph can be considered as a global property on the click graph. Based on this understanding, we develop a weighting scheme to weight the link between a query and an URL in the bipartite click graph by incorporating the user click frequency, and the link structures of the query and URL from both sides with global consistency in a consistent manner. We conduct experiments on the AOL search engine query log dataset and evaluate the query suggestions by estimating the similarity between the user query and suggested query using the knowledge of the *Dmoz* open directory project. The results turns out that our global consistency scheme achieved better performance than the current entropy-biased model.

キーワード Query logs, bipartite graph, global property, click frequency, *Dmoz*

1 Introduction

Extensive research has been conducted on query log analysis nowadays, since the exploitation of the clicks of past users from the query log has been proven to be an effective method to improve the search result. Since users have been found to struggle formulating queries based on keywords given their limited vocabulary, query suggestion by mining query logs plays an important role in this regard. A query log is generally represented as a bipartite graph on a query set and a URL set. An edge connects a query and a URL, and the edge value generally corresponds to the raw click frequency as the semantic relation.

The objective of the query suggestion is to find semantically similar queries for the given query by mining search engine query logs, in which we have a query-URL bipartite graph, and the queries and URLs. The problem we address is how to utilize and leverage both the graph and content information, so as to improve the precision of the retrieved queries.

Most traditional approaches used the raw click frequency to weigh the link between a query and a URL on the click graph which suffers from two problems: raw click frequency does not favor unpopular queries or URLs, and ranking model based on raw click frequency often favor long-tail query and URLs. In order to alleviate the spurious effects of raw click frequency, some entropy-biased model [7] inspired by TF-IDF in text retrieval by incorporating raw click frequency with the inverse query frequency or inverse URL frequency [13] was proposed as the weighting scheme for query representation. Since user frequency in click graph is the implicit feedback from user in compare to the term importance in the document, utilizing user frequency in the same manner as TF-IDF may not be appropriate in the context of user click graph. Our assump-

tion is that less clicked URLs tend to be more relevant to a given query, and the query's link structure and URL's link structure has great importance on the final semantic relevancy between the query and URL. For example, let consider two queries *map* and *weather*, where they may be co-linked by some URLs such as "www.google.com" and "weather.com". It is obvious that the query *weather* more appropriate to the specific URL "weather.com" than "www.google.com". As the general URL "www.google.com" is associated with many queries, it can aggregates large relevance score by applying link analysis method and eventually query *weather* may have large relevance score for this general URL. In this case, if we consider the content information of URL "www.google.com" for the query *weather* including the link information of query and URL, the query *weather* may have low relevance score with the general URL "www.google.com", and similarly "weather" may have large relevance score with the specific URL "weather.com".

In this paper, we observe that popular query and URLs has long-tail. The implicit user click frequency is the content information which can be considered as local property of the URL. The link structures of the query and URL can reduce adverse effect of the popular queries and URLs which can be considered as the global property of the click graph. Based on this understanding, we introduce a weighting scheme to weight the link between a query and an URL in the bipartite click graph by incorporating the user click frequency, and the link structures of the query and URL from both sides of the bipartite graph with global consistency in a consistent manner.

The rest of this paper is organized as follows. Section 2 describes the systematic review of the related work while our proposed method is defined in section 3. Section 4 includes the experimental evaluation and the results we obtained. Finally, concluding remarks and some future directions of our work are described in Section 5.

2 Related Work

Extensive research has been carried out on query recommendation based on click-through data from query logs [4, 10]. In these methods, the association between query and documents in the search graph is mined to infer related queries. More recently, random walk frameworks have been proposed to rank documents and queries using hitting time [11], and based on query-document frequency in the graph [5]. There were more works ranging from selecting queries to be suggested from those appearing frequently in query session [9], to use clustering to devise similar queries on the basis of cluster membership [2, 1, 3], to use click-through data to devise query similarity [14, 6]. An algorithm has been proposed based on link analysis to suggest the similar query by incorporating the bipartite graph with the content information from both sides as well as the constraints of relevance [8].

In contrast, less work has been carried out on the study of query representation on the click graphs. The entropy-biased model for query representation has been proposed [7] to replace raw click frequency on the click graph. It assumed that less clicked URLs are more effective in representing a given query than heavily clicked ones. Thus, the raw click frequency was weighed by the inverse query frequency of the URL. However, the entropy-biased model utilized raw click frequency and inverse query frequency in the same manner as TF-IDF does, which may not be appropriate in the context of click graph. Our work is closely related to [13], which introduced inverse URL frequency as global weight of URL while our contribution is to study how to combine the user frequency, the link structure of query and URL in a consistent way for query representation.

3 Methods

In this section, we describe the scenario in which query representation is studied and the method proposed to mine query suggestions on user click graph.

3.1 Preliminaries

Many web data can be modeled as bipartite graph. For instance, search engine query log data can be modeled as Query-URL bipartite graph. Let us consider a

bipartite graph: $\mathcal{G} = (\mathcal{Q} \cup \mathcal{U}, \mathcal{E})$, where the query set $\mathcal{Q}$ and the URL set $\mathcal{U}$ are connected by edges in $\mathcal{E}$. In graph $\mathcal{G}$, each edge in $\mathcal{E}$ connects a query vertex $q \in \mathcal{Q}$ and a URL vertex $u \in \mathcal{U}$ with an edge value $f_{q,u}$; that is, there is no edge between two vertices in the same set. The edge value $f_{q,u}$ denotes the semantic relations between query q with URL u. In most cases, the edge value $f_{q,u}$ corresponds to the raw click frequency $cf_{q,u}$ or user frequency $uf_{q,u}$ between a query q and a URL u. The raw click frequency $cf_{q,u}$ is the number of times the users click on URL u when u is presented to the users in search engine result page ($SERP$) for query q, and the user frequency $uf_{q,u}$ is the number of different users who issued query q and clicked on URL u.

Let $\mathcal{Q} = \{q_1, q_2, \cdots, q_{\mathcal{M}}\}$ and $\mathcal{U} = \{u_1, u_2, \cdots, u_{\mathcal{N}}\}$ be two sets of $\mathcal{M}$ and $\mathcal{N}$ unique entities. Now, the bipartite graph $\mathcal{G}$ can be represented as rank $\mathcal{M} \times \mathcal{N}$ matrix $\mathcal{W}$, with the entry (i, j) which contains the edge value f_{q_i,u_j} of query q_i with URL u_j. Thus, a query q_i can be represented as a row vector of $\mathcal{W}$, and a URL u_j corresponds to a column vector of $\mathcal{W}$.

Given $q_i \epsilon \mathcal{Q}$ and $u_j \epsilon \mathcal{U}$, if there is an edge connecting q_i and u_j, the transition probabilities w_{q_i,u_j} and w_{u_j,q_i} are positive, where w_{q_i,u_j} denotes the transition probability from a query q_i to a URL u_j, and w_{u_j,q_i} denotes the transition probability from a URL u_j to a query q_i; otherwise, $w_{q_i,u_j} = w_{u_j,q_i} = 0$. Since the transition probability from state i to all other states must be 1, we have $\sum_{u_j \epsilon \mathcal{U}} w_{q_i,u_j} = 1$ and $\sum_{q_i \epsilon \mathcal{Q}} w_{u_j,q_i} = 1$.

For a bipartite graph, there is a natural random walk on the graph with the transition probability as discussed above. Let $\mathcal{W}^{\mathcal{QU}} \epsilon\ \mathcal{R}^{\mathcal{M}\times\mathcal{N}}$ denote the transition matrix from query set $\mathcal{Q}$ to URL set $\mathcal{U}$, whose entry (i, j) contains a value w_{q_i,u_j} from a query q_i to a URL u_j. Let $\mathcal{W}^{\mathcal{UQ}} \epsilon\ \mathcal{R}^{\mathcal{N}\times M}$ be the transition matrix from URL set $\mathcal{U}$ to query set $\mathcal{Q}$, whose entry (j, i) contains a value w_{u_j,q_i} from a URL u_j to a query q_i.

3.2 Entropy-Biased Model

The entropy-biased model from [7] proposed *inverse query frequency* for the URL, which are similar to TF-IDF term weighting scheme in text retrieval. It has discriminative ability to lower the weight of general URLs. Therefore, it incorporates inverse query frequency with user click frequency to weigh the edge between a query and a URL as follows:

$$weight(q_i, u_j) = uf_{q_i,u_j} \times \log(\frac{|\mathcal{Q}|}{l_{u_j}}) \qquad (1)$$

表 1 Preliminaries and Notations

$\lvert\mathcal{Q}\rvert$	Total number of queries
$\lvert\mathcal{U}\rvert$	Total number of URLs
uf_{q_i,u_j}	Total number of users clicked on URL u_j for query q_i
w_{q_i,u_j}	Transition probability from query q_i to URL u_j
w_{u_j,q_i}	Transition probability from URL u_j to query q_i
l_{u_j}	Number of district queries associated with URL u_j
l_{q_i}	Number of district URLs associated with query q_i
$Avgl_{\mathcal{Q}}$	Average number of district URLs associated all queries
$Avgl_{\mathcal{U}}$	Average number of district Queries associated with all URLs

3.3 Global Weight Model

Traditionally, a query can be represented by the edge values of the associated URLs, such as click frequency $cf_{i,j}$ or user frequency $uf_{i,j}$. Previous works argued that different query-URL pairs should be treated differently [7]. As some queries and URLs in the click graph are very diverse, we can consider two aspects for weighting the edge between a query and a URL in the click graph. The first aspect is the propagation of weight from the query to the URL, for example, in Figure 1 the edge 'map - www.google.com' and the edge 'map - www.wikipedia.com' have the same user frequency ($uf_{q_4,u_1} = uf_{q_4,u_3} = 5$), moreover, 'map - www.google.com' should be more relevant to the query 'map' than 'map - www.wikipedia.com'. To overcome this problem, inverse query frequency (IQF) or inverse url frequency (IUF) has been proposed [7, 13] to improve the weighting scheme in a consistent manner. The second aspect is the propagation of weight from the URL to the query, for example, in Figure 1 'weather - weather.com' and 'news - weather.com' have the same

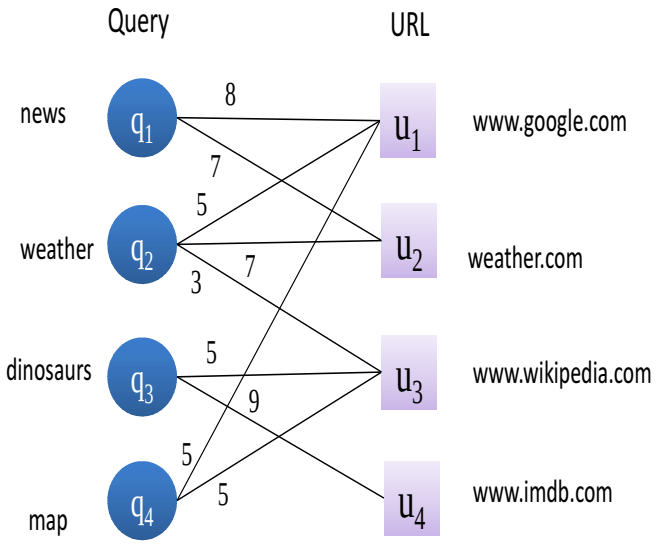

図 1 A query-URL bipartite graph example

user frequency ($uf_{q_2,u_2} = uf_{q_1,u_2} = 7$), moreover, 'weather - weather.com' is more relevant to the URL 'weather.com' than the query 'news'. However, the general queries or URLs should have low relevance score than the unique or more specific queries or URLs. In order to avoid the adverse effect of noisy edges among popular queries and URLs, we consider the user frequency, the structure of query, and the structure of the URL in the click graph to weigh an edge between a query and a URL.

To estimate the weight between a query and a URL, we incorporate the relevancy of the query and the URL from both sides of the bipartite graph by including content information and link structures. The weight between a query q_i and a URL u_j is defined using our proposed globally consistent weight scheme as follows:

$$weight(q_i, u_j) = \frac{Avgl_{\mathcal{Q}}}{l_{u_j}} \cdot w_{q_i,u_j} + \frac{Avgl_{\mathcal{U}}}{l_{q_i}} \cdot w_{u_j,q_i} \quad (2)$$

where l_{u_j} is the length of the URL u_j i.e. the number of queries associated with it, $Avgl_{\mathcal{U}}$ is the average length of all the URLs in $\mathcal{U}$, l_{q_i} is the length of the query q_i i.e. the number of URLs associated with it, $Avgl_{\mathcal{Q}}$ is the average length of all the queries in $\mathcal{Q}$. The w_{q_i,u_j} is the $(i,j)^{th}$ entry of the transition matrix $\mathcal{W}^{\mathcal{QU}}$ and w_{u_j,q_i} is the $(j,i)^{th}$ entry of the transition matrix $\mathcal{W}^{\mathcal{UQ}}$.

3.4 Query Similarity Matrix

To consider the vertices in one side of the bipartite graph, such as query-to-query graph in query logs, then a hidden similarity weight $w^{\mathcal{QQ}}_{q_i,q_j}$ from query q_i to query q_j can be introduced as:

$$w^{\mathcal{QQ}}_{q_i,q_j} = \sum_{k \in \mathcal{U}} w^{\mathcal{QU}}_{q_i,u_k} w^{\mathcal{UQ}}_{u_k,q_j} \quad (3)$$

In equation 5, the similarity score between two query q_i and q_j is calculated through their common clicked URLs. Therefore, we computed a query-to-query similarity matrix $\mathcal{W}^{\mathcal{QQ}}$ by exploiting the relations from Query-URL bipartite graphs applying One Mode Projection. In this similarity matrix, we only found the non-sparse query and their similar queries. The dimension of this similarity matrix $\mathcal{W}^{\mathcal{QQ}}$ is much lower the original matrix $\mathcal{W}^{\mathcal{QU}}$. We use this matrix for suggesting similar queries for a user given query.

4 Experimental Evaluation

In the following experiments, we conduct an empirical evaluation on the tasks of mining query logs. The task is defined as follows: Given a query and a query-URL bipartite graph, the system has to identify a list of queries which are most similar or semantically relevant to the given query. In the rest of this section, we introduce the data collection, the assessments and evaluation metrics, and present the experimental results.

4.1 Data Collection

The experimental dataset is selected from the AOL search engine query log [12]. The log data comprises twenty million search queries from 650,000 users over three months. The entire collection consists of 19,442,629 user click-through events. These records contain 10,154,172 unique queries and 1,632,789 unique URLs. As shown in table 2, each record of the click contains the same information in the form of UserID, Query, Time, Rank, and ClickURL. This data set is the raw data recorded by the search engine, and contains a lot of noises. Hence, we clean the data by removing queries which length is less than three. If the same query is issued by the same user, we only consider it as a single click-through event. After cleaning, the data collection con-

表 2 Samples of the AOL query log dataset

UserID	Query	Time	Rank	ClickURL
217	lottery	2006-03-01 11:58:51	1	www.calottery.com
2005	google	2006-03-24 21:25:10	1	www.google.com
2178	baby names	2006-05-20 15:02:38	1	www.babynames.com
2421	poker	2006-03-14 4:58:14	4	www.pokerstars.com

sists of $4,811,111$ queries and $1,091,637$ URLs. After the construction of the query-URL bipartite graph, we further filter out the queries which is issued less than 5 users. After filtering the click graph, we observe that a total of $5,81,612$ queries and $3,35,441$ URLs exist, and each queries has 5.135 district clicks, and each URL is clicked by 4.486 district queries.

4.2 Assessment and Evaluation Metric

It is difficult to evaluate the quality of query similarity/relevance rankings due to the scarcity of data that can be examined publicly. We employ the same method as used in [4] for automatic evaluation, using *Dmoz* Open Directory Project (*ODP*)[1] to represent each query. As the reader may know, when a user types a query in *ODP*, besides site matches, we can also find categories matches in the form of directory paths. Moreover, these categories are ordered by relevance. For instance, the query "Bangladesh" would provide the category "Regional/ Asia/ Bangladesh/ Business and Economy/ Shopping" (among others), while one of the results for "Chittagong" would be "Regional/ Asia/ Bangladesh/ Localities/ Chittagong/ Business and Economy". Hence, to measure how similar two queries are, we can use a notion of similarity between the corresponding categories (directory paths) as provided by the *ODP*. In particular, the similarity between two directory paths d_i and d_j is defined as the length of the longest common prefix P(d_i, d_j) divided by the length of the longest path between d_i and d_j. More precisely, the similarity between two directory paths d_i and d_j is defined as:

$$Sim(d_i, d_j) = \frac{|P(d_i, d_j)|}{max(|d_i|, |d_j|)} \tag{4}$$

where $|d_i|$ denotes the length of a path. For instance, the similarity between the above two queries is 3/6 since they share the path "Regional/ Asia/ Bangladesh" and the longest one is made of six directories, while the similarity between "Bangladesh" and "Dhaka" of which directory path is "Regional/ Asia/ Bangladesh/ Government/ Embassies and Consulates/ Foreign", is 3/6. We evaluate the similarity between the aggregated categories of the two queries, among the top 5 answers provided by *ODP*.

To give a fair assessment, we randomly select 200 distinct queries from the data collection, then retrieve the top 5 similar queries for each query from the query-to-query similarity matrix $\mathcal{W}^{\mathcal{QQ}}$. We further retrieve the top 5 *ODP* categories for each query terms. The similarity of the two groups of directory paths is chosen to the most similar categories between the two queries. For evaluation of the task, we adopt the precision at rank n to measure the relevance of the top n results of the retrieved list with respect to a given query q_m, which is defined as follows:

$$P@n = \frac{\sum_{i=1}^{n} Sim(q_m, q_i)}{n} \tag{5}$$

where $Sim(q_m, q_i)$ means the similarity between q_m and q_i. In our experiments, we report the precision from $P@1$ to $P@5$, and take the average precision over all the 200 district queries.

4.3 Performance Analysis

We have tested the performance of our method and report the average precision from P@1 to P@5. Our evaluation result is depicted in Figure 2. We could not explicitly compare our method with entropy-biased approach [13] because they uses the google directory paths instead of *ODP* directory paths, and another reason is that we applied very strong filtering strategy on the query-URL bipartite graph to remove the spareness of the graph. It turns out that our method

[1] http://www.dmoz.org

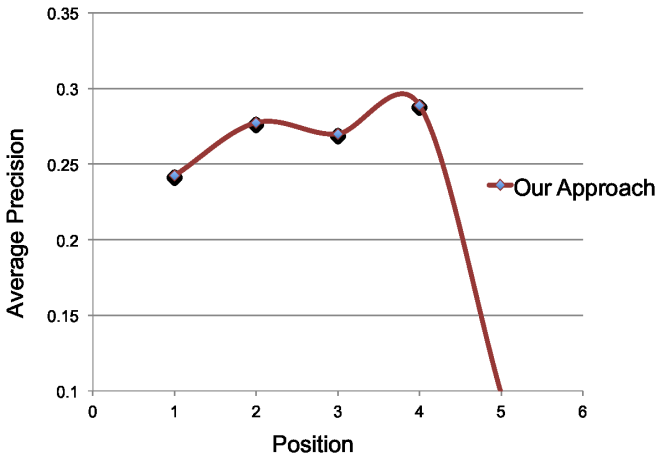

図 2 Average Precision at different positions

produces a good result in terms of precision in the filtered graph. Since, every day huge query logs data are recorded in the search engine, our result might perform much better with the most updated logs.

5 Conclusion

In this paper, we proposed a method for query suggestions by exploiting the knowledge from the query logs on user click graph. We incorporated the content information and link information of the query-URL bipartite graph in our method. The content information is the user click frequency as the local property and link information are the link structure of the query and URL as the global property. The query suggestions are automatically evaluated by estimating the precision at different levels using the category paths from *ODP*. In future, we would like to apply some score propagation method in the click graph and boost the performance of our system. We will further try to infer the query intent by analyzing the user session boundary and applying query clustering based approach.

参考文献

[1] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query clustering for boosting web page ranking. In *Advances in Web Intelligence*, pages 164–175. Springer, 2004.

[2] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query recommendation using query logs in search engines. In *Current Trends in Database Technology-EDBT 2004 Workshops*, pages 588–596. Springer, 2005.

[3] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Improving search engines by query clustering. *Journal of the American Society for Information Science and Technology*, 58(12):1793–1804, 2007.

[4] R. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 76–85. ACM, 2007.

[5] N. Craswell and M. Szummer. Random walks on the click graph. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 239–246. ACM, 2007.

[6] S. Cucerzan and R. W. White. Query suggestion based on user landing pages. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 875–876. ACM, 2007.

[7] H. Deng, I. King, and M. R. Lyu. Entropy-biased models for query representation on the click graph. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 339–346. ACM, 2009.

[8] H. Deng, M. R. Lyu, and I. King. A generalized co-hits algorithm and its application to bipartite graphs. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–248. ACM, 2009.

[9] B. M. Fonseca, P. B. Golgher, E. S. de Moura, and N. Ziviani. Using association rules to discover search engines related queries. In *Web Congress, 2003. Proceedings. First Latin American*, pages 66–71. IEEE, 2003.

[10] W. Gao, C. Niu, J.-Y. Nie, M. Zhou, K.-F. Wong, and H.-W. Hon. Exploiting query logs for cross-lingual query suggestions. *ACM Transactions on Information Systems (TOIS)*, 28(2):6, 2010.

[11] Q. Mei, D. Zhou, and K. Church. Query suggestion using hitting time. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 469–478. ACM, 2008.

[12] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *InfoScale*, volume 6, pages 1–7. Citeseer, 2006.

[13] D. Zhang, R. Zhu, S. Men, and V. Raychoudhury. Query representation with global consistency on user click graph. *arXiv preprint arXiv:1305.5981*, 2013.

[14] Z. Zhang and O. Nasraoui. Mining search engine query logs for query recommendation. In *Proceedings of the 15th international conference on World Wide Web*, pages 1039–1040. ACM, 2006.