Clustering of Search Engine Keywords using Access Logs

Shingo Otsuka† and Masaru Kitsuregawa†

†Institute of Industrial Science, The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
{otsuka,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract. It the becomes possible that users can get kinds of information by just inputting search keyword(s) representing the topic which users are interested in. But it is not always true that users can hit upon search keyword(s) properly. In this paper, by using Web access logs (called panel logs), which are collected URL histories of Japanese users (called panels) selected without static deviation similar to the survey on TV audience rating, we study the methods of clustering search keywords. Different from the existing systems where the related search keywords are extracted based on the set of URLs viewed by the users after input of their original search keyword(s), we propose two novel methods of clustering the search words. One is based on the Web communities (set of similar web pages); the other is based on the set of nouns obtained by morphological analysis of Web pages. According to evaluation results, our proposed methods can extract more related search keywords than that based on URL.

1 Introduction

Users search information they are interested in by using search engines in cyberspace. Due to the improvement of searching accuracy with development of technologies, it becomes possible that users can get various kinds of information by just inputting keywords representing the topic which users are interested in. However, it is not always true that users can hit upon search keywords properly. In some search engines like *Google*, you can get some results and some spelling suggestion even if you misspelled its search keywords. For example, '*I want to search one bank but forgot its name.*', '*I forget the search keywords but it's related to bank.*' and so on. In this case, just submitting '*bank*' as a search keyword to search engines will not produce satisfactory results since the search keyword is too general. It is important to present some related words to hit on as search for users who are unfamiliar to search engines.

On the other hand, it is possible to extract search keywords (inputted by users) and URLs accessed after users checking the logs recorded by the search engine sites. It is hard to collect this information because they are not open to the public. Recently, similar to survey on TV audience rating, a new kind of business appeared, which collects URL histories of Japanese users (called panel) who are selected without statistic deviation. By analyzing these logs (called panel logs) which are merged from accessing history of panels, it becomes possible to collect all the web pages (URLs) accessed as well as search keywords inputted by users.

In this paper, we propose two novel methods of clustering of search engine keywords by using access logs in order to find related search keywords associated

with the search keywords submitted by users. One is based on the web communities (set of similar web pages)[1]; the other is based on the set of nouns obtained by morphological analysis of web pages. According to evaluation results, our proposed methods can extract more related search keywords than previous methods that based on URLs. Experiment results also show that the methods based on web community as well as nouns have different characteristic while extracting the related search keywords.

The rest of the paper is organized as follows. Section 2 will review related works. In Section 3 we will explain technology which is necessary to understand our proposed methods. Our proposed methods of clustering search keywords using panel logs will be discussed in Section 4. Section 5 will show experimental results and evaluation, while Section 6 will give the conclusion.

2 Related Works

Until now, many works have been done based on web access logs as follows[1, 2]:

- users' behavior.[3, 4]
- the relationship between web pages.[5, 6]
- search engine sites.[7–9]
- access logs visualization.[10, 11]

Most of previous works are focused on user behavior by analyzing access logs from a certain web server. [12] uses proxy logs which are similar to the panel logs. To the best of our knowledge, we are the only research using the panel logs, detailed research is not done in others[4].

Results of works related search keywords clustering are not opened to the public because these results are directly connect to e-commerce business and it is hard to get search keywords data. [9] describes extracting related information like '*summer* and *vacation*' using search logs inputted in NTT[2] DIRECTORY. The purpose of this work is to extract synonyms based on search keywords frequency and intervals of input search keywords during a certain fixed period. Lycos and Microsoft publish search keywords research using access logs from search engine sits[7, 8]. These works classify search keywords based on the set of URLs and directories visited by the users after input of their original search keywords. Our methods use contents analysis of web pages visited by the users and community technique. Therefore these researches are also different from our methods.

3 Technology for Search Keyords Clustering

In this section, we describe concepts of panel logs, web community and web pages archive, which is necessary to understand the proposed methods.

3.1 Panel Logs

We use web access logs provided by '*Video Research Interactive Inc.*' in this paper and we call these access logs '*panel logs*'. This company is one of internet rating company. Figure 1 is the outline of panel log collection.

[1] In this paper, *community* means *web community*.
[2] Nippon Telegraph and Telephone Corporation.

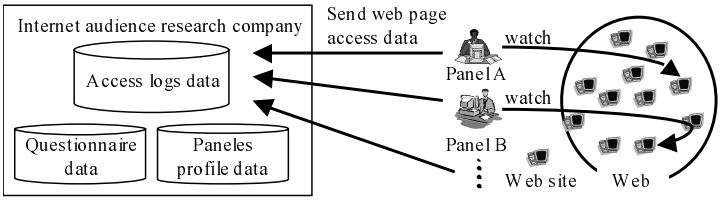

Fig. 1. Collection method of panel logs

Table 1. The details of the panel logs

An amount of data	about 10(Giga byte)
A term of collecting data	45(weeks)
A number of access	55,415,473(access)
A number of session	1,148,093(session)
A number of panels	about 10,000(persons)
A kind of URL	7,776,985(variety)
A kind of search keywords	334,232(variety)

– This company selects users based on RDD (Random Digit Dialing) and requests to become panels.
– Panels reply to some questionnaires and are requested to install the software in his (her) computer if they agree to become a panel.
– This software sends automatically panel's perusal information on web pages to the server of this company.

We do not use the questionnaire data and the panels profile data as in Figure 1 due to privacy reasons.

Details of the data are shown in Table 3.1. The panel logs consist of *panel ID, access time of web pages, reference second of web pages, URLs of accessed web page and so on*. The data size of panel logs we used is 10GB and all used panels are in Japanese. Panel ID is a unique ID which is assigned to each panel, and it is specific to an individual panel. Notice that panel logs also include search keywords submitted to search engines.

Usually, analysis of access logs uses the concept of *session* which is a sequence of web accesses. A session is defined as a set of URLs visited by a panel during a web browsing activity. We employed a well-known 30 minutes threshold for the maximum interval[13], such that two continuous accesses within 30 minutes interval are regarded as in a same session.

3.2 Web Community

In this paper, we define a web community as '*a set of relating web pages which are connected by hyperlinks*'. Most studies on web communities can be roughly classified into two kinds. One study is extracting dense subgraphs[14] and the other is extracting complete bipartite graphs[15]. The former one determines the borderline between inside and outside of web community using the theorem of "Maximum Flow Minimum Cut" based on network theory. The latter one

extracts complete bipartite graphs in web snapshot since hyperlinks between web pages which convey the message of common interest topics represented by complete bipartite graphs.

In our previous work, we created a web community chart[16] based on the complete bipartite graphs, and extracted communities automatically from a large amount of web pages.

3.3 Web Pages Archive

We periodically crawl web page written in Japanese. 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. Since the time of the web page crawling for the web communities is in between the duration of panel logs collection, there are some web pages which are not covered by the crawling due to the change and deletion of pages which were accessed by the panels.

Thus we define *matching factor* as follows to examine matching ratio between the URLs belonging to web-communities and the URLs included in panel logs.

$$matching\ factor = \frac{\begin{array}{c}the\ matching\ number\ of\ URLs\ belong\ to\ communities \\ and\ included\ in\ panel\ logs\end{array}}{the\ number\ of\ URLs\ included\ in\ panel\ logs}$$

We measured the *matching factor* and the result was only about 19%. If we delete the directory (file) part in URLs, the matching factor increases about 40% and when we delete the 'subdomain part', the matching factor improves further about 8%. By modifying URLs, about 65% of the URLs included in panel logs are covered by the URLs in the web communities. The details are mentioned in [4]

Our proposed methods require analyzing web pages are visited by various panels and these pages are one million. Therefore we check our web pages archive in order to examine whether the web page at the time of panel log collection exist. As a result, about 68 hundred thousand web pages at the time of panel log collection are saved in archive.

4 Methods of Search Keywords Clustering with Panel Logs

The search results in search engine sites (e.g. Yahoo!, Google, Lycos and so on.) are usually present as the lists of URLs related with the search keywords following page titles and abstracts of the pages. The users (who inputted search keywords in search engine sites) click and view his (her) interest pages, after reading page titles and abstracts. We consider that these clicked (viewed) pages (we call *clicked page sets*) are high relevance to the search keywords. Therefore, we extract many sets of a search keywords and clicked pages in panel logs, and we cluster the search keywords using these sets.

We remove multiple search keywords [3] because most search keywords are one word in panel logs as results of our preliminary experiment. We don't discuss about the methods of using multiple search keywords in this paper.

[3] Actually, in our experiments, we remove multiple search keywords inputted in Japanese. Therefore, the search keywords translated from Japanese to English may become multiple search keywords. For example, a word 'exchange rate' is single word in Japanese.

4.1 Definition of Feature Spaces

We newly define three feature spaces as *noun space*, *community space* and *URL space* in order to cluster search keywords. The noun space is created using nouns extracted from texts of clicked page sets. The community space is created using web community technique as mentioned in Session 3.2. The URL space is using previous works as mentioned in Session 2 and we define this feature space in order to compare with the above two feature spaces.

4.2 Definition of Similarity

We define A as a universal set of all search keywords:

$$A = \{a_1, a_2, \ldots, a_x, \ldots, a_n\}$$

(a_x is any search keywords and n is the number of total search keywords.)

We also define T_x which is feature space of a_x as follows.

$$T_x = \{t_{x1}, t_{x2}, \ldots, t_{xm}\}$$

(t_x is URL if feature space is the URL space, t_x is community ID[4] if feature space is the community space and t_x is the noun if feature space is noun. And m is a number of total feature space.)

Similarity of any two search keywords a_x, a_y in A is defined as:

$$K_{xy} = \frac{|T_x \cap T_y|}{|T_x \cup T_y|}$$

Moreover, it is possible to get frequency of URLs visited by different users by clicking information in panel logs. Therefore, we define similarity considering the frequency. Let T_x and T_y be the feature space of a_x and a_y which are any words and its' intersection of set are T_z. Then we define frequency space H_z considering the frequency as

$$H_z = \{(h_{z1}, h_{z2}, \ldots, h_{zj})\}$$

(h_{z1} is the total number of frequency of T_x and T_y, and j is the number of the feature space in the intersection of sets of T_x and T_y.)

Then there is similarity $K\mathrm{f}_{xy}$ considering the frequency as follows.

$$K\mathrm{f}_{xy} = \frac{The\ total\ number\ of\ H_z}{The\ total\ number\ of\ frequency}$$

We define *high frequency elements* as URLs, communities and nouns contained in any clicked page sets. For example, high frequency elements of URLs are *Yahoo!, MSN, Google and so on*, and high frequency elements of nouns are *I, today, news and so on*. We define a similarity *Kd* as the results of calculation excluded high frequency elements from feature space of T_x and T_y and a similarity *Kfd* as the results of calculation excluded high frequency elements from frequency space of H_z. Therefore, it is possible to say that similarity *Kfd* is the concept of tf*idf taken in similarity space *K*.

[4] Consider each communities have unique ID.

Table 2. Character of evaluated search keywords

Search keywords	A number of input frequency	Ranking of input times	A number(variety) of veiwed by the users after input search keywords		
			URL	Community	Noun
Bank	330(times) 94(sessions)	679	24(times) 20(variety)	24(times) 10(variety)	6,591(times) 1,725(variety)
University	799(times) 195(sessions)	168	31(times) 8(variety)	31(times) 5(variety)	5,255(times) 483(variety)

5 Experiments

In this paper, we experiment for search keywords inputted 4 times or more in panel logs since the small number of times of input is inaccurate. These search keywords are 30,000 and a number of high frequency elements of noun space is 4,565 words, in case of URL space is 4 URLs and in case of community space is 9 communities.

In this paper, we used panel logs which are collected from Japanese people. Therefore, all results have been translated from Japanese vocabulary items.

5.1 Search Keywords Cluster Viewer

We calculated similarity of 30,000 search keywords and make the *search keywords cluster viewer (SKCV)* which displays search keywords related to search keywords inputted by users (we call *target keywords*). The results of using the SKCV are shown in Figure 2,3. It is possible to adjust slide bar in the lower center of Figure 2,3. An edge will be connected to two words when relevance to nodes (which is extracted related search keywords) is high. And we setup the edges to become short when relevance to nodes is high.

Nodes with high relevance are displayed near each other although it is meaningless in the position of each node. It is possible to understand results to be divided into a group of banks and economic terms from a result of Figure 2. And in Figure 3, related search keywords are grouped as: Major banks group[5] (lower left side in the Figure 3), regional banks group (left side), internet banks (upper side), and economic terms (right side).

5.2 Evaluation

We evaluate three feature spaces and four similarities (*K, Kf, Kd and Kfd*) on three hundred thousand search keywords obtained in the experiment.

In the experiments, we test on two general words *Bank* and *University*. The details of these search keywords are shown in Table 2. We extract related search keywords related with *Bank* and *University* from three hundred thousand search keywords using three feature spaces and four similarities, and we evaluate relevance of search keywords and extracting related search keywords. We define the judgment of relevance to extract related search keywords based on search keywords (Bank and University) as follows.

[5] Well known banks in Japan.

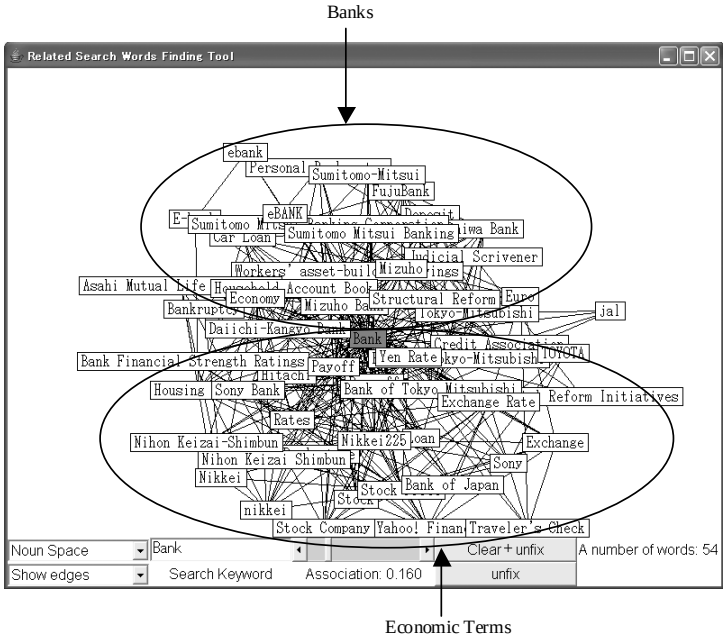

Fig. 2. An example of expression using noun space.

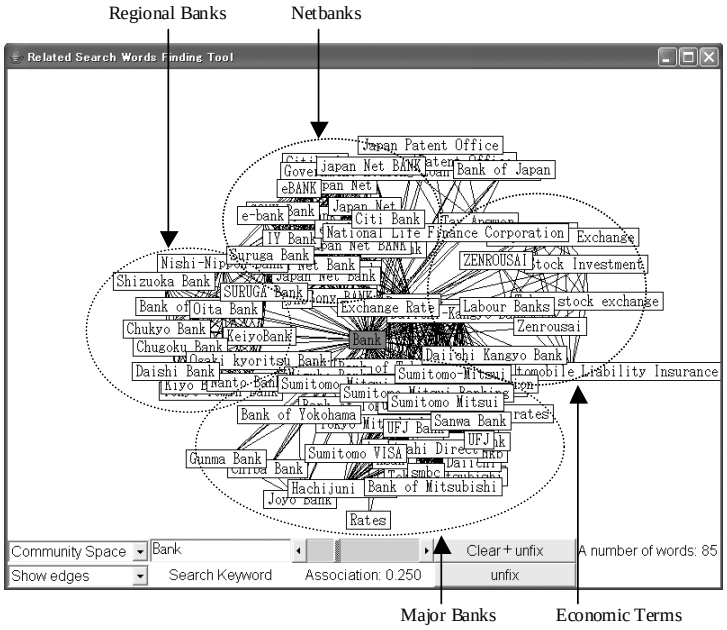

Fig. 3. An example of expression using community space.

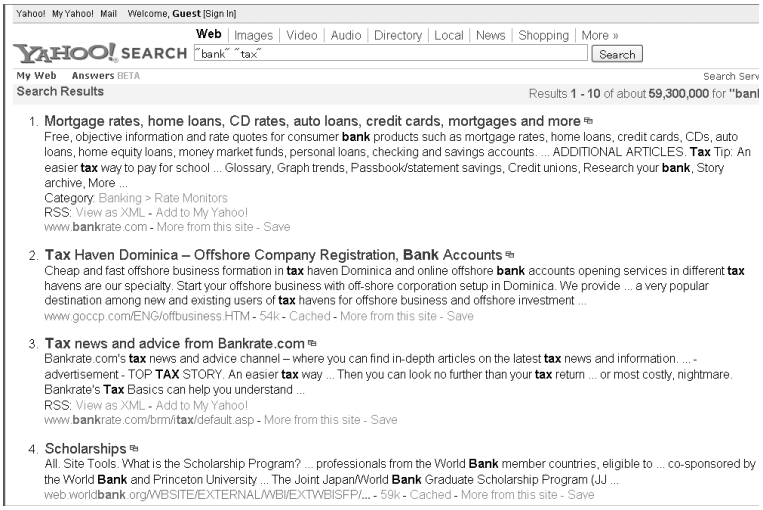

Fig. 4. An example of search keywords 'bank' and 'tax' in the Yahoo! site.

Category1,2 We decide *Category1* if the relation between related search keywords and search keywords is high. And we regard *Category2* if it is judged that there was a certain relation although the relationship is not higher than *Category1*.

Yahoo! judgment In Yahoo! site, we can see the results with page titles and brief abstracts as Figure 4 when users do search. We consider that the relevance between target keywords and displayed keywords (in SKCV) is high if both keywords exist simultaneously in the results of page titles or brief abstracts. We judge true when page titles include target keywords and displayed keywords simultaneously. Generally, brief abstracts consists of sentences in various places in a web page and each sentence is divided with *'...'*. We judge true in case of existing target keywords and displayed keywords appear in the same single sentence as two or more topics on one page may be treated. For example, we show a result search keywords 'bank' and 'tax' at the same time in the Yahoo! sites in Figure 4. We judge true because the page title or brief abstracts of the second result in Figure 4 includes 'bank' and 'tax' simultaneously. we judge false in the case of the first result in Figure 4 because the title does not include both keywords and they don't appear simultaneously in any of the single sentence of the brief abstract. We also judge false in the case of the third and the fourth result in Figure 4 because these titles and brief abstracts include only one side of keywords. Although Category1,2 are subjective because of our judgment, Yahoo! judgment is more objective than Category1,2.

5.3 Similarity and Feature Spaces

First, we show the results in Figure 5 when the target keyword is *Bank* and the number of displayed keywords is 10 and 100 . We define precision as *'the number of displayed keywords judged as Category1 or 2 divided by a total number.'* in

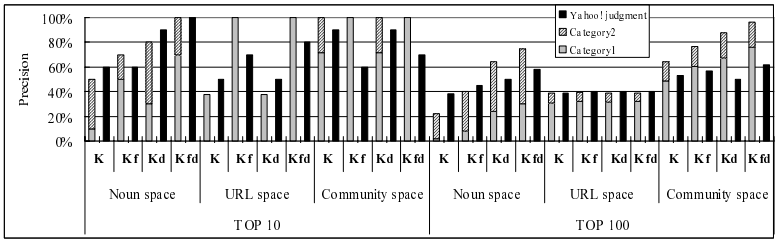

Fig. 5. Precision of 'bank'.

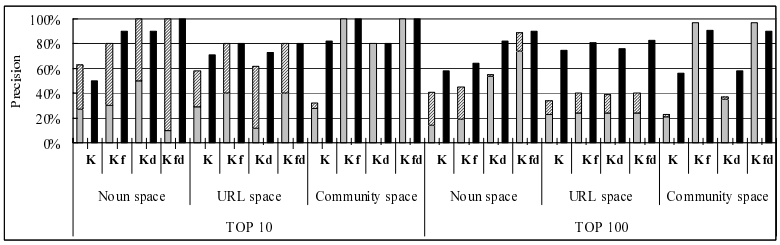

Fig. 6. Precision of 'university'.

case of Category1,2 and as *'a number of displayed keywords Yahoo! judgment divided by a total number.'* in case Yahoo! judgment, respectively. The noun space tends to extract Category2 more than other feature spaces. And we can get good precisions using similarity *Kfd* regardless of feature space and a number of extracted search keywords.

Next, we examine search keywords *University* and show the results in Figure 6[6]. The similarity *Kfd* is as good as the results of *Bank*. Similarity *Kfd* is the best precision unless the result of community space in case of target keywords 'bank'. And our proposed feature spaces (noun space and community space) are better precision than existing feature space (URL space) in almost all the cases.

6 Conclusion

In this paper we proposed two novel feature spaces and the method of similarity to cluster search keywords using panel logs based on web community and noun. We also show our tool for viewing search keywords cluster and evaluate our proposed methods.

As an application of this tool, it is possible to indicate related search keywords like *payoff* and *Workers' asset-building savings* using related search keywords extracting with noun space when the users do not remember search keywords related to bank.

[6] We omit a result of Category1,2 because their results show the same tendency as Yahoo! judgment.

Acknowledgment

We would like to thank Jun Hirai from Systems Integration Technology Center, Toshiba Solutions Corporation. And we also wish to thank Video Research Interactive, Inc. for providing the panel logs.

References

[1] Eirinaki, M., Vazirgiannis, M.: Web mining for web personalization. ACM Transactions on Internet Technology (ACM TIT) **3**(1) (2003) 1–27

[2] Cooley, R., Mobasher, B., Srivastava, J.: Web mining: Information and pattern discovery on the world wide web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97) (1997)

[3] Ungar, L., Foster, D.: Clustering methods for collaborative filtering. AAAI Workshop on Recommendation Systems (1998)

[4] Otsuka, S., Toyoda, M., Hirai, J., Kitsuregawa, M.: Extracting user behavior by web communities technology on global web logs. Proc. of 15th International Conference on Database and Expert Systems Applications (DEXA'2004) (2004) 957–968

[5] Su, Z., Yang, Q., Zhang, H., Xu, X., Hu, Y.: Correlation-based document clustering using web logs. 34th Hawaii International Conference on System Sciences (HICSS-34) (2001)

[6] Tan, P., Kumar, V.: Mining association patterns in web usage data. International Conference on Advances in Infrastructure for e-Business, e-Education, e-Science, and e-Medicine on the Internet (2002)

[7] Beeferman, D., Berger, A.: Agglomerative clustering of s earch engine query log. The 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2000) (2000)

[8] Wen, J., Nie, J., Zhang, H.: Query clustering using user logs. ACM Transactions on Information Systems (ACM TOIS) **20**(1) (2002) 59–81

[9] Ohkubo, M., Sugizaki, M., Inoue, T., Tanaka, K.: Extracting information demand by analyzing a www search log. IPSJ Journal **39**(7) (1998) 2250–2258

[10] Koutsoupias, N.: Exploring web access logs with correspondence analysis. Methods and Applications of Artificial Intelligence, Second Hellenic (2002)

[11] Prasetyo, B., Pramudiono, I., Takahashi, K., Kitsuregawa, M.: Naviz: Website navigational behavior visualizer. Advances in Knowledge Discovery and Data Mining 6th Pacific-Asia Conference (PAKDD2002) (2002)

[12] Zeng, H., Chen, Z., Ma, W.: A unified framework for clustering heterogeneous web objects. The Third International Conference on Web Information Systems Engineering (WISE2002) (2002)

[13] Catledge, L., Pitkow, J.: Characterizing browsing behaviors on the world-wide web. Computer Networks and ISDN Systems (27(6)) (1995)

[14] Flake, G., Lawrence, S., Giles, C.L., Coetzee, F.: Self-organization and identification of web communities. IEEE Computer **35**(3) (2002) 66–71

[15] Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.: Trawling the web for emerging cyber-communities. Proc. of the 8th WWW conference (1999) 403–416

[16] Toyoda, M., Kitsuregawa, M.: Creating a web community chart for navigating related communities. In: Conference Proceedings of Hypertext 2001. (2001) 103–112