Query Recommendation Using Large-scale Web Access Logs and Web Page Archive

Lin Li [1], Shingo Otsuka [2], and Masaru Kitsuregawa [1]

[1] Dept. of Info. and Comm. Engineering, The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
{lilin, kitsure}@tkl.iis.u-tokyo.ac.jp
[2] National Institute for Materials Science
1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan
otsuka.shingo@nims.go.jp

Abstract. Query recommendation suggests related queries for search engine users when they are not satisfied with the results of an initial input query, thus assisting users in improving search quality. Conventional approaches to query recommendation have been focused on expanding a query by terms extracted from various information sources such as a thesaurus like WordNet[1], the top ranked documents and so on. In this paper, we argue that past queries stored in query logs can be a source of additional evidence to help future users. We present a query recommendation system based on large-scale Web access logs and Web page archive, and evaluate three query recommendation strategies based on different feature spaces (i.e., noun, URL, and Web community). The experimental results show that query logs are an effective source for query recommendation, and the Web community-based and noun-based strategies can extract more related search queries than the URL-based one.

1 Introduction

Keyword based queries supported by Web search engines help users conveniently find Web pages that match their information needs. A main problem, however, occurs for users: properly specifying their information needs through keyword-based queries. One reason is that queries submitted by users are usually very short [8]. The very small overlap of the query terms and the document terms in the desired documents will retrieve Web pages which are not what users are searching for. The other reason is that users might fail to choose terms at the appropriate level of representation for their information needs. Ambiguity of short queries and the limitation of user's representation give rise to the problem of phrasing satisfactory queries.

The utilization of query recommendation has been investigated to help users formulate satisfactory queries [1, 3, 4, 8–10]. For example, new terms can be expanded to the existing list of terms in a query and users can also change some or all the terms in a query. We summarize the main process of query recommendation in the following three steps:

[1] http://wordnet.princeton.edu/

1. choosing information sources where most relevant terms as recommendation candidates can be found, given a current query;
2. designing a measure to rank candidate terms in terms of relatedness;
3. utilizing the top ranked relevant terms to reformulate the current query.

The selection of information sources in the first step plays an important role for an effective query recommendation strategy. The general idea of existing query recommendation strategies [3, 4, 9, 10] has focused on finding relevant terms from *documents* to existing terms in a query based on the hypothesis that a frequent term from the documents will tend to co-occur with all query terms. On the Web, this hypothesis is reasonable, but not always true since there exists a large gap between the Web document and query spaces, as indicated by [4] which have utilized query logs to bridge the query and Web document spaces. In this paper, different from the above researches, we think that *past queries* stored in the query logs may be a source of additional evidence to help future users. Some users who are not very familiar with a certain domain, can gradually refine their queries from related queries that have been searched by previous users, and hence get the Web pages they want.

In this paper, we present and evaluate a query recommendation system using past queries that provide a pool of relevant terms. To calculate the relatedness between two queries, we augment short Web queries by three feature spaces, i.e., noun space, URL space, and community (*community* means *Web community* in this paper) space respectively. The three feature spaces are based on Web access logs (the collected 10GB URL histories of Japanese users selected without static deviation) and a 4.5 million Japanese Web page archive. We propose a query recommendation strategy using a community feature space which is different from the noun and URL feature spaces commonly used in the literature [1, 3, 4, 8, 9]. The evaluation of query recommendation is labor intensive and it is not easy to construct an object test data set for it at current stage. An evaluation is carefully designed to make it clear that to which degree different feature based strategies add value to the query recommendation quality. We study this problem and provide some preliminary conclusions. The experimental results show that query logs are an effective source for query recommendation, and community-based and noun-based methods can extract more related search keywords than the URL-based one.

The rest of this paper is organized as follows. Firstly, we introduce three query recommendation strategies in Section 2. Then, we describe the details of experiment methodology and discuss the experimental results in Section 3 and Section 4 respectively. Lastly, we conclude our work in Section 5.

2 Query Recommendation Strategies

The goal our recommendation system is to find the related past queries to a current query input by a Web user, which means we need to measure the relatedness between queries and then recommend the top ranked queries. Previous queries having common terms with the input query are naturally recommended.

However, it is possible that queries can be phrased differently with different terms but for the same information needs while they can be identical but for the different information needs. To more accurately measure relatedness between two queries, most of existing strategies augment a query by terms from Web pages or search result URLs [1, 5, 8]. We think that the information related to the accessed Web pages by Web users are useful sources to augment original queries because the preferences of a user are reflected in form of her accesses. One important assumption behind this idea is that the accessed Web pages are *relevant* to the query. At the first glance, although the access information is not as accurate as explicit relevance judgment in the traditional relevance feedback, the user's choice does suggest a certain level of relevance. It is therefore reasonable to regard the accessed Web pages as relevant examples from a statistical viewpoint.

2.1 Three Feature Spaces for Query Recommendation

Given these accessed Web pages, we define three feature spaces as *noun space*, *URL space*, and *community space* to augment their corresponding Web queries and then estimate the relatedness between two augmented queries.

Noun Space As we know, nouns in a document can more accurately represent the topic described by the document than others. Therefore, we enrich a query with the nouns extracted from the contents of its accessed Web page sets, which intends to find the topics hidden in the short query. Our noun space is created using ChaSen, a Japanese morphological analyzer [2]. Since we have already crawled a 4.5 million Japanese Web page archive, we can complete the morphological analysis of all the Web pages in advance.

URL Space The query recommendation strategy using the noun feature space is not applicable, at least in principle, in settings including: non-text pages like multimedia (image) files, documents in non-HTML file formats such as PDF and DOC documents, pages with limited access like sites that require registration and so on. In these cases, URLs of Web pages are an alternate source. Furthermore, because the URL space is insensitive to content, for online applications, this method is easier and faster to get recommendation lists of related past queries than the noun feature space. Our URL space consists of the hostnames of accessed URLs in our Web logs. The reason that we use hostnames of URLs will be explained in Sectioin 3.2.

Community Space The noun and URL spaces are straightforward and common sources to enhance a query. In this paper, we utilize the Web community information as another useful feature space since each URL can be clustered into its respective community. The technical detail of creating community is in our previous work [7] which created a web community chart based on the complete bipartite graphs, and extracted communities automatically from a large amount of web pages. We labeled each Web page by a community ID. As thus, these community IDs constitute our community space.

[2] http://chasen-legacy.sourceforge.jp/

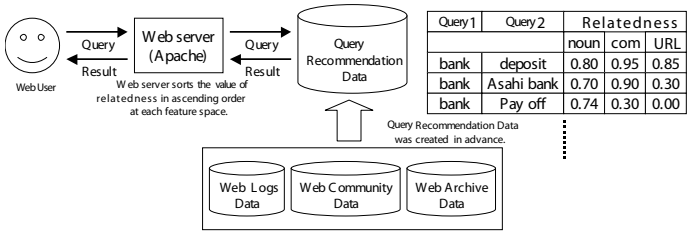


Fig. 1. The architecture of our system

2.2 Relatedness Definition

In this section we discuss how to calculate relatedness between two queries enriched by the three feature spaces. The relatedness is defined as

$$R_{q_x,q_y} = \frac{\sum_{e_i \in q_x \& e_i \in q_y} f_{q_x}(e_i) + \sum_{e_i \in q_x \& e_i \in q_y} f_{q_y}(e_i)}{2},$$

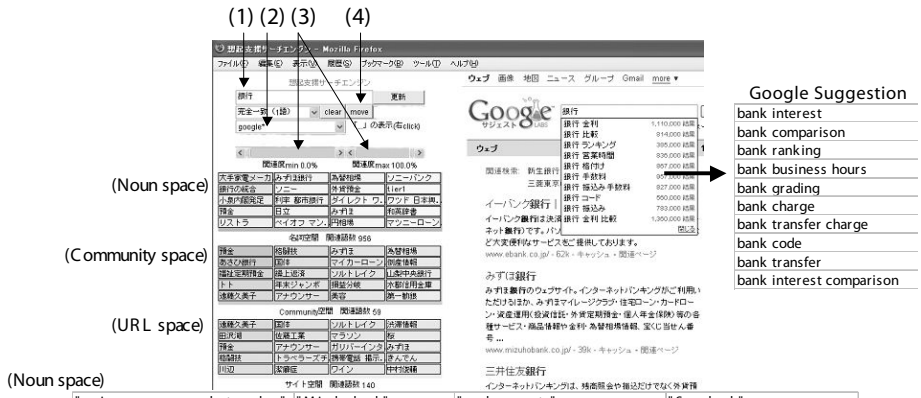
where q_x and q_y are queries, and Rq_x, q_y is the estimated relatedness score between q_x and q_y. Let $Q = \{q_1, q_2, \ldots, q_x, \ldots, q_n\}$ be a universal set of all search queries where n is the number of total search queries. We augment the query q_x by adding one of the three feature spaces(i.e., noun, URL, and community ID), denoted as: $q_x = \{e_1, e_2, \ldots, e_x, \ldots, e_m\}$ where e_x is an element of the used feature space and m is the total number of elements. The frequencies of elements in a query are denoted as $f_{q_x} = \{f_{e_1}, f_{e_2}, \ldots, f_{e_x}, \ldots, f_{e_m}\}$ for a single feature space. For URL space, we can get the frequencies of URLs visited by different users using access information in our Web logs.

In addition, we create an excluded set which stores the *highly frequent elements* of URLs, communities and nouns contained in accessed Web page sets. For example, the highly frequent elements of URLs are *Yahoo!, MSN, Google* and so on, and the highly frequent elements of nouns are *I, today, news* and so on. We exclude a highly frequent element e_h from the frequency space of f_{q_x} if the number of the test queries which feature spaces include e_h are more than half of the number of all the test queries used in our evaluation.

3 Experiment Methodology

3.1 Our System Overview

The goal of our query recommendation system is to find the related queries from past queries given an input query. Figure 1 shows a sketch of the system architecture consisting of two components. One is "Web Server(Apache)" where a user interactively communicates with our system. The user inputs a query to the Web server, and then the Web server returns the list of related past queries to her.



(Noun space)

" major consumer product maker"	" Mizuho bank"	" exchange rate"	" Sony bank"
" unification of the banks"	" Sony"	" foreign currency deposits"	(name of financial commodity)
" Koizumi Cabinet inauguration"	" rate" " megabank"	(name of financial commodity)	(name of financial commodity)
" deposit"	" Hitachi"	" Mizuho" (bank name)	" Japanese-English dictionary"
" restructure"	" pay off"	" Yen rate"	" my car loan"

(Community space)

" deposit"	" combative sport"	" Mizuho" (bank name)	" exchange rate"
" Asahi bank"	" national athletic meet"	" my car loan"	" bankruptcy information"
" welfare fixed deposit"	" advanced payment"	" Saltlake" (city name)	" Yamanashi Chuo bank"
" toto" (football lot)	(a name of public lottery)	" break-even point"	" Suito shinkin bank"
" Kumiko Endo"	" announcer"	" beauty"	" Dai-Ichi Kangyo bank"

(URL space)

" Kumiko Endo"	" national athletic meet"	" Saltlake" (city name)	" traffic jam information"
" Tazawa lake"	(company name)	" marathon"	" cherry tree"
" deposit"	" announcer"	(company name)	" Mizuho" (bank name)
" combative sport"	" traveler's check"	" cellular phone" " bulletin board"	(company name)
" Kawabe"	" cleanness freak"	" wine"	" Syunsuke Nakamura"

Fig. 2. The user interface of our system

The other is "Query Recommendation Data" storing the recommendation results of the three strategies discussed in Section 2. Our Web logs, Web community, and Web page archive data ensure the richness of information sources to augment Web queries. The interface of our query recommendation system is illustrated in Figure 2. A user can input a search query in Figure 2(1) while the related queries recommended by the component "Query Recommendation Data" are shown below and divided by using different feature spaces. Then, the user can choose one recommended query to add or replace the initial query and submit the reformulated query to a search engine selected from a dropdown list as shown in Figure 2(2). Finally, the search results retrieved by the selected search engine are listed in the right part. Furthermore, in Figure 2(3), there are two slide bars which can adjust the lower and upper bounds of relatedness scores. For each feature space, the maximal number of recommended queries is 20. If the user wants more hints, she can click a button shown in Figure 2(4) to get more recommended queries ordered by their relatedness scores with the initial query. The more the recommended query is related to the initial query, the query is displayed as a deeper red. Figure 2 presents recommendation of the query "Bank" as an example. Since in this study we utilize Japanese Web data, the corresponding English translation is in the bottom of this figure. If some

UserID	AccessTime	RefSec	URL
1	2002/9/30 00:00:00	4	http://www.tkl.iis.u-tokyo.ac.jp/welcome_j.html
2	2002/9/30 00:00:00	6	http://www.jma.go.jp/JMA_HP/jma/index.html
3	2002/9/30 00:00:00	8	http://www.kantei.go.jp/
4	2002/9/30 00:00:00	15	http://www.google.co.jp/
1	2002/9/30 00:00:04	6	http://www.tkl.iis.u-tokyo.ac.jp/Kilab/Welcome.html
5	2002/9/30 00:00:04	3	http://www.yahoo.co.jp/
6	2002/9/30 00:00:05	54	http://weather.crc.co.jp/
2	2002/9/30 00:00:06	11	http://www.data.kishou.go.jp/maiji/
3	2002/9/30 00:00:08	34	http://www.kantei.go.jp/new/kousikiyotei.html
5	2002/9/30 00:00:07	10	http://search.yahoo.co.jp/bin/search?p=%C5%B7%B5%A4
5	2002/9/30 00:00:10	300	http://www.tkl.iis.u-tokyo.ac.jp/Kilab/Members/members-j.html

(a)

Fig. 3. A part of our panel logs (Web access logs)

queries are only available in Japanese, we give a brief English explanation. For example, the query "Mizuho" is a famous Japanese bank.

3.2 Data Sets

Web Access Logs Our Web access logs, also called "*panel logs*" are provided by *Video Research Interactive Inc.* which is one of Internet rating companies. The collecting method and statistics of this panel logs are described in [6]. Here we give a brief description. The panel logs consist of *user ID, access time of Web page, reference seconds of Web page, URL of accessed Web page* and so on. The data size is 10GB and the number of users is about 10 thousand. Figure 3 shows the details of a part of our panel logs. In this study, we need to extract past queries and related information from the whole panel logs. We notice that the URL from a search engine (e.g., Yahoo!) records the query submitted by a user, as shown in Figure 3(a). We extract the query from the URL, and then the access logs followed this URL in a session are corresponding Web pages browsed by the user. The maximum interval to determine the session boundary is 30 minutes, a well-known threshold [2] such that two continuous accesses within 30 minutes interval are regarded as in a same session.

Japanese Web Page Archive The large-scale snapshot of a Japanese Web page archive we used was built in February 2002. We crawled 4.5 million Web pages during the panel logs collection period and automatically created 17 hundred thousand communities from one million selected pages [7]. Since the time of crawling the Web pages for the Web communities is during the time of panel logs collection, there are some Web pages which are not covered by the crawling due to the change and deletion of pages accessed by the panels. We did a preliminary analysis that the full path URL overlap between the Web access logs and Web page archive is only 18.8%. Therefore, we chopped URLs to their hostnames and then the overlap increases to 65%.

3.3 Evaluation Method

We invited nine volunteers(users) to evaluate the query recommendation strategies using our system. They are our lab members who usually use search engines

Table 1. Search queries for evaluation

Test Query	Accessed Web Pages	Group	Test Query	Accessed Web Pages	Group
lottery	891	A	bank	113	C
ring tone	446	B	fishing	64	A
movie	226	C	scholarship	56	B
hot spring	211	A	university	50	C
soccer	202	B			

Table 2. Evaluation results of our query recommendation system

Relevance of queries	Noun space	Community space	URL space
irrelevant	**0.037**	0.244	0.341
lowly relevant	0.043	0.089	0.107
relevant	0.135	0.131	0.106
highly relevant	**0.707**	0.480	0.339
relevant and highly relevant	**0.843**	0.611	0.444
un-judged	0.078	0.056	0.107

to meet their information needs. We also compare these strategies with the query recommendation service supplied by Google search engine. Nine test queries used in our evaluation are listed in Table 1. The total number of queries evaluated by users are 540 because they evaluate the top 20 of recommended queries according to their relatedness values on each of the three feature spaces [3]. To alleviate the workload on an individual user, we divided the nine users to three group (i.e., A, B, and C) as shown in Table 1. We ask each group to give their relevance judgments on three queries. The relevance judgment has five levels, i.e., irrelevant, lowly relevant, relevant, highly relevant, and un-judged.

4 Evaluation Results and Discussions

4.1 Comparisons of Query Recommendation Strategies

The evaluation results of the recommended queries related with all search queries are shown in Table 2 where each value denotes the percentage of a relevance judgment level on each feature space. For example, The value with the "highly relevant" judgment using noun space is 0.707 which means the percentage of the "highly relevant" judgment chosen by the users in all recommended queries[4] using the noun based strategy.

In Table 2, when the noun space based strategy is applied, there are more recommended queries evaluated "highly relevant" and fewer queries evaluated "irrelevant" than the other two feature spaces. Furthermore, if we combine the "highly relevant" judgments and the "relevant" judgments, we still gain the best results in the noun space while the percentage of the recommended queries judged as "irrelevant" using the URL space (i.e., 0.341) is higher than those

[3] 9 search queries * 20 recommended queries * 3 feature spaces = 540 evaluated queries
[4] 9 search queries * 20 recommended queries = 180 evaluated queries

using other two spaces. In the community space, there are many recommended queries evaluated "highly relevant" and "relevant". Although the community based strategy produces less related queries than the noun based strategy, it is more than the URL based strategy. By using the URL space, the number of recommended queries judged as "irrelevant" is more than that of queries judged as "highly relevant" (e.g., "irrelevant" vs. "highly relevant" = 0.341 vs. 0.339 in Table 2). In general, the noun and community spaces can supply us with satisfactory recommendation while the URL space based strategy cannot stably produce satisfactory queries related to a query.

4.2 Case Study with "Google Suggestion"

We compare the result of our case study in Figure 2 with the query recommendation service provided by *Google Suggestion*. In Figure 2 our system presents some good recommended queries related to *bank* in the noun and community spaces such as "deposit", "my car loan", "toto" and so on that are not given by *Google Suggestion*.

5 Conclusions and Future Work

In this paper, we design a query recommendation system based on three feature spaces, i.e., noun space, URL space, and community space, by using large-scale Web access logs and Japanese Web page archive. The experimental results show that the community-based and noun-based strategies can extract more related search queries than the URL-based strategy. We are designing an optimization algorithm for incremental updates of our recommendation data.

References

1. D. Beeferman and A. L. Berger. Agglomerative clustering of a search engine query log. In KDD, pages 407–416, 2000.
2. L. Catledge and J. Pitkow. Characterizing browsing behaviors on the world-wide web. *Computer Networks and ISDN Systems*, (27(6)), 1995.
3. P.-A. Chirita, C. S. Firan, and W. Nejdl. Personalized query expansion for the web. In SIGIR, pages 7–14, 2007.
4. H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Query expansion by mining user logs. *IEEE Trans. Knowl. Data Eng.*, 15(4):829–839, 2003.
5. N. S. Glance. Community search assistant. In IUI, pages 91–96, 2001.
6. S. Otsuka, M. Toyoda, J. Hirai, and M. Kitsuregawa. Extracting user behavior by web communities technology on global web logs. In DEXA, pages 957–968, 2004.
7. M. Toyoda and M. Kitsuregawa. Creating a web community chart for navigating related communities. In HT, pages 103–112, 2001.
8. J.-R. Wen, J.-Y. Nie, and H. Zhang. Query clustering using user logs. *ACM Trans. Inf. Syst.*, 20(1):59–81, 2002.
9. J. Xu and W. B. Croft. Query expansion using local and global document analysis. In SIGIR, pages 4–11, 1996.
10. Y. Zhu and L. Gruenwald. Query expansion using web access log files. In DEXA, pages 686–695, 2005.