# Dynamic Approach for Confidentiality Protection in Web Search

Bonam Sirisha
PG Student, B V C Engineering College,Odalarevu,AP, India.

N Krishnaiah
Associate Professor, B V C Engineering College,Odalarevu, India.

**Abstract – Web search engines (WSEs) collect and store information about their users in order to tailor their services better to their users' needs. Nevertheless, while receiving a personalized attention, users lose control over their own data. Search logs can disclose sensitive information and the real identities of users, thus creating serious risks of privacy breaches. Privacy preserving techniques seek to limit these risks by modifying the data. Although privacy is preserved, the data utility is reduced in a consequence of the data modifications. Achieving a good trade-off between privacy and utility can be a difficult task. In the present thesis we discuss the problem of limiting privacy disclosure risks in search logs while preserving enough data utility. The first part of this thesis focuses on the methods to prevent the gathering of information by WSEs. Since search logs are convenient in order to receive an accurate service, the aim is to provide logs that are still suitable to provide personalization. To that end, we propose a protocol which uses a social network in order to hide the queries submitted by a user. Results shows that users achieve good levels of privacy, meanwhile the response time of the protocol is acceptable.**

**Index Terms – Privacy, Private information retrieval, Social networks, Web search.**

## 1. INTRODUCTION

In recent years, the problem of properly protecting personal information has received a lot of attention due to the vast amount of information that is collected by Internet companies. Thanks to the evolution of information technologies, every transaction performed by an individual is stored, analyzed or even shared or disseminated. The queries submitted to a Web Search Engine are an example. WSEs have become one of the most successful services on Internet, responding to the demand for information facilities and services of the Information Society. By providing an easy way to access the Web, WSEs receive several hundred million queries each day. For example, during 2011, the WSE Google received near 5 000 million queries per day [4].

All these transactions are collected and stored as search or query logs.

The reasons why WSEs maintain the search logs can be classified into the three following categories: *Personalization*. WSEs provide their users result pages related to their searches in the web search process. From the huge amount of results, often thousands, only some links are relevant to the user's needs; meanwhile the other ones are irrelevant. *Improving search* Past searches are invaluable resources to im- prove the quality of search results. By knowing the frequencies of most formulated queries and most selected results, WSEs are able to improve the ranking algorithms [2] and to suggest reformulated queries that can add specificity to the user's initial query [6]. *Sharing data a*side from the information retrieval role, WSEs can act as an information source for third parties.

Web search engines (WSE) have become an essential tool for searching information on the Internet. In order to provide personalized search results for the users, WSEs store all the queries which have been submitted by the users and the search results which they have selected. The AOL scandal in 2006 proved that this information contains personally identifiable information which represents a privacy threat for the users who have generated it. In this way, AOL released a file containing twenty million queries made by 658,000 persons and several of those users were successfully tracked. In this paper, we propose a P2P protocol that exploits social networks in order to protect the privacy of the users from the profiling mechanisms of the WSEs. The proposed scheme has been designed considering the presence of users who do not follow the protocol (i.e., adversaries). In order to evaluate the privacy of the users, we have designed a new measure (the profile exposure level (PEL)). Finally, we have used the AOL's file in order to simulate the behavior of our scheme with real queries which have been generated by real users. Our tests show that our scheme is usable in practice and that it preserves the privacy of the users even in the presence of adversaries.

## 2. BACKGROUND DATA

The release of query logs from AOL with poor protection was a mistake normally regarded as a bad initiative taken by the company. Our objective is to provide a stronger anonymization so query logs collected by search engine companies do not pose a risk to the privacy of their users. In a standard WSE scenario,

the protection of query logs limiting improper disclosures can be addressed from two different points:

• Client-side. WSEs have no interest to give users control over the collected data because: (i) query logs helps to improve the information retrieval service and therefore the users' satisfaction; (ii) the business model of the WSEs is based on advertisements, whose efficacy relies on their personalization. Moreover, once the personal data are gathered, users can do nothing to prevent the WSE from using them for commercial purposes, putting their privacy at risk. The AOL case is an example of this. Server-side. The WSE wants to share or outsource the collected query logs without putting the privacy of users at risk. For this reason, it anonymizes the query logs using techniques such as privacy preserving data mining and statistical disclosure control. Statistical Disclosure Control techniques are needed to limit the risks of information inference in micro data. SDC techniques seek to disseminate statistical information in such a way that no confidential information about a specific individual can be inferred. To that end, data are modified to provide sufficient protection while trying to keep the information loss at minimum. Internet services collect and store information about their users in order to tailor their services better to their users' needs. Nowadays, these data are usually released in form of micro data because they have the advantage to be more flexible than the aggregated macro data. A micro data set is usually represented in a tabular form, where each record contains attributes (data) of an individual respondent (user). These attributes can be either numerical or categorical, and they are usually classified in the following categories, which are not mutually exclusive, depending on their content:

From the operational point of view, micro aggregation is defined in terms of partition and aggregation: *Partition* Records are partitioned into several clusters, each of them consisting of at least k records. *Aggregation* For each of the clusters a representative is computed, and then original records are replaced by the representative of the cluster to which they belong. Privacy concerns have a long history. They already existed in information retrieval from public databases. In this field, when a user submits a query, she is also exposing her interests to the database operator. Since the early 1980', several proposals to hide this personal information have emerged in order to address this situation. Those proposals can be classified according to the level of privacy that is offered to users: (i) schemes that provide perfect privacy but no personalized service; and (ii) schemes that partially protect the privacy of users and provide a certain degree of personalized service. We next summarize the different existing schemes of each type, starting with the ones that offer perfect privacy. A PIR (private information retrieval) protocol allows a user to retrieve a certain item from a database without allowing the latter to know which item is being acquired. Trivially, PIR can be achieved sending a copy of the entire database to the user, but this is very inefficient and unfeasible in practice.

However, it requires the existence of at least two copies of the same database. Besides, those databases cannot communicate between them. Accordingly, this proposal cannot work in a single server scenario like WSEs.

## 3. PROPOSED DESIGN

The behavior of the method is depicted in Algorithm 1. Data partition begins by calculating the centroid of the whole dataset and selecting the most distant record ($x_r$) to it. Then, a cluster is constructed with the k-1 least distant records to $x_r$. After that, the most distant record $x_s$ to $x_r$ is selected and a new cluster is constructed. The process is repeated until less than 2k records remain ungrouped. Remaining records are grouped together in a last cluster.

1. Require: X: original data set, k: integer
2. Ensure: X′: anonymized data set
3. X = X′
4. /*Data Partition*/
5. while $|X| \geq 3 \times k$ do
6. Compute centroid cx of all records in X
7. Find the most distant record xr to centroid cx
8. Form a cluster in X′ that contains xr together with its k−1 least distant
9. records
10. Remove these records from X
11. Find the most distant record xs to xr
12. Form a cluster in X′ that contains xs together with its k−1 least distant
13. records
14. Remove these records from X
15. end while
16. if $|X| \geq 2 \times k$ then
17. Compute centroid cx of all records in X
18. Find the most distant record xr to centroid cx
19. Form a cluster in X′ that contains xr together with its k−1 least distant
20. records
21. Remove these records from X
22. end if
23. Form a cluster in X′ with the remaining records
24. /*Data anonymization*/
25. for each cluster q in X′ do
26. Compute centroid cq of all records in q
27. Replace all records of q in X′ by their centroid cq
28. end for

The evaluation process includes two different types of tests: The first type checks the equitable distribution of messages around the network. Each user generates a unique query and sends it many times. This is the worst possible case because all the queries which are submitted by the same user are equal and

different from the queries sent by other users. As a result, these queries can be easily linked together. Nevertheless, if the system works correctly, the user who has generated all these queries remains hidden among the set of users who have submitted them. This kind of tests use synthetic queries (queries generated at random by a computer). Besides, the results of these tests provide the optimal values for. The second type of tests use real queries in order to evaluate the privacy level achieved by the users. These queries were extracted from the AOL file [4]. This file shows which queries were submitted by each AOL user (note that the real identity of each AOL user is not disclosed, only her queries). In this way, in our tests, a certain simulated user gets the personality of a certain AOL user. Therefore, the simulated user only sends the queries which were generated by her assigned AOL user. Note that each user submits a different number of queries. Depending on this number, their privacy might vary.

Number of neighbors they have. We have completed the table with the number of users who have an uncertainty percentage above 70% and 80%. Generally, the users with fewer connections are the most exposed ones. However, there are some users with many connections who also expose their profile to the WSE. This case can occur when the number of queries of these users is small or when their neighbors have sent them a small number of queries.

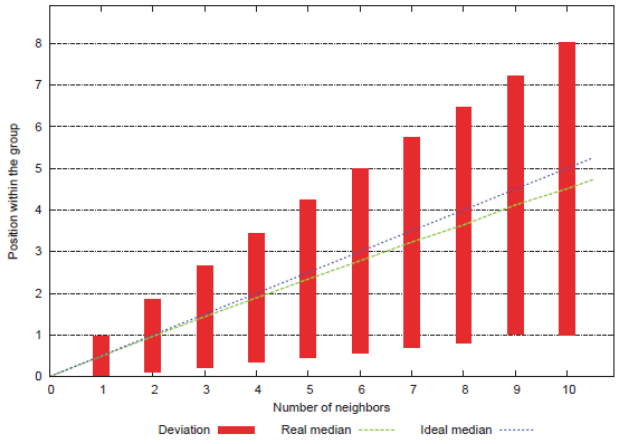


Figure 1-CLIENT-SIDE ANONYMIZATION

In the following sections we measure the privacy achieved by our method, and the utility of the protected data. We also evaluate the utility of the protected data in data mining processes, and provide an analysis of the frequency of queries and words. For each user id we have her original set of queries ' and the corresponding protected ones '', which have been protected by means of our micro aggregation method. Note that ' and '' can be seen as two random variables, which can take so many values as different queries they have and with probability proportional to the number of repetitions.

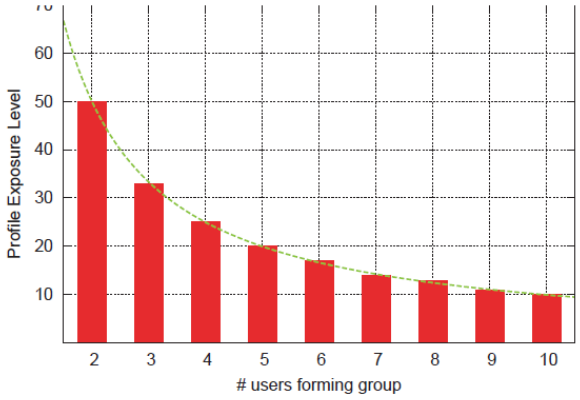


Figure 2-SERVER-SIDE ANONYMIZATION

Query logs are normally used in data mining processes for their analysis. To evaluate the utility of our protection method in data mining, we have considered clustering as a generic data mining process. There are several data mining techniques, from which clustering is one of the most popular [9, 10, 5]. Although the clustering of query logs is normally performed with some customized and more elaborated clustering, we show our results in a simple clustering just to give a generic idea. To provide micro aggregation at a user level, we have defined a new user distance and aggregation operator. The user aggregation described and it was designed in order to be as computationally efficient as possible. Note that the most important part is the aggregation of the queries since it is the information that will be more valuable in future analysis. Note also that queries are aggregated separately. An alternative could be to actually mix the terms of queries from different users to end up with new queries that somehow summarize all the users' queries. We opted for the first approach given the complexity that the second one imposes, and also because it already produced satisfactory results.

## 4. IMPLEMENTATION

Given a set of records, each one corresponding to the set of queries performed by each user, the basic steps of the method are: 1. Query processing and conceptual mapping: in order to semantically interpret textual queries, these are processed so that syntactical constructions (i.e. noun phrases) can be mapped to their conceptual abstractions modeled in a knowledge base. 2. Semantic data partition: clusters of query logs of at least k-users are created (fulfilling k-anonymity) by means of the MDAV micro aggregation algorithm. The cluster construction process and the centroid calculus method, on which the MDAV method relies, have been adapted in order to consider query semantics and the distributional properties of set-valued data. 3. Semantic query anonymization: clustered query logs are replaced by a synthetic set of queries that represent both their meaning and their distribution. Synthetic query logs are

constructed minimizing the information loss and the disclosure risk owing to the replacement.

Conceptual mapping In order to map individual queries to their conceptual abstractions in a knowledge base, we look for query-concept label matching's. Since words and NPs could be expressed (both in the queries and in the knowledge base) with different linguistic/morphological variations (e.g. water sport, water sports, this water sports, etc.), we apply additional analyses to detect equivalent formulations of the same concept. 1. Domain-independent words with very general meanings like determinants, prepositions and adverbs, called stop words, are removed from NPs (e.g. this water sports = water sports). 2. Both queries and concept labels in the knowledge based are stemmed to remove derivational affixes of the same root word, identifying equivalent terms. 3. When a query composed by several words is not found in the knowledge base, we look for simpler query forms by progressively removing adjectives/nouns starting from the one most on the left.

As a result of the query processing and the conceptual mapping, the query log of each user is represented by a set of categories with their corresponding taxonomical generalizations. Hence, we propose a measure that computes the semantic distance between categories, according to their taxonomical trees. In the client-side method, the quality of the service is related to the reliability of the interests of the user. We will consider using a specialized social network in order to get more homogeneous shared interests between users. This enhancement should improve the quality of the service. Nevertheless, there are some privacy issues that must be investigated. The vast amount of queries WSEs receive every day, should be taken into consideration in order to apply the presented methods. While the methods offer good privacy and data utility, their performance dealing with large datasets has not been evaluated. The way to deal with vast volumes of queries should be studied. The interpretation of queries is conditional on the knowledge base. Different knowledge can have different query' interpretations, thus providing different distances between two queries. Therefore, the obtained results by using our methods can be altered if we change the knowledge base. Different ways to compare search logs that do not depend on external sources should be investigated.

## 5. CONCLUSIONS

The existing micro aggregation techniques for query logs do not usually take the semantic proximity between users into account, which is negatively reflected in the usefulness of the resulting data. We have presented a new micro aggregation method for query logs, based on a semantic clustering algorithm. To that end, we use ODP as knowledge base to interpret the queries' terms and its hierarchy as metric space to define a distance operator. Aggregation is performed selecting randomly queries inside the same cluster. We have tested our proposal using real query logs from AOL. As we have seen, we obtain good results, both in terms of information loss and in terms of protection, which is guaranteed because our method ensures k-anonymity at user level.

## REFERENCES

[1] Z. Dou, R. Song, and J.-R. Wen, "A Large-Scale Evaluation and Analysis of Personalized Search Strategies," Proc. Int'l Conf. World Wide Web (WWW), pp. 581-590, 2007.

[2] J. Teevan, S.T. Dumais, and E. Horvitz, "Personalizing Search via Automated Analysis of Interests and Activities," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 449-456, 2005.

[3] M. Spertta and S. Gach, "Personalizing Search Based on User Search Histories," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI), 2005.

[4] B. Tan, X. Shen, and C. Zhai, "Mining Long-Term Search History to Improve Search Accuracy," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2006.

[5] K. Sugiyama, K. Hatano, and M. Yoshikawa, "Adaptive Web Search Based on User Profile Constructed without any Effort from Users," Proc. 13th Int'l Conf. World Wide Web (WWW), 2004.

[6] X. Shen, B. Tan, and C. Zhai, "Implicit User Modeling for Personalized Search," Proc. 14th ACM Int'l Conf. Information and Knowledge Management (CIKM), 2005.

[7] X. Shen, B. Tan, and C. Zhai, "Context-Sensitive Information Retrieval Using Implicit Feedback," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), 2005.

[8] F. Qiu and J. Cho, "Automatic Identification of User Interest for Personalized Search," Proc. 15th Int'l Conf. World Wide Web (WWW), pp. 727-736, 2006.

[9] J. Pitkow, H. Schu¨ tze, T. Cass, R. Cooley, D. Turnbull, A Edmonds, E. Adar, and T. Breuel, "Personalized Search," Comm. ACM, vol. 45, no. 9, pp. 50-55, 2002.

[10] Y. Xu, K. Wang, B. Zhang, and Z. Chen, "Privacy-Enhancing Personalized Web Search," Proc. 16th Int'l Conf. World Wide Web (WWW), pp. 591-600, 2007.

[11] K. Hafner, Researchers Yearn to Use AOL Logs, but They Hesitate, New York Times, Aug. 2006.

[12] A. Krause and E. Horvitz, "A Utility-Theoretic Approach to Privacy in Online Services," J. Artificial Intelligence Research, vol. 39, pp. 633-662, 2010.

[13] J.S. Breese, D. Heckerman, and C.M. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence (UAI), pp. 43-52, 1998.

[14] P.A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschu¨ tter, "Using ODP Metadata to Personalize Search," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), 2005.

[15] A. Pretschner and S. Gauch, "Ontology-Based Personalized Search and Browsing," Proc. IEEE 11th Int'l Conf. Tools with Artificial Intelligence (ICTAI '99), 1999.