Website Community Mining from Query Logs with Two-phase Clustering *

Lidong Bing, Wai Lam, Shoaib Jameel, and Chunliang Lu

Key Laboratory of High Confidence Software Technologies
Ministry of Education (CUHK Sub-Lab)
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Shatin N.T., Hong Kong
{ldbing, wlam, msjameel, cllu}@se.cuhk.edu.hk

Abstract. A website community refers to a set of websites that concentrate on the same or similar topics. There are two major challenges in website community mining task. First, the websites in the same topic may not have direct links among them because of competition concerns. Second, one website may contain information about several topics. Accordingly, the website community mining method should be able to capture such phenomena and assigns such website into different communities. In this paper, we propose a method to automatically mine website communities by exploiting the query log data in Web search. Query log data can be regarded as a comprehensive summarization of the real Web. The queries that result in a particular website clicked can be regarded as the summarization of that website content. The websites in the same topic are indirectly connected by the queries that convey information need in this topic. This observation can help us overcome the first challenge. The proposed two-phase method can tackle the second challenge. In the first phase, we cluster the queries of the same host to obtain different content aspects of the host. In the second phase, we further cluster the obtained content aspects from different hosts. Because of the two-phase clustering, one host may appear in more than one website communities.

Keywords: Website Community, Query Logs, Tow-phase Clustering.

1 Introduction

The World Wide Web has been extensively developed since its first appearance two decades ago. Various applications on the Web have unprecedentedly changed humans' life. Web search provides us a fast and accurate access to the useful information on the Web. Online encyclopedia contains human knowledge in different areas and makes it accessible to every Web user. Online shopping saves us the time for searching items in the malls. Corresponding to various

* The work described in this paper is also supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: CUHK413510).

applications on the Web, some important and challenging research topics have attracted a lot of attention from the research community. To facilitate a better Web search experience for the users, several directions are well studied such as search result ranking [20, 13], query log analysis [12, 23] and query reformulation [28, 4]; Online encyclopedias such as Wikipedia are employed to upgrade the performance of document classification and clustering methods [8, 17, 27] as well as to generate a hybrid ontology by merging them with the expert-edit ontology [22]; Different methods in Web information extraction (IE) are proposed to tackle the problem of extracting useful information from different types of Web pages, such as product description details extraction [29] and Web data record extraction [30].

Although the explosive growth and spread of the Web have resulted in a large scale information repository, it is still under-utilized due to the difficulty in automatically processing the information. One important task is to automatically organize and categorize a large amount of websites on the Web into different website communities. A website community refers to a set of websites that concentrate on the same or similar topics. Website community is different from online directories, such as Yahoo! Directory[1]. Online directories have predefined architecture, while the website community is not predefined and is adaptable according to the evolution of the real Web. There are two major challenges in website community mining task. First, the websites in the same topic may not have direct links among them because of competition concerns. One way to solve this problem is to calculate the content similarity of each pair of websites. This way is time consuming and cannot be generalized well. Second, a website may contain information about several topics. Accordingly, the website community mining method should be able to capture such phenomena and assigns such website into different communities.

In this paper, we propose a method to automatically mine website communities by exploiting the large scale Web query logs. Typically, each record in a query log data has user query, the clicked URL and other fields. Query log data can be regarded as a comprehensive summarization of the real Web. The queries that resulting a particular website clicked can be regarded as the summarization of the website content. Two websites in the digital video domain are clicked because of similar queries and they are thus connected by the queries indirectly. The strong summarization characteristic of query log data can help us overcome the first challenge mentioned above. The proposed two-phase method can tackle the second challenge. In the first phase, we cluster the queries of the same host to obtain different content aspects of the host. Then we construct sub-host vectors based on the obtained aspects of a host. In the second phase, we further cluster the sub-host vectors from different hosts and construct website communities. Because of the two-phase clustering, one host may appear in more than one website communities. The results of our website community mining can help us find the most important and popular website communities, the most popular hosts or URLs in these communities, as well as the most popular queries for

[1] http://dir.yahoo.com/

reaching these popular destinations. It also lets users know about the reliability of a website, because the number of users visiting a website can automatically tell the reliability of the information contained in the website.

2 Website community mining

2.1 Data representation

In Web query log, a query q and a host h are the two basic items of an effective log record. q is the query issued by a search engine user and h is the website clicked by the user after browsing the list of the search results. After aggregating the click through results of different users for the same query q_j, the clicking information of query q_j can be denoted as a vector $\boldsymbol{q_j} : (h_{1j}, h_{2j}, \cdots, h_{ij}, \cdots)$, where h_{ij} is the weight of host h_i in this vector. Similarly, by aggregating the click through results of different users on the same host h_i, the clicking information of host h_i can be denoted as a vector $\boldsymbol{h_i} : (q_{i1}, q_{i2}, \cdots, q_{ij}, \cdots)$, where q_{ij} is the weight of query q_j in this vector.

Each of h_i and q_j can be regarded as a pseudo-document. The weight of each term in the pseudo-document can be calculated with the classical term frequency-inverse document frequency (TFIDF) method, which is a numerical value and reflects how important a term is to a document in a collection or corpus. We adopt the TFIDF method to calculate the weights in $\boldsymbol{h_i}$. Let f_{h_i,q_j} denote the raw frequency of query q_j resulting the host h_i clicked in the search query logs. The term frequency tf_{h_i,q_j} is calculated as:

$$tf_{h_i,q_j} = \begin{cases} 1 + \log f_{h_i,q_j} \; if \; f_{h_i,q_j} > 0, \\ 0 \qquad\qquad otherwise. \end{cases} \tag{1}$$

The inverse document frequency idf_{q_j} of q_j is a measure of whether q_j is common or rare among all hosts. It is calculated as:

$$idf_{q_j} = \log \frac{|H|}{|\{h \in H : q_j \in h\}|}, \tag{2}$$

where H is the collection of all hosts and $q_j \in h$ indicates the host h was clicked after at least one user issued query q_j. Finally, the weight q_{ij} of query q_j in the host vector $\boldsymbol{h_i}$ is calculated as:

$$q_{ij} = tf_{h_i,q_j} \times idf_{q_j}. \tag{3}$$

Similarly, the weight h_{ij} of host h_i in the query vector $\boldsymbol{q_j}$ is calculated as:

$$h_{ij} = tf_{q_j,h_i} \times idf_{h_i}, \tag{4}$$

where tf_{q_j,h_i} is the term frequency of host h_i clicked after the query q_j issued by users, and idf_{h_i} is the inverse document frequency of h_i.

2.2 Website community mining with two-phase clustering

As we know, one site may contain several content topics. For example, Yahoo! portal contains information covering military, economy, etc. As a result, the queries related to Yahoo! are very diverse. In other words, the elements of vector $\boldsymbol{h_i}$ cover several topics. If $\boldsymbol{h_i}$ is directly fed into the community mining method, the host h_i will be grouped into a single community and the obtained community is not topic-cohesive. To tackle this problem, we propose a two-phase clustering method. In the first phase, we cluster the queries of the same host to obtain different content aspects of the host. Then we construct sub-host vectors based on the obtained aspects of a host. In the second phase, we perform clustering on sub-host vectors from different hosts and construct website communities. Therefore, the obtained website communities cover topics in a much finer granularity and a single host may appear in more than one website community according to its content aspects.

First phase clustering: In the first phase, we aim at mining different content topics covered by a particular host h_i. Let Q_{h_i} denote the set of queries that have non-zero weight in the vector $\boldsymbol{h_i}$. We construct a matrix as shown below:

$$\begin{pmatrix} \boldsymbol{q_1} \\ \boldsymbol{q_2} \\ \cdot \\ \cdot \\ \boldsymbol{q_{|Q_{h_i}|}} \end{pmatrix} = \begin{pmatrix} h_{11} & h_{21} & \cdots\cdots & h_{|H|1} \\ h_{12} & h_{22} & \cdots\cdots & h_{|H|2} \\ & \cdots & \cdots & \\ & \cdots & \cdots & \\ h_{1|Q_{h_i}|} & h_{2|Q_{h_i}|} & \cdots\cdots & h_{|H||Q_{h_i}|} \end{pmatrix}. \tag{5}$$

Each row of the matrix is a query vector and each query results in at least one clicking on host h_i. Suppose q_i and q_j are issued for searching different topics. Although it is possible that some general websites such as Yahoo! will be clicked for both of them, the host sets clicked for them should be significantly different due to a large number of websites in the topics related to each of the queries. Based on this observation, we can get different topics in the host h_i by performing clustering on the query set Q_{h_i} with the feature vectors given in the above matrix. After this phase, the query set of a host is partitioned into several clusters, and each cluster represents a content aspect covered by the host h_i.

***k*-means algorithm:** k-means algorithm is employed to conduct the clustering operation in this work. k-means is a method of cluster analysis and aims at partitioning a set of instances into k clusters. Each instance belongs to the cluster with the nearest mean. Given a set of instances $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\}$, where each instance is a real vector, k-means clustering partitions the n instances into k clusters ($k \leq n$) $\boldsymbol{C} = \{C_1, C_2, \cdots, C_k\}$ so as to minimize the within-cluster sum of squares:

$$\underset{\boldsymbol{C}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\boldsymbol{x_j} \in C_i} \|\boldsymbol{x_j} - \boldsymbol{\mu_i}\|, \tag{6}$$

Input:
// A set of instance to be clustered
$I = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\}$
k // Number of clusters
Output:
$\boldsymbol{C} = \{C_1, C_2, \cdots, C_k\}$ // Obtained clusters
Procedure k-means:
// Without replacement sampling
Sample an $\boldsymbol{x_i}$ as mean $\boldsymbol{\mu_j}$ of each C_j
For each $\boldsymbol{x_i} \in I$
Put $\boldsymbol{x_i}$ into $C_j = \text{argmin}_{C \in \boldsymbol{C}}\, distance(\boldsymbol{x_i}, \boldsymbol{\mu})$
End
While the mean of any cluster changes
For each $j \in \{1..k\}$
Recompute $\boldsymbol{\mu_j} = \frac{1}{|C_j|} \sum_{\boldsymbol{x_i} \in C_j} \boldsymbol{x_i}$
End
For each $\boldsymbol{x_i} \in I$
Put $\boldsymbol{x_i}$ into $C_j = \text{argmin}_{C \in \boldsymbol{C}}\, distance(\boldsymbol{x_i}, \boldsymbol{\mu})$
End
End
Return $\boldsymbol{C}$

Algorithm 1: Pseudocode for k-means.

where $\boldsymbol{\mu_i}$ is the mean of instances in C_i.

The problem is computationally NP-hard and an efficient heuristic algorithm is employed to obtain a local optimum, which is described in Algorithm 1. The distance function $distance()$ is defined based on the commonly used cosine similarity, which is used to measure the similarity between two vectors and is calculated as follows:

$$cosine(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{\boldsymbol{x_i} \cdot \boldsymbol{x_j}}{\|\boldsymbol{x_i}\| \|\boldsymbol{x_j}\|}. \tag{7}$$

The distance function is defined as:

$$distance(\boldsymbol{x_i}, \boldsymbol{x_j}) = 1 - cosine(\boldsymbol{x_i}, \boldsymbol{x_j}). \tag{8}$$

We intend to find at most 20 and at least 4 finer topics from a single host. Therefore, we can use the following formula to decide the number k in the k-means algorithm in the first phase:

$$k = \begin{cases} 4 & if\ |Q_{h_i}| = 20, \\ round(\frac{16}{1980} \times (|Q_{h_i}| - 20) + 4) & if\ 20 < |Q_{h_i}| \leq 2000, \\ 20 & if\ |Q_{h_i}| > 2000. \end{cases} \tag{9}$$

Second phase clustering: Based on the clusters of queries obtained in the first phase, we compose sub-host vectors for each host h_i. Let C_l denote a cluster of queries. The sub-host vector of h_i constructed based on C_l can be denoted as

$\boldsymbol{h_i^l} : (q_{i1}^l, q_{i2}^l, \cdots, q_{ij}^l, \cdots)$, where q_{ij}^l denotes the weight of q_j in the cluster C_l and is calculated as:

$$q_{ij}^l = \begin{cases} q_{ij} & if\ q_j \in C_l, \\ 0 & otherwise. \end{cases} \tag{10}$$

Then a topic matrix of a host is composed as below:

$$\begin{pmatrix} \boldsymbol{h_i^1} \\ \boldsymbol{h_i^2} \\ \cdot \\ \cdot \\ \boldsymbol{h_i^k} \end{pmatrix} = \begin{pmatrix} q_{i1}^1\ q_{i2}^1 \cdots\cdots q_{i|Q|}^1 \\ q_{i1}^2\ q_{i2}^2 \cdots\cdots q_{i|Q|}^2 \\ \cdots \quad \cdots \\ \cdots \quad \cdots \\ q_{i1}^k\ q_{i2}^k \cdots\cdots q_{i|Q|}^k \end{pmatrix}, \tag{11}$$

where Q is the set of all queries in the query log data. Some cluster in $\boldsymbol{C}$ may contain very small number of queries and it is not the main topic of the host. We remove such small clusters i.e., the corresponding sub-host vectors, so that to avoid the possible noise. First we sort the clusters in a descending order of their size. Then we accumulate the queries from the top cluster to the bottom one. If the number of accumulated queries exceeds a fixed percentage threshold, the remaining clusters are pruned. Here we set the threshold to be 60%.

After the sub-host vectors are constructed for each host, we perform k-means clustering again on the entire set of sub-host vectors from different hosts to we obtain a set of clusters, i.e., website communities, composed of different sub-hosts. Our two-phase clustering method can capture more finer topics covered by the same host and it can cluster one host into different communities according to these finer topics. This provides more flexibility in the automatic community mining task.

3 Experiments

3.1 AOL query log data

The query log data used in our experiments is the AOL data from [21] spanning three months from 1 March, 2006 to 31 May, 2006. The raw data is composed of queries and clicks recorded by a search engine. Each log record has 5 fields, namely, Anonymous User ID, Query, Query Time, Clicked URL, and URL Rank. The details of the fields are explained in Table 1. In the raw data, there are 19,442,629 lines of log records, 4,802,520 unique user ID, 1,416,831 unique URLs and 3,710,809 unique queries.

3.2 Data preprocessing

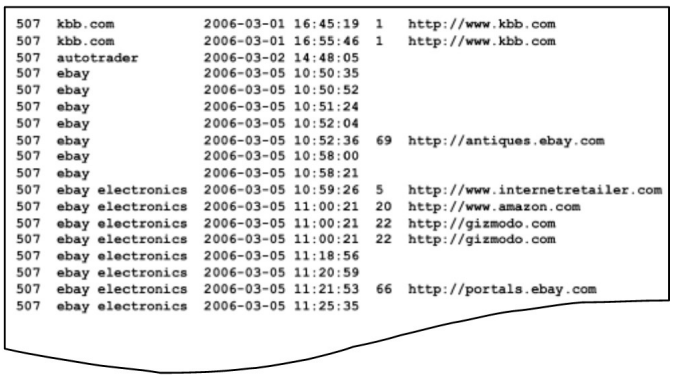
One fragment of the raw data is given in Fig.1. It can be seen that the raw data contains a lot of noise records so that we first conduct data cleansing before the data is fed into our method. The cleansing operations include incomplete record removal, navigation query removal, and trivial host/query removal. As we can

Table 1. The details of the fields.

	Meaning	Example
Anonymous User ID	Each search action has a unique Anonymous User ID, no user's information is revealed for privacy protection.	78253443
Query	The query issued by the user, it is case insensitive and the punctuation is removed.	lottery
Query Time	The time when the query was submitted for search.	2006-03-01 11:58:51
Clicked URL	If the user clicked on a search result, the domain portion (host) of the URL in the clicked result is stored.	www.calottery.com
URL Rank	If the user clicked on a search result, the rank of the clicked result is stored.	1

see in Fig.1, the Clicked URL field of some records is null, which indicates that the user did not click any result after he or she submitted the query. This kind of record is called incomplete record and should be removed. We also notice that quite a few queries are issued with the purpose of navigation, such as query "kbb.com" in Fig.1. These navigation queries are not helpful in semantic analysis and they are also removed. Finally, we also remove the trivial host and query. Trivial host means the host which has very few related queries in the entire log data. Similarly trivial query can be defined. We set the thresholds to be 20 and 10 for removing trivial hosts and trivial queries, respectively. This is an iterative cleansing process, because some trivial hosts are removed, some nontrivial queries may become trivial. After these cleansing operations, the remaining data set contains 3,031 hosts and 12,386 queries.

In Web search, different users have different searching habits. Even when they are searching for the same information, the queries issued may be different.



```
507  kbb.com           2006-03-01 16:45:19  1   http://www.kbb.com
507  kbb.com           2006-03-01 16:55:46  1   http://www.kbb.com
507  autotrader        2006-03-02 14:48:05
507  ebay              2006-03-05 10:50:35
507  ebay              2006-03-05 10:50:52
507  ebay              2006-03-05 10:51:24
507  ebay              2006-03-05 10:52:04
507  ebay              2006-03-05 10:52:36  69  http://antiques.ebay.com
507  ebay              2006-03-05 10:58:00
507  ebay              2006-03-05 10:58:21
507  ebay electronics  2006-03-05 10:59:26  5   http://www.internetretailer.com
507  ebay electronics  2006-03-05 11:00:21  20  http://www.amazon.com
507  ebay electronics  2006-03-05 11:00:21  22  http://gizmodo.com
507  ebay electronics  2006-03-05 11:00:21  22  http://gizmodo.com
507  ebay electronics  2006-03-05 11:18:56
507  ebay electronics  2006-03-05 11:20:59
507  ebay electronics  2006-03-05 11:21:53  66  http://portals.ebay.com
507  ebay electronics  2006-03-05 11:25:35
```

Fig. 1. A fragment of the AOL query log data.

Therefore, we need to do query normalization to merge together the queries that have the same semantic meaning. Firstly, we remove the stop words in a query, such as "a", "of", "the", etc. Then, each remaining word in the query is lemmatized to its base form. Finally, because the search engines are keyword matching based, we can resort the words alphabetically in a query without affecting the searching results. For example, the query "store of movies dvd" will be converted into "store movies dvd" after stop words removing, then it is lemmatized into "store movie dvd", and finally it is reordered as "dvd movie store".

3.3 Results and discussions

After the first phase, we obtained 7,875 sub-host vectors from the retained 3,031 hosts obtained after the cleansing. These sub-host vectors were fed into the second phase to mine the website communities. We set the number of clusters k, i.e. the number of website communities, to be 100.

Six generated website communities are presented in Table 2. From the results, we can observe that the hosts belonging to the same topic are grouped in the same community. Cluster 1 is about "tattoos" and most of the websites in it are related to this type of information. Only one of the results in Cluster 1 has gone astray, namely, "cheats.ign.com". Cluster 2 includes a set of websites on movies and most of the sites are very popular ones such as "movies.yahoo.com", "hollywood.com", and "ibdb.com". Cluster 3 includes a set of websites on musics, such as "music.aol.com", "mp3.com", etc.

It can be seen that the same URL "tattoojohnny.com" has occurred twice in the Cluster 1. This is due to the two-phase clustering of our method that has been done. In the first phase of clustering, this host is divided into several sub-hosts, such as "tattoojohnny.com$^{\beta}$" and "tattoojohnny.com$^{\alpha}$", covering different content aspects of this host. The obtained sub-hosts from different hosts may have different topic granularity so that the sub-hosts from the same host may be clustered into the same community in the second phase. Therefore, our method is not sensitive on the difference among the hosts and it can automatically adjust the concerned granularity of topics.

The host "cars.com" appears in two clusters, namely, Cluster 4 and Cluster 6. From the hosts in Cluster 4, we may easily observe that this community focuses on providing some guidance in car trading (e.g. "internetautoguide.com" and "auto.consumerguide.com") as well as some mobile magazine (e.g. "automobilemag.com"). While the community obtained in Cluster 6 focuses on providing information about classic cars (e.g. "classiccar.com") and old car trading (e.g. "oldcartrader.com"). After checking the website of "cars.com" manually, we observe that this website provides information for both new cars and used cars. Because of this nature of the website, it is assigned into two communities by our two-phase mining method. Thus, different topics of the same site are well revealed in different communities.

The size of the mined communities is given in Fig. 2. The smallest community contains 30 hosts and the largest community contains 165 hosts. The diversity of the size is relatively small, which also suggests that the two-phase clustering

Table 2. Some example clusters, i.e., website communities.

Cluster 1, 165 hosts	Cluster 2, 95 hosts
www.rankmytattoos.com	movies.yahoo.com
tattoo.about.com	movies.go.com
www.bullseyetattoos.com	www.the-numbers.com
www.vanishingtattoo.com	www.themovieinsider.com
www.tattoojohnny.com$^{\beta}$	www.hollywood.com
www.cheats.ign.com	movies.aol.com
www.cheatscodesguides.com	www.countingdown.com
www.tattoojohnny.com$^{\alpha}$	www.imdb.com
www.tattoonow.com	filmforce.ign.com
www.canismajor.com	movies.monstersandcritics.com
Cluster 3, 81 hosts	**Cluster 4, 95 hosts**
music.aol.com	www.internetautoguide.com
music.msn.com	www.carsearch.com
www.mp3.com	www.autobytel.com
www.allthelyrics.com	www.autotrader.com
www.rottentomatoes.com	www.modernracer.com
www.sing365.com	www.kbb.com
www.lyricsfreak.com	auto.consumerguide.com
www.lyricsdir.com	www.epinions.com
www.artistdirect.com	www.cars.com
www.lyricsdownload.com	www.automobilemag.com
Cluster 5, 48 hosts	**Cluster 6, 97 hosts**
www.123greetings.com	www.cars-on-line.com
www.bluemountain.com	www.oldride.com
www.hallmark.com	www.classicsandcustoms.com
yahoo.americangreetings.com	www.oldcars.com
www.dgreetings.com	www.antiquecar.com
www.1lovecards.com	www.classiccar.com
www.regards.com	www.vintagecars.about.com
www.egreetings.com	www.cars.com
www.americangreetings.com	www.oldcartrader.com
www.marlo.com	www.collectorcartraderonline.com

strategy and small sub-host removal are effective to generate more reasonable clustering result.

4 Related Work

A number of papers have been published describing characteristics of query logs coming from some of the most popular search engines, including [1, 2, 9–11, 14, 16, 18, 19, 25]. Silverstein et al. [24] is the first to analyze a large query log of the AltaVista search engine containing about a billion queries submitted in a period of 42 days. Similarly to other works, their results showed that the majority of the users (in this case about 85%) only visit the first page of results. They also

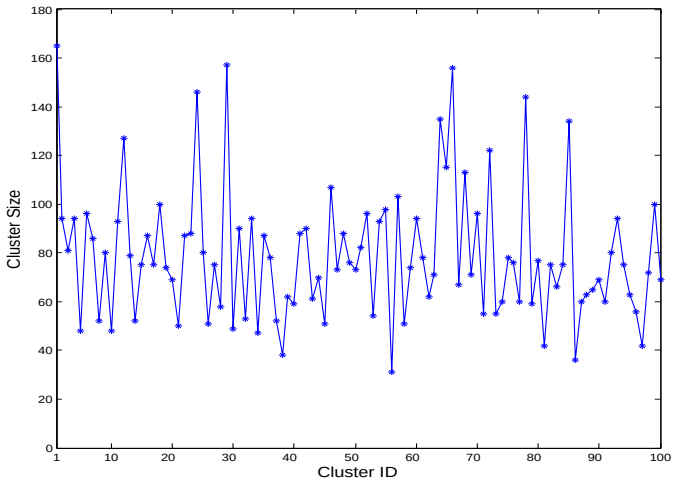


Fig. 2. Cluster size.

showed that 77% of the users' sessions end up just after the first query. As the authors stated, a smaller log could be influenced by ephemeral trends in querying (such as searches related to news just released, or to a new record released by a popular singer).

In [15] and [6], the authors analyzed a log made up of 7,175,648 queries issued to AltaVista during the summer of 2001. This second AltaVista log covers a time period almost three years after the first study was presented by Silverstein et al. and it is not as large as the first AltaVista log. Categorizing queries into topics is not a simple task. There are papers showing techniques for assigning labels to each query. Recent papers on the topic, including [3, 5, 7, 26], adopt a set of multiple classifiers subsequently refining the classification. Classification of the excite queries made in [25] shows that in no way is pornography a major topic of web queries, even though the top ranked query terms may indicate this. One sixth of the web queries have been classified as sex related. Web users look interested on a wide range of different topics, such as commerce, travel, and employment. Close to 10% of queries are about health and science.

5 Conclusions

In this paper, we proposed a method to automatically mine website communities by exploiting a large scale Web query logs. Query log data can be regarded as a comprehensive summarization of the real Web. The queries that lead to a particular website clicked are the summarization of the website content. The websites in the same topic are indirectly connected by the queries that convey information need in this topic. A two-phase method is proposed. In the first phase, we cluster the queries of the same host to obtain different content aspects

of the host. In the second phase, we further cluster the obtained content aspects. Because of the two-phase clustering, one host can appear in more than one website communities.

The results of our website community mining can help us find the most important and popular website communities, the most popular hosts or URLs in these communities, as well as the most popular queries for reaching to these popular destinations. It also lets users know about the reliability of a website. The amount of user visiting to a website can automatically tell the reliability of the information contained in the website.

References

1. Beitzel, S.M., Jensen, E.C., Chowdhury, A., Frieder, O., Grossman, D.: Temporal analysis of a very large topically categorized web query log. J. Am. Soc. Inf. Sci. Technol. 58(2), 166–178 (Jan 2007)
2. Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., Frieder, O.: Hourly analysis of a very large topically categorized web query log. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 321–328 (2004)
3. Beitzel, S.M., Jensen, E.C., Frieder, O., Lewis, D.D., Chowdhury, A., Kolcz, A.: Improving automatic query classification via semi-supervised learning. In: Proceedings of the Fifth IEEE International Conference on Data Mining. pp. 42–49 (2005)
4. Bing, L., Lam, W., Wong, T.L.: Using query log and social tagging to refine queries based on latent topics. In: Proceedings of the 20th ACM international conference on Information and knowledge management. pp. 583–592 (2011)
5. Broder, A.Z., Fontoura, M., Gabrilovich, E., Joshi, A., Josifovski, V., Zhang, T.: Robust classification of rare queries using web knowledge. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 231–238 (2007)
6. Fagni, T., Perego, R., Silvestri, F., Orlando, S.: Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data. ACM Trans. Inf. Syst. 24(1), 51–78 (Jan 2006)
7. Gravano, L., Hatzivassiloglou, V., Lichtenstein, R.: Categorizing web queries according to geographical locality. In: Proceedings of the twelfth international conference on Information and knowledge management. pp. 325–333 (2003)
8. Hu, J., Fang, L., Cao, Y., Zeng, H.J., Li, H., Yang, Q., Chen, Z.: Enhancing text clustering by leveraging wikipedia semantics. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 179–186 (2008)
9. Jansen, B.J., Spink, A.: An analysis of web searching by european alltheweb.com users. Inf. Process. Manage. 41(2), 361–381 (Mar 2005)
10. Jansen, B.J., Spink, A.: How are we searching the world wide web?: a comparison of nine search engine transaction logs. Inf. Process. Manage. 42(1), 248–263 (Jan 2006)
11. Jansen, B.J., Spink, A., Koshman, S.: Web searcher interaction with the dogpile.com metasearch engine. J. Am. Soc. Inf. Sci. Technol. 58(5), 744–755 (Mar 2007)

12. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 133–142 (2002)
13. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46(5), 604–632 (Sep 1999)
14. Koshman, S., Spink, A., Jansen, B.J.: Web searching on the vivisimo search engine. J. Am. Soc. Inf. Sci. Technol. 57(14), 1875–1887 (Dec 2006)
15. Lempel, R., Moran, S.: Predictive caching and prefetching of query results in search engines. In: Proceedings of the 12th international conference on World Wide Web. pp. 19–28 (2003)
16. Mat-Hassan, M., Levene, M.: Associating search and navigation behavior through log analysis: Research articles. J. Am. Soc. Inf. Sci. Technol. 56(9), 913–934 (Jul 2005)
17. Ni, X., Sun, J.T., Hu, J., Chen, Z.: Cross lingual text classification by mining multilingual topics from wikipedia. In: Proceedings of the fourth ACM international conference on Web search and data mining. pp. 375–384 (2011)
18. Ozmutlu, H.C., Spink, A., Ozmutlu, S.: Analysis of large data logs: an application of poisson sampling on excite web queries. Inf. Process. Manage. 38(4), 473–490 (Jul 2002)
19. Ozmutlu, S., Spink, A., Ozmutlu, H.C.: A day in the life of web searching: an exploratory study. Inf. Process. Manage. 40(2), 319–345 (Mar 2004)
20. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. In: Proceedings of the 7th International World Wide Web Conference. pp. 161–172 (1998)
21. Pass, G., Chowdhury, A., Torgeson, C.: A picture of search. In: InfoScale'06 (2006)
22. Ponzetto, S.P., Navigli, R.: Large-scale taxonomy mapping for restructuring and integrating wikipedia. In: Proceedings of the 21st international jont conference on Artifical intelligence. pp. 2083–2088 (2009)
23. Radlinski, F., Joachims, T.: Query chains: learning to rank from implicit feedback. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. pp. 239–248 (2005)
24. Silverstein, C., Marais, H., Henzinger, M., Moricz, M.: Analysis of a very large web search engine query log. SIGIR Forum 33(1), 6–12 (Sep 1999)
25. Spink, A., Ozmutlu, H.C., Lorence, D.P.: Web searching for sexual information: an exploratory study. Inf. Process. Manage. 40(1), 113–123 (Jan 2004)
26. Vogel, D., Bickel, S., Haider, P., Schimpfky, R., Siemen, P., Bridges, S., Scheffer, T.: Classifying search engine queries using the web as background knowledge. SIGKDD Explor. Newsl. 7(2), 117–122 (Dec 2005)
27. Wang, P., Domeniconi, C.: Building semantic kernels for text classification using wikipedia. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 713–721 (2008)
28. Wang, X., Zhai, C.: Mining term association patterns from search logs for effective query reformulation. In: Proceedings of the 17th ACM conference on Information and knowledge management. pp. 479–488 (2008)
29. Wong, T.L., Bing, L., Lam, W.: Normalizing web product attributes and discovering domain ontology with minimal effort. In: Proceedings of the fourth ACM international conference on Web search and data mining. pp. 805–814 (2011)
30. Zhai, Y., Liu, B.: Web data extraction based on partial tree alignment. In: Proceedings of the 14th international conference on World Wide Web. pp. 76–85 (2005)