2009 International Conference on Advanced Information Networking and Applications Workshops

Temporal Dynamics of User Interests in Web Search Queries

Aysegul Cayci, Selcuk Sumengen,
Cagatay Turkay
Sabanci University
{aysegulcayci,selcuk,turkay}@su.sabanciuniv.edu

Selim Balcisoy,
Yucel Saygin
Sabanci University
{balcisoy, ysaygin}@sabanciuniv.edu

Abstract

Web search query logs contain valuable information which can be utilized for personalization and improvement of search engine performance. The aim in this paper[1] is to cluster users based on their interests, and analyze the temporal dynamics of these clusters. In the proposed approach, we first apply clustering techniques to group similar users with respect to their web searches. Anticipating that the small number of query terms used in search queries would not be sufficient to obtain a proper clustering scheme, we extracted the summary content of the clicked web page from the query log. In this way, we enriched the feature set more efficiently than the content crawling. We also provide preliminary survey results to evaluate clusters. Clusters may change with the user flow from one cluster to the other as time passes. This is due to the fact that users' interests may shift over time. We used statistical methods for the analysis of temporal changes in users' interests. As a case study, we experimented on the query logs of a search engine.

1. Introduction

Query logs of search engines are an indispensable source of information to understand web search behavior. Query logs are analyzed in particular for search result re-ranking, search result clustering, query suggestion, or modification. In this work, we cluster users based on their queries and clicks in the query log made available by the AOL[2] search engine. We first cluster users with respect to their query terms and then analyze temporal variations among clusters. We use three statistical measures: overall overlap, distinct overlap and Pearson correlation proposed in [4] to assess the temporal changes in user interests.

We use document clustering technique presented in [10] by incorporating feature enrichment. Existing query clustering methods proposed in the literature deal with the problem of short queries since query clustering is similar to document clustering, but the small number of terms used in the queries prevents generation of meaningful clusters. One method to handle this problem is to consider the URL that the user has clicked as an implicit feedback, and use the content of the page addressed by the URL to enrich feature set of clustered queries [2][8]. Another method is to incorporate the number of same URLs that users have clicked into the similarity measure [3][5][9]. Wen, et. al, [9] compared the precision and recall of query clusters produced by different similarity measures. They report that clusters produced by the similarity of keywords together with shared number of clicks have the highest precision and recall. Beeferman and Berger, [3] proposed a content-ignorant query clustering method where similarity by keyword is not considered but queries with similar clicks are clustered. Chan, et. al's [5] query clustering algorithm improves Beeferman and Berger's by eliminating noisy user clicks. Clusters can also be improved by taking advantage of the fact that contents of clicked URLs provide more keywords, by including the keywords either from the top ranked result snippets[8] or from the top ranked result pages[2]. Fonseca, et. al, [6] use association rule mining to find related queries. Ross and Wolfram, [7] categorized term pairs in the queries and use query clustering to discover common terms of categories.

A cluster of users submitting similar queries share common interests and constitute a group of users. It has been shown by Beitzel, et. al, [4] that category popularity of queries change over the hours in a day. We investigate the flow between user groups over six time slices of a day. There exist few temporal query

[1] This work was partially funded by The Scientific and Technological Research Council of Turkey (TUBITAK) under grant number 105E185.

[2] Note that later on AOL withdrew this data from their website due to privacy concerns. In this paper, we do not intent to breach the privacy of individuals. The research results in this paper are only aggregate information which do not leak personal information.

978-0-7695-3639-2/09 $25.00 © 2009 IEEE
DOI 10.1109/WAINA.2009.71

log analysis studies in the literature. Adar, et. al, [1] analyze data on two query logs, a blog dataset, a news dataset and a TV dataset to correlate them on a hourly basis for detecting trends.

Previous study to understand user behavior from query logs, examine different aspects of queries. Re-finding behavior of users have been examined [12][13]. Lau et al. analyze query refinement habits of users [15]. Wang et al. try to understand web users' query behavior by linguistic analysis [16]. Wedig et al. analyze persistence of web search users and their topical interests [14].

Our contribution in this paper can be summarized as: (1) Extracting content summary of clicked URLs as descriptive keywords, (2) Performing a case study on clustering users with respect to their web search queries, (3) Statistical analysis of temporal dynamics of user interests. To the best of our knowledge this is the first work to study the temporal dynamics of user clustering.

2. User Clustering

Users issuing similar queries have common interests and belong to the same interest group. Since a user may have searched on several topics, she can be included in more than one interest group. We have used not only the keywords from the user queries but also the clicked URLs in our clustering process to cope with short queries which may prevent a proper clustering scheme to be discovered. The procedure that we used for extracting user interest groups can be outlined as:

1. Preprocessing query data to filter out queries that are not informative for clustering.
2. Collecting keywords from queries of users who clicked the same URL to obtain sets of query keywords which belong to the same page.
3. Using query keyword sets (obtained in Step 2) to cluster web pages with similar content.
4. Associating users to clusters of web sites (generated in Step 3) with respect to their queries.

Clusters resulting from the four steps outlined above are groups of users who submitted queries on similar topics. Details of each step are explained in the following subsections. We first introduce a formal definition of query set.

Definition: Let $q_1, q_2, \dots, q_n$ be the queries in query set Q, where an individual query is defined as:

$$q = \langle userid, \vec{w}, timestamp, url \rangle$$

And $userid$ is anonymous id of the user who issued the query, $timestamp$ is the query submission time, and if exists, url is the address of the website visited after the query issued. $\vec{w}$ is query keyword vector defined as follows:

$$\vec{w} = \langle w_1, \dots, w_d \rangle$$

2.1. Data Preprocessing

As we are mainly concerned with user's search behavior and interests, we use query keywords for clustering. We remove stop words and navigational queries (i.e. queries containing www or http) as they do not provide any contextual insight. Since URLs clicked by users are used for clustering as well as user queries, we also remove queries without clicked URL information. AOL data consists of about 20 million queries collected from approximately 650K users. After the preprocessing step, approximately 5 million queries are filtered out and our final set $\tilde{Q}$ consists of nearly 15 million queries.

2.2. Content Extraction

Web queries consist of 2-3 terms on average which may not be sufficient for clustering. For example; *user A* searching for "web mining" and *user B* searching for "query clustering" may have similar interests. However, capturing such similarities may require semantic analysis of queries. Instead of a complex semantic analysis, we use a simple approach by utilizing the clicked URL information. Clicks on the same URL by different users can be an indication that these users have common interests although they express it with different queries. In our example, if *user A* and *user B* issuing different search terms both click the URL of this paper, then we conclude that these queries are associated semantically. We can group the users directly according to the URLs they click, however this method has a disadvantage that contents of web pages are not considered during clustering. We first group the URLs where each group represents a specific topic. Grouping URLs requires extracting information from the corresponding web page to capture the topic of the URL which incurs an overhead of accessing these pages. In addition, not all the keywords in the clicked URL may be related with the users' interests visiting this page as a result of search. As an alternative, the contents of the web pages related to user interests can be extracted from the query log by using the related query keywords. Collection of query keywords associated with same URL can be considered as summary content of that web page. A formal definition of associating query words to web pages is as follows:

Definition: Let $P = \{p_1, p_2, \dots, p_m\}$ be the set web pages with URLs existing in Q, where a single web page is defined as:

$$p = \langle \vec{w}, url \rangle$$

Each web page has an associated content summary vector $p[\vec{w}]$, which is formed by collecting keywords of queries with the same URL of web page p.

$$p[\vec{w}] = \bigcup_{p[url]=q_i[url]} q_i\,[\vec{w}]$$

By collecting all the keywords users issue to get the link of a certain web page in the search results, we acquire the content of that web page enriching the feature set used during clustering.

Most of the URLs are clicked by a small number of users, so their associated keyword vectors, $p[\vec{w}]$'s, are generally very short (e.g. average number of keywords associated with URLs clicked by 1 to 10 users is around 4). A comparative result showing the range of user counts clicked on a URL, percentage of queries covered, and the resulting web pages' average $p[\vec{w}]$ vector lengths are presented in Table 1.

Table 1. Statistics on keywords collected for an URL

URL clicked by	Percentage of queries covered	Number of keywords collected
1-10 users	82.99%	3.5519
10-100 users	14.96%	29.5057
100-1000 users	1.84%	231.7192
1000-10000 users	0.12%	1725.0670

2.3. Clustering

We used an open source document clustering library [10]. Each vector of keywords $p[\vec{w}]$, associated with a specific web page p is provided as input to the library as an individual document.

Definition: Let $C^{url} = \{c_1^{url}, c_2^{url}, \ldots, c_k^{url}\}$, be the set of clusters of web pages that have similar contents formed by document clustering function f defined as:

$$f: P \rightarrow C^{url}$$

And output of clustering function f is a set C^{url}, where each cluster $c^{url} = \langle \vec{w}, \overrightarrow{url} \rangle$ is formed by a vector of descriptor keywords, and a vector of URLs.

Jaccard, Dice and cosine similarity measures are generally used metrics that have proven to be successful for document clustering [11]. In our experiments, Jaccard distance metric is used to form the clusters. We tried several similarity thresholds to observe how this affects our clustering performance (Figure 1-2). In Figure-1, the effect of different similarity threshold values on the number of clusters and the number of sites in those clusters can be observed. Throughout the clustering process, some of the web pages are not marked as members of any cluster since their distance to existent clusters are below the threshold. As expected, using higher similarity threshold values produce smaller sized clusters, and total number of clusters decreases as well.

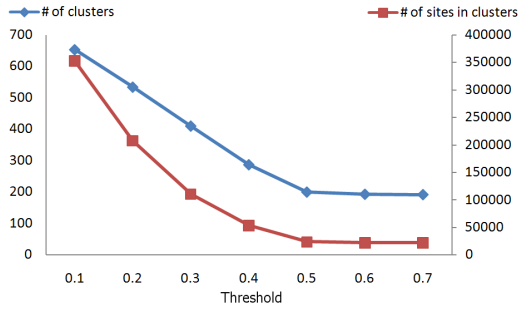

Figure 1. Number of clusters vs number of sites falling in clusters over varying threshold

As we decrease the threshold value, the number of web sites entering each cluster increase remarkably, thus increasing the number of queries successfully projected onto these clusters (Figure 2). A similarity threshold value of 0.1 causes 634 million pair-wise comparisons of web sites. This requires approximately 36 hours using a computer equipped with 2 dual-core AMD Opteron CPU's running at 2.6 GHz. Since average distinct keyword count for each URL is not large enough, choosing a smaller threshold value would not have a considerable effect on clustering performance. We choose the threshold value of 0.1 for clustering that utilizes more than 5% of query logs.

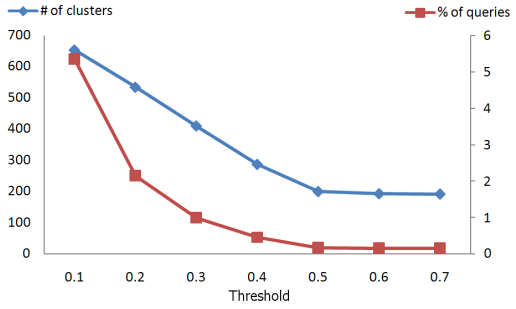

Figure 2. Number of clusters vs percent of queries projected onto clusters over varying threshold

After the clustering process, each cluster $c^{url} = \langle \vec{w}, \overrightarrow{url} \rangle$ consists of a vector of websites and described by a vector of words called best descriptors. These descriptors are chosen based on their frequency and their correlation with other frequent descriptors.

2.4. Associating Users to Clusters

Clusters of similar pages corresponding to URLs are formed as a result of the previous step. The last

step is to associate users who clicked those pages by looking at the query log.

Definition: Let $C^{userid} = \{c_1^{userid}, c_2^{userid}, \dots, c_k^{userid}\}$, be the set of clusters of users that have clicked on web pages of the same cluster, where each cluster $c^{userid} = \langle \vec{w}, \overrightarrow{userid} \rangle$ is a tuple formed by using corresponding URL cluster c^{url}, and query set Q with the following equations:

$$c_i^{userid}[\vec{w}] = c_i^{url}[\vec{w}] \qquad (1)$$

$$c_i^{userid}[\overrightarrow{userid}] = \bigcup_{q_j[url] \in c_i^{url}[\overrightarrow{url}]} q_j[userid] \qquad (2)$$

A manually selected subset of the resulting clusters C^{userid} is shown in Table-2. Each row represents a cluster with its associated descriptors $c^{userid}[\vec{w}]$, and the number of users projected onto these clusters. The frequency of descriptors in the query set is given in brackets beside each descriptor. As the total number of clusters generated was high, we only include top ranked clusters according to user counts and also eliminate few of them due to their similar characteristics with the existing ones.

Table 2. A subset of resulting clusters with descriptors

Rank	Number of users	Best Descriptors
1	18987	school[5472], high[2987], district[258], middle[227], elementary[161], public[148], central[129], catholic[107], unified[79], independent[78], …
2	17607	free[7950], online[292], music[198], printable[150], sex[140], web[130], porn[118], games[97], flash[93], download[92], email[73], …
3	16517	county[6651], tax[186], orange[179], property[167], public[161], ohio[145], sheriff[82], humane[77], franklin[60], suffolk[59], jefferson[59], …
4	14829	credit[2212], federal[930], union[246], card[137], employees[60]
5	14605	new[6558], york[1179], jersey[446], england[213], mexico[190], orleans[177], city[149], state[136], zealand[107], hampshire[105], …
6	13181	real[4810], estate[2708], sale[239], estatereal[158], realty[97], prudential[82], remax[58], banker[48]
9	7307	american[3146], native[219], african[185], indian[74], idol[69], express[61], association[57], legion[43]
10	7144	church[3704], baptist[633], catholic[272], st.[214], methodist[191], united[186], lutheran[174], christ[171], presbyterian[141], christian[109], episcopal[107], grace[86], bible[75], calvary[70], orthodox[70], god[61], trinity[59], hope[47]
15	5290	club[2035], country[257], golf[128], night[72], yacht[49]
47	3272	world[1322], war[84], warcraft[48], cup[45]

We have performed a survey to evaluate the quality of the resulting clusters. In our survey, 25 subjects (among graduate students and professors) were given first five clusters shown in Table 2, and were asked to "Score the relevancy of words in each group if they use a combination of them while issuing a query in a web search engine". Clusters have an average relevancy score of 65.2. Results of the survey can be seen in Table 3, where each cell contains the total number of subjects who marked that relevancy score interval for the corresponding cluster.

Table 3. Results of Cluster Quality Survey

Cluster Rank	0-20	20-40	40-60	60-80	80-100
1	0	1	3	9	12
2	1	1	4	11	8
3	3	5	10	6	1
4	0	1	2	12	10
5	2	4	6	7	6

Precision of clusters can be further improved using a phrase based clustering approach which we left as a future work. This approach might solve issues of projecting distinct phrases having common words to the same cluster. Cluster 47 in Table 2 that contains phrases "world war", "world of warcraft" and "world cup", is an example of the mentioned problem.

2.5. An Illustrative Snapshot

Stages of clustering are described with a sample flow in Figure 3. Blocks in the figure represent: sample set of queries from the query log, summary content of web pages, web page clusters and user clusters.

A subset of user queries is shown in Figure 3-a. Figure 3-b is the summary content of web pages which is obtained after content extraction phase described in Section 2.2. Words like "rankings", "high", and "school" constitute the summary content of the web page "prepreview.com". Clusters seen in Figure 3-c are produced by URL clustering phase which is described in section 2.3. The first cluster C^{url} in Figure 3-c contain the web pages like "topschools.com", "preview.com" whose summary contents contain the words "school", "high", etc. Users residing in C^{userid} in Figure 3-d have interests in topics best described by the corresponding descriptors. User clusters are obtained after associating users to URL clusters which is described in section 2.4.

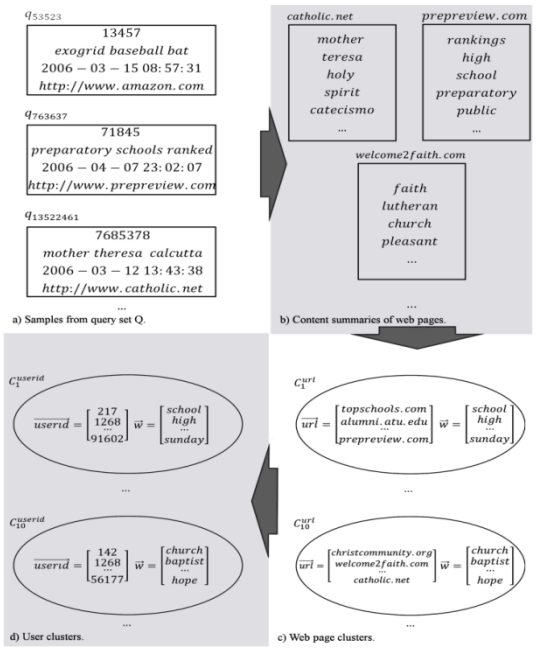

Figure 3. An illustrative snapshot of user clustering

3. Temporal Analysis

Temporal analysis of query logs was previously performed for measuring query repetitions over hours of day by overlapping and correlating query sets in a recent work by Beitzel, et. al [4]. In this work, our aim is to find a relation between user interest groups in successive time intervals. We adopted the statistical measures proposed in [4]: distinct overlap, overall overlap and Pearson correlation, for calculating the overlap ratio of clusters and find correlation between user sets.

Calculations are done on a temporal basis, between cluster pairs. For the sake of simplicity, we split hours of a day into six time slices (12 a.m. – 4 a.m., … , 8 p.m. – 12 a.m.); however, this analysis can be further investigated with varying time slices. For each time slice, we obtained a $n \times n$ correlation matrix M, of clusters $c_1^{userid}, c_2^{userid}, \ldots, c_n^{userid}$. M_{ij} refers to the correlation between clusters c_i^{userid} and c_j^{userid}, where the value indicates the magnitude of user flow between clusters.

In Figure 4, average correlations between all pairs of user clusters are given in a temporal basis for each statistical measure which is computed by:

$$\sum_{i<j} M_{ij}/n^2 - n$$

In our approach, we consider all the queries of a user in a given time slot, and obviously these queries don't have to be on the same topic. Consequently, a user might be a member of several clusters at the same time. Therefore, high value of average correlation between clusters means that users are shared among clusters, and Figure 4 can be interpreted as an indicator of user interest dispersion. Furthermore, this result shows that users issue queries on different topics throughout the day, while they are more focused during night.

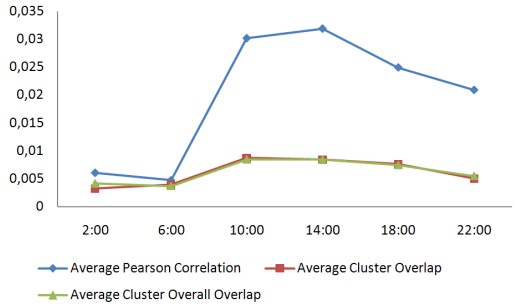

Figure 4. Average correlations between clusters in successive time slots

After the comparison of average similarities between successive time slots, Pearson Correlation appears to be the most distinctive measure, so we extracted the clusters having the maximum Pearson Correlation values (Figure 5).

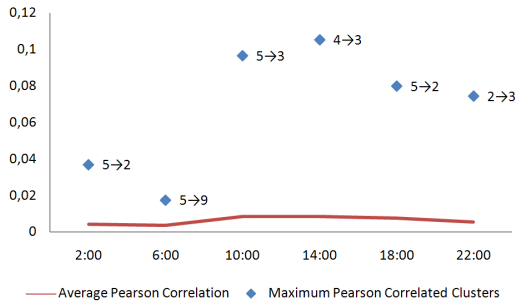

Figure 5. Maximum user flow between clusters for successive time slots

Amount of correlation between clusters in consecutive time slots is directly proportional to the magnitude of user flow between these clusters. Given two clusters c_A^{userid} and c_B^{userid} in successive time slots t_1 and t_2, we evaluate the user flow from c_A^{userid} to c_B^{userid} between t_1 and t_2 by calculating correlation between c_A^{userid} at t_1 and c_B^{userid} at t_2. Comparing these values provides us an overview of temporal changes in user interest groups.

As can be seen on Figure 5, a group of users from cluster with rank 5 (with best descriptors *new, york, jersey*, …) migrate to cluster with rank 2 (with best descriptors (*free, online, music,…*) and this is the highest flow at time interval 12 p.m. to 4 a.m.

4. Conclusion

Clustering is a well-researched problem with its vast application areas and Web is one of them. However, studies on extracting user interest from query logs, and clustering these users according to their interest are rare. Moreover, temporal dynamics of these user clusters did not draw the attention of data mining researchers much. In this paper, we first apply clustering techniques for grouping users with similar interests with respect to their web search queries. To accomplish this, we extract the content summaries of web pages with the proposed technique. Then, we cluster web pages to associate users to interest groups and study temporal dynamics of user clusters with well-known statistical methods. Our experiments show that query words alone are not enough to generate meaningful clusters, as expected. Since contents of the web pages visited in response to a query result enrich the feature set, we extract summary of the web page content. In our method, we use information in the query log for this purpose which is more efficient than crawling.

Statistical analysis of temporal dynamics of clusters verified that there exist groups of users whose interests shift from one common topic to other common topic between successive time slots. We believe that the results obtained from this study are useful for improving personalized search.

5. References

[1] Adar, E., Weld, D. S., Bershad, B.N., and Gribble, S. D. 2007. Why we search: Visualizing and predicting user behavior. In Proceedings of the 16th International Conference on World Wide Web.

[2] Baeza-Yates, R., Hurtado, C., and Mendoza, M. 2004. Query recommendation using query logs in search engines. In International Workshop on Clustering Information over the Web.

[3] Beeferman, D., and Berger, A. 2000. Agglomerative clustering of a search engine query log. In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, New York, NY, 407-416.

[4] Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., and Frieder, O., 2004. 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. ACM Press, New York, NY, 321-328.

[5] Chan, W.S., Leung, W.T., and Lee D.L. 2004. Clustering search engine query log containing noisy clickthroughs. In Proceedings of the 2004 International Symposium on Applications and the Internet. IEEE Computer Society Press, Washington, D.C., USA, 305-308.

[6] Fonseca, B.M., Golgher, P.B., Moura, E.S., and Ziviani, N. 2003. Using association rules to discover search engines related queries. In Proceedings of the 1st Latin American Web Congress. IEEE Computer Society Press, Washington, D.C., USA, 66-71.

[7] Ross, M. C. M., and Wolfram, D. 2000. End user searching on the internet: An analysis of term pair topics submitted to the excite search engine. Journal of the American Society for Information Science, 51(10):949-958, 2000.

[8] Wang, X., and Zhai, C.X. 2007. Learn from Web search logs to organize search results. In Proceedings of the 30th Annual International CAN SIGIR Conference on Research and Development on Information Retrieval.

[9] Wen, J.-R., Nie, J.-Y., and Zhang, H.-J. 2001. Clustering user queries of a search engine. In Proceedings of the 10th International Conference on the World Wide Web (Hong Kong, 2001). WWW'01. ACM Press, New York, NY, 162-168.

[10] Lemoine, J. (2008, July 17) C++ Clustering Library. [Online]. Available: wikipedia-clustering.speedblue.org/clustering.php

[11] Salton, G. and McGill, M. J. 1983. Introduction to Modern Information Retrieval. McGraw-Hill, New York.

[12] J. Teevan, E. Adar, R. Jones, and M. A. S. Potts, "Information re-retrieval: Repeat queries in Yahoo's logs," in Proc. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 151–158.

[13] M. Sanderson and S. Dumais, "Examining Repetition in User Search Behavior," LECTURE NOTES IN COMPUTER SCIENCE, vol. 4425, p. 597, 2007.

[14] S. Wedig and O. Madani, "A large-scale analysis of query logs for assessing personalization opportunities," in Proc. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 742–747.

[15] T. Lau and E. Horvitz, "Patterns of Search: Analyzing and Modeling Web Query Refinement, Courses and Lectures-International centre for Mechanical Sciences, pp. 119–128, 1999.

[16] P. Wang, M. W. Berry, and Y. Yang, "Mining longitudinal web queries: Trends and patterns," Journal of the American Society for Information Science and Technology, vol. 54, iss. 8, pp. 743–758, 2003.