Ilija Subasic
ilija@eccf.su.ac.yu
Zita Bosnjak
bzita@eccf.su.ac.yu
Saša Bosnjak
sasa.bosnjak@sora.co.yu

Faculty of Economics Subotica
Segedinski put 9-11
24000 Subotica

SEARCH PROFILES BASED ON USER TO CLUSTER SIMILARITY

Abstract
Privacy of web users' query search logs has, since last year's AOL dataset release, been treated as one of the central issues concerning privacy on the Internet. Therefore, the question of privacy preservation has also raised a lot of attention in different communities surrounding the search engines. Usage of clustering methods for providing low level contextual search, while retaining high privacy/utility is examined in this paper. By using only the user`s cluster membership the search query terms could be no longer retained thus providing less privacy concerns both for the users and companies. The paper brings lightweight framework for combining query words, user similarities and clustering in order to provide a meaningful way of mining user searches while protecting their privacy. This differs from previous attempts for privacy preserving in the attempt to anonymize the queries instead of the users.

Keywords
Search queries, clustering, privacy.

ACM classification
H.3 INFORMATION STORAGE AND RETRIEVAL, H.3.3 Information Search and Retrieval

JEL classification
M1 – Business Administration, M15 – IT Management

INTRODUCTION

The privacy on the Internet has been its weak spot since its mass adoption. In the last year much due to the AOL search query dataset release, the issues of web search privacy became one of the major issues. User queries are kept in transition logs which are *electronic record of interactions that have occurred during a searching episode between a Web search engine and users searching for information on that Web search engine* [6]. Performing data analysis on this data can provide an insight into the way users exploit search engines. Log analysis has been employed in implicit feedback for web search ranking, query spelling correction, query suggestion, personalization and monitoring [7]. In the past, these analysis provided benefit mostly for the search engine, but recently, the 4 generation search engines use different algorithms for retrieving information in context, best suited for the given user. To achieve this, search engines obviously need a lot of date from different sources including query logs. Therefore, the search engine must keep some records of the user`s queries if it is to produce personalized and context sensitive results. On the other side users would like search engines not to keep any of their personal searches, since it can contain some personal data, but also the users would like a better suited results to be provided by search engines. This leads to the Privacy-Utility tradeoff [9]. This tradeoff simply states that the more data from the log are eliminated; the query log mining produces less useful results. Determining the exact ratio of this tradeoff is very difficult. Despite of this, it is clear that some level of privacy, such as credit card numbers or other similar information must be removed from the logs. Some of the users are willing to move the ratio to utility side if it will bring better results, but some users want the ratio to be completely on the privacy side and want that the search engine does not keep any query logs. Different search engine functionalities make use of different types of information from the users. For example, they can use only the time gap between subsequent searches or they

can make more advanced inferences based on the user`s complete search history.

In order to protect user privacy different schemes for anonymizing query logs are devised. Most of these schemes can be classified into two major groups [9], database and network oriented groups. The most of them employ the notion of k-anonymity and outliner detection. All different approaches to anonymizing query logs hide users, not the search words, which users actually want to protect. No user is going to object to the logging of his visit to Google as long as his query is protected. Instead of masking keywords, most algorithms use some heuristics to remove them. This paper focuses on the idea of masking the keywords used in a query. This is done by grouping the keywords together into clusters and then using these clusters as the basis for providing different utility.

The paper continues by examining the privacy issues of search query logs in section 2, than explains the proposed keywords clustering in section 3, and than describes one experiment and its results (section 4).

1. SEARCH QUERY PRIVACY ISSUES

Last August AOL released what is known as 600k+ user search records, triggering a massive dispute over data privacy protection. The data consisted of three months (March - May) search queries with a user id, links followed, date and ranks. Although the data was anonymous, the famous NY Times article found that it is possible to identify some of the users by examining queries. Three highly ranking employees resigned over what *Cnnmoney* magazine listed as 57 dumbest business moves in 2006. The notion that someone's private search data does not belong to himself/herself hit the world bringing the issues of web search log privacy in the spotlight of internet community as well as governing bodies such as European Commission. Looking from one year perspective it can be said that the AOL search database release had forced other companies to clarify the way they are retaining user data.

If user tracking raises so many questions, why would the search engines (or other) store them? There are a number of reasons for this and Google [5] highlights 3 reasons for keeping data, which appear to be very general.

The first reason is the improvement of service using log data. Discovering patterns from log data has been around for about 10 years and is an indispensable technique for major e-commerce organizations. Since web servers are logging every (or almost every) user request, it is possible to track user behavior and adapt the content of a web page to specific user. Log keeping in e-commerce is identical to the brick and mortar store where someone is noting when did you enter and exit, what products did you bay, what products did you look at and for how long, what did you ask the sellers, etc. The same situation happens with search engines, which collect query words and associate them with IP and cookies. In this way queries can be associated with user and used for service improvements such as personalized search.

The second reason stated by Google is the creating of additional privacy safeguards and providing more certainty about data retention practices. To anonymize logs search engines use IP changing and cookies which protects the privacy in some extent. But if someone uses 8 digits of IP, you can still search for the users, and the search expands to only 256 computers. The way of anonymizing differs from search engine to search engine. Another issue is for how long one should keep this anonymized data. Several of the largest search engines are urged, under the force of public organizations, to disclose the way of anonymizing the data and the period of data keeping. Although they have not yet do this, more is known about their privacy policy. Google, which covers more than 70% of users (Nielsen Net-Ratings), has planned to keep cookie data until the year 2038, but in March disclosed the new privacy policy stating that data will be kept between 18 and 24 months, and later on, faced with The Article 29 Working Part letter, cut this period to 18 months (unless legally required to retain data longer). Ask.com is probably the leader in privacy protection, claiming that it would not retain user's Web search history at all, if the searcher didn't want it to. Two other big search engines ran by Microsoft and Yahoo! claim that they did not receive any formal

request from the EU body, although Microsoft later stated that it keeps data for 18 months. In July 2007 Ask.com and Microsoft formed a consortium that was to work on bringing the standard both for anonymizing and data keeping.

Microsoft published five basic principles for data privacy: user notice, user control, search data anonymization, minimizing privacy and data protection, legal requirements and industry best practices. The need for an industry wide standard is clear, but without collaboration of other major search engines, as well as non profit organizations, the chance for its adoption is very slim. Recently the Privacy international group, that looks over the privacy issues, urged all big companies to meet, and start discussing the cross industry standard.

Abiding the data retention requirements is the final reason for keeping logs. Log keeping in telecommunication is regulated in EU by the Directive on Mandatory Retention of Communications Traffic Data, published in May 2006, and in the US Electronic Communication Transactional Records Act of 1997, which requires that the data should be kept for 90 days. It is clear that governments need these data to examine potentially illegal activities of users. This leads to real problem of privacy. User log data contains track of its online behavior and governments can use it to predict if someone is likely to commit some crime (terrorism, sexual abuse, etc.). In order to predict something like this, government agencies are turning to data mining techniques for classification. These techniques are trying to find patterns inside data that are showing the membership of a given object to one of the defined groups. Although they have been around for a while and successfully applied in various fields, they are not quite appropriate. The main reason is the inaccuracy of classification. Most data mining models employ some kind of machine learning algorithm to find a function for classifying data and these functions are built within what is called the PAC (Probably Approximately Correct) learning framework. This basically means that the function is trying to classify new examples while tolerating a certain level of error. It is considered that a classifier with an 85 or 90 percent error rate is performing well in fields like marketing, banking or even machine vision. If we use the same classification for direct mailing problems, than this error rate of 15 to 10 percent means that 10 to 15 percent of all mails will be sent to a recipient with no real interest in them, or if we are trying to classify some illegal activity such as terrorist activities, 10 to 15 percent of people suspected and monitored will have absolutely nothing to do with terrorism. This rational was behind the Total Information Awareness program ran by DARPA. Although the funding of TIA was cancelled, some of its projects are still running. The main connection between data retention and these kinds of projects in the US is in the legal protection of data privacy. Basically, the US government can not gather personal data, but can obtain them through collaboration with private companies.

Therefore, the way data are kept and analyzed is more important today than ever before. It is up to companies to find out a way of anonymizing the data they gather and up to the scientist to develop data mining methods which can protect privacy of data they use.

2. ANONYMIZING KEYWORDS

When using a search engine users often get an almost endless list of web pages ordered in different way depending on the ranking algorithm that the search engine uses. In order to overcome this, an increased interest in user search queries was raised. Each time a user submit a query, a large amount of data is stored, that can be used for getting insight into what the user searches for. Mapping user queries to a specific topic can bring many improvements both in efficiency and effectiveness of a search engine. If we could in some way know in what category the query was, the answer to the query would be much more precise. The precise topic classification of queries is the issue which is widely dealt within the IR community. To our knowledge, best results are provided by semi supervised learning approach presented in [1], but some other approaches are also proposed [10, 11, 12]. As this paper has no intention of analyzing different approaches, they will not be examined any further, and the assumption that each query can be put into one or more classes is made.

Once the topics for queries are known, they are matched against user queries. In this way the keyword can be stored in some demilitarized zone, away from potential treats from outside. The next step is creation of a user-topic proportion matrix. This matrix can be formally written as:

$$A(a_{ij})_{m\times n}$$

where m are users, n are topics, and a_{ij} is the proportion of queries for a topic to all queries made by the user:

$$a(nj) = \frac{t_u}{N}$$

where t_u is the number of queries of the user that belong to class u and N is the total number of user's queries that is matched to the predefined queries.

The proportion given above is used instead of the number of queries because of the high variance. Using the matrix A clusters are built by applying the k-means algorithm. In [2] different clustering schemes are compared.

After the initial step of building the user-topic proportion matrix, the next step is its updating. Herein the problem of the existence of synonyms arises. If the query is classified into several classes (synonyms) the system should somehow decide which class the user is more likely to choose. To update the matrix in a correct way each cluster center of the keyword topics is updated in respect to the previous proportion of the given topic for a user. This can be formally written as:

$$a(nj)' = \frac{a(nj) + \left(\frac{1}{N} \times a(nj) \times p(n)\right)}{N}$$

where $a(nj)'$ is the new proportion of topics, N is the total number of user's queries and $p(n)$ is the probability of query word belonging to the class n.

Once the user clusters have been found, the next step tries to associate each user with the cluster in which he fits. This is done by calculating the similarity between user-topic proportions and cluster centers. Finding similarity of users is another issue. Several measures have been introduced to measure the similarity of the users by comparing them among themselves. This approach has 2 major problems. First is the sparsely distributed query, meaning that among a large number of users the probability of matching query to topic is low and the second problem is computation time. If one user is compared to a millions of others, that takes some time and effort. These problems can be avoided to some extent if the user is not compared to all the users but rather to the most representing ones (clusters). After the cluster centers are obtained user similarity to clusters can be calculated. In [3] similarity among the users from web logs is examined. The authors in [3] distinguish between 4 different similarity measures: usage based, frequency based, viewing time based and visiting order based. Since clusters represent topics, the page is of no importance for the analysis, and the frequency based similarity measure is used:

$$SM(u_i c_i) = \frac{\sum\sum a(u_i) \times a(c_i)}{\sqrt{\sum\sum (a(u_i))^2} \times \sqrt{\sum\sum (a(c_i))^2}}$$

where $a(c_i)$ is the frequency for the cluster topics and $a(u_i)$ for the user topics.

In this way we can obtain the similarity between a user and a cluster. This can be used to calculate the probability of the topic user is searching once he entered a query that is classified into more than a one class. Once the membership of cluster is found it can be used to provide more utility for the user. For example, if a user enters a query that is classified into 2 classes than based on his similarity to the certain cluster and its topic proportion the results can be presented with respect to this.

3. EXPERIMENT

The dataset used in this experiment is the AOL search data set released in August 2006, which is made up of about 600k+ of AOL users and their queries from May to March 2006. Before going further, some assumptions must be made. To make computation faster AOL data set has been reduced to a 1000 (randomly picked) users for model building. Only those examples which query could be correctly associated with a class have been used for classification. Cluster number is not examined in detail. A data set provided by [1] AOL editors have been used for query classification. It is a dataset of about 20k queries classified by humans into 20 topics (autos, business, computing, entertainment, games, health, holidays, home, misspell, news, organizations, other, personal finances - pf, places, porn,

research, shopping, sports, travel and url). Since some of the categories are more difficult to classify then others, misspelling and url classes have been removed.

Since the goal was not to build the best clusters the k-means algorithm was used. The other reason for using the k-means algorithm was that only frequency data were used, not the content of the pages. The number of used clusters was 10. The number of clusters was obtained using a normal distribution of the data, and than by measuring the within-cluster distance. The cluster centers are shown in table 1.

To obtain the relative importance of each category, cluster centers are normalized. This is shown in table 2.

Figure 1 shows the same proportion.

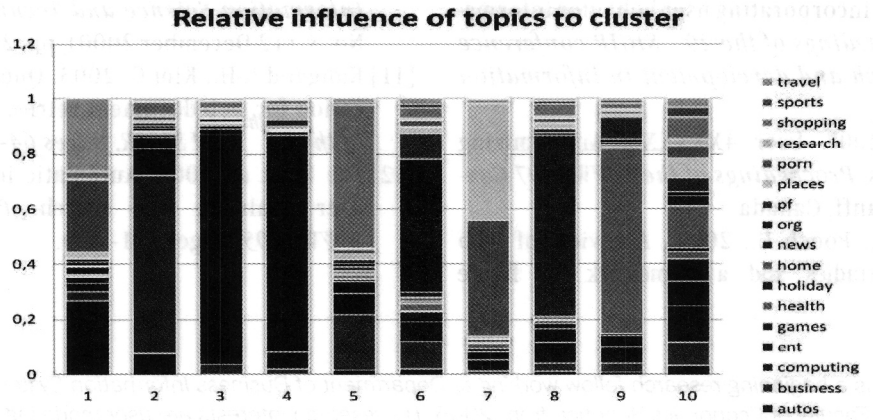
Table 1. **Topics cluster centers**

	1	2	3	4	5	6	7	8	9	10
autos	0,5441	0,0805	0,0061	0,0034	0,014	0,0193	0,0003	0,0089	0,0065	0,0302
business	0,0693	0,6808	0,0285	0,0165	0,1993	0,0945	0,0182	0,0324	0,033	0,0815
computing	0,066	0,1026	0,8325	0,0597	0,0744	0,0694	0,0674	0,0731	0,0578	0,1217
ent	0,0405	0,044	0,0188	0,782	0,0408	0,0335	0,0483	0,0489	0,0398	0,0842
games	0,013	0,0087	0,0123	0,0137	0,0101	0,0013	0,0184	0,019	0,0064	0,0792
health	0,0144	0,0111	0,01	0,0078	0,0157	0,0279	0,0317	0,0116	0,6557	0,0249
holiday	0,0529	0,0007	0,0002	0,0003	0,0033	0,006	0	0,0022	0,0015	0,0122
Home	0,0097	0,0035	0,0046	0,0027	0,0112	0,015	0,0096	0,0109	0,0091	0,1441
news	0,0263	0,0121	0,0111	0,0048	0,0144	0,0184	0,015	0,6509	0,008	0,0364
org	0,0223	0,0089	0,0045	0,0028	0,011	0,4633	0,0034	0,0097	0,0115	0,0133
pf	0,0084	0,013	0,0087	0,0247	0,0132	0,0874	0,0116	0,0153	0,0773	0,0482
places	0,0447	0,007	0,0077	0,033	0,0208	0,0225	0,0057	0,0198	0,0108	0,0673
porn	0,5381	0,0158	0,0117	0,0131	0,0176	0,0237	0,6979	0,0156	0,0116	0,0432
research	0,006	0,0071	0,006	0,0097	0,0042	0,0057	0,7046	0,0084	0,0052	0,0221
shopping	0,058	0,0477	0,0231	0,0147	0,5109	0,0365	0,0225	0,0453	0,022	0,0754
sports	0,0029	0,0046	0,0051	0,0027	0,0082	0,0149	0,0066	0,0103	0,0107	0,0354
travel	0,5381	0,0087	0,0061	0,0034	0,014	0,0193	0,0003	0,0089	0,0065	0,0302

Table 2. **Relative influence of a topic to cluster**

	1	2	3	4	5	6	7	8	9	10
autos	0,265	0,076	0,006	0,003	0,014	0,020	0,000	0,009	0,007	0,032
busine ss	0,034	0,644	0,029	0,017	0,203	0,099	0,011	0,033	0,034	0,086
computing	0,032	0,097	0,835	0,060	0,076	0,072	0,041	0,074	0,059	0,128
ent	0,020	0,042	0,019	0,786	0,042	0,035	0,029	0,049	0,041	0,089
games	0,006	0,008	0,012	0,014	0,010	0,001	0,011	0,019	0,007	0,083
health	0,007	0,011	0,010	0,008	0,016	0,029	0,019	0,012	0,674	0,026
holiday	0,026	0,001	0,000	0,000	0,003	0,006	0,000	0,002	0,002	0,013
home	0,005	0,003	0,005	0,003	0,011	0,016	0,006	0,011	0,009	0,152
news	0,013	0,011	0,011	0,005	0,015	0,019	0,009	0,657	0,008	0,038
org	0,011	0,008	0,005	0,003	0,011	0,483	0,002	0,010	0,012	0,014
pf	0,004	0,012	0,009	0,025	0,013	0,091	0,007	0,015	0,079	0,051
places	0,022	0,007	0,008	0,033	0,021	0,023	0,003	0,020	0,011	0,071
porn	0,262	0,015	0,012	0,013	0,018	0,025	0,420	0,016	0,012	0,045
resear ch	0,003	0,007	0,006	0,010	0,004	0,006	0,424	0,008	0,005	0,023
shoppi ng	0,028	0,045	0,023	0,015	0,520	0,038	0,014	0,046	0,023	0,079
sports	0,001	0,004	0,005	0,003	0,008	0,016	0,004	0,010	0,011	0,037
travel	0,262	0,008	0,006	0,003	0,014	0,020	0,000	0,009	0,007	0,032

Figure 1. **Relative influence of a topic to cluster**



CONCLUSION

Release of AOL search data into the public showed how far the internet community moved from the original ideas of information privacy. The key idea is that it is up to users to determine how their information is communicated to others. Internet as the global information system is based on user collaboration and trust is one of its main principles. AOL case showed that this principle can be easily broken. In whatever way release of AOL database may had hurt some (if not all) of its users it eventually pointed out some of the major issues in internet and media worlds. Although the privacy protection was always an important issue for academia and government, only the AOL case raised more public attention. The notion that users can be directly affected by some similar action forced the largest search engines to disclose the way they are using the collected data. Considered by some as just a PR stunt it is definitely a step into the right direction. In the last year there was more public debate over the search engine data privacy than there was since the introduction of the first search engine.

Although we are still far from the uniform and "complete" privacy protection, there is definitely a lot of progress in this area. To reach open standards for privacy, protection research must be taken in different areas. First of all, a uniform standard of data collection that will anonymize users must be developed. After that it is up to the legislation bodies to produce a legal framework that will guaranty user privacy. And the final step should be done by researches in developing methods, both technology and data based, that will allow for the analysis of the data in a privacy conscious way. This paper brings a data mining approach to preserving privacy by trying to aggregate query keywords into clusters.

Future research will go into direction of multiple clustering memberships. This means that based on the classes which are similar user can be included into several clusters.

REFERENCES

[1] Beitzel, S.M. et al., 2005. Improving automatic query classification via semi-supervised learning. *Proceedings of the Fifth IEEE International Conference on Data Mining 2005, Pages: 42 - 49*

[2] Zamir O., Etzioni O. 1999: A Dynamic Clustering Interface to Web Search Results. *Proceedings of the Eighth International World Wide Web Conference (WWW8)*, Toronto, Canada, May 1999.

[3] Xiao J. et al., 2001, Measuring Similarity of Interests for Clustering Web-Users, *Proceedings of the 12th Australasian Database Conference (ADCI01), 2001.*

[4] Witten, I. H., Frank, E., 2005, *Data Mining Practical Machine Learning Tools and Techniques second edition*. Morgan Kaufmann Publishers, San Francisco, USA.

[5] Google Log Retention Policy FAQ, 2007, *www.seroundtable.com/google_log_retention_policy_faq.pdf* (accessed on 09.10.2007), Google Inc.

[6] Jansen, B.J., 2006. Search log analysis: What it is, what's been done, how to do it, *Library & Information Science Research 28*(2006) 407–432

[7] Xiong, L., Agichtein E., 2007. Towards Privacy-Preserving Query Log Publishing, *Proceedings of the WWW2007 Conference*, Banff, Canada

[8] Agichtein,E. et al, 2006. Improving web search ranking by incorporating user behavior information, *Proceedings of the 29th SIGIR conference on Research and development in information retrieval.*

[9] Adar E., 2007. User 4XXXXX9: Anonymizing Query Logs, *Proceedings of the WWW2007 Conference,* Banff, Canada.

[10] Jansen B.J., Pooch U., 2000., A review of Web searching studies and a framework for future research, *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 3. (12 December 2000), pp. 235-246.

[11] Kangand I.-H., Kim G. 2003. Query type classification for web document retrieval. *Proceedings of the 26th ACM SIGIR, pages 64–71.*

[12] Lee U. et al, 2005. Automatic identification of user goals in web search. *Proceedings of WWW2005,* pages 391–400.

Biography:

Ilija Subašic *is a beginning research fellow working at Department of Business Information Systems and Quantitative Methods, Faculty of Economics Subotica from 2005. His research interests are user modeling, web mining and information retrieval.*

Zita Bošnjak *is a full professor employed at the Faculty of Economics Subotica. She is teaching courses in Intelligent systems, Data Mining and Management information systems. She holds a PhD in Information Systems and her research are soft computing, artificial intelligence and data mining.*

Saša Bošnjak *is a associate professor of Computer Science at the Faculty of Economics Subotica. He holds a range of courses in information engineering. His research interests are expert systems, artificial intelligence, soft computing, data mining, ebusiness, software development, fuzzy methods. He earned a PhD degree in Information Systems Faculty of Economics Subotica in 1995.*