Volume 3, Issue 5, May 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Publishing Search Information of User Protected Information

***R. Ramesh*[1], *Dr. T.Nalini* [2]**

[1]*Student, Final Year, M.Tech.,(CSE)*

Department of Computer Science & Engineering, Bharath University, Chennai-73, TN, India,

[2]*Professor*

Department of Computer Science & Engineering, Bharath University, Chennai-73, TN, India,

Abstract: Search engine companies gathering database information and histories of the users from search queries. These search details are useful for the researchers. Search engine companies are releasing different search details in order not to release perceptive information of the users. In this paper, I analyze the various algorithms for publishing relevant keywords, queries, and clicks of search details. I have to show how methods that achieve difference of k-anonymity are vulnerable to active attacks. Then i demonstrate that the stronger guarantee ensured by e-differential secured unfortunately does not provide any utility for this issue. I then implement an algorithm and show how to set its parameters to achieve probabilistic privacy. My paper concludes with a detailed study using real applications where the algorithm and previous work that achieves k-anonymity in search details releasing. My results show that the algorithm gives comparable utility to k-anonymity while at the same time achieving much stronger secured details.

Keywords: information storage and retrieval, db management, web search, IT and systems Security, protection.

1. Introduction

Search engines doing the major role for getting any information. Today's search engines not only collect and index webpage, but also collect and keep secured information about their users. They save the queries, clicks, IP-addresses, and other information about the interactions with users is called search information. Search information hold important information that search engines use their services better to their users' requirements. They enable the finding of trends, patterns, and anomalies in the search behavior of users, and they can be used in the development and checking of new algorithms to develop search performance and quality. In the world, scientists are like to acquire this valuable information for their own research purpose, search engine companies do not release them because they contain personal information of the users. For example, searches for political dealings, lifestyle choices, personal interest and search for diseases.

2. Related Work

The AOL was released the search information only in 2006, and it went into the dispute of technical history as one of the great tragedy in The search industry of AOL was released three months of search information of 650,000 users. The only result to protect user privacy was the replacement of user-ids with random numbers utterly insufficient protection as the "New York Times" showed by identifying a user from Lilburn, Georgia whose search queries not only contained identifying information but also important information about the friends' illness. The AOL search information release shows that simply replacing user-ids with random numbers does not prevent information disclosure. Other *ad hoc* methods have been studied and found to be similarly insufficient, such as the removal of names, age, zip codes, and other identifiers and the replacement of keywords in search queries by random numbers. In this paper, we compare formal methods of limiting disclosure when publishing frequent keywords, queries, and clicks of the users search details. The methods vary in the guarantee of disclosure limitations they provide and in the amount of useful information they provide and in the amount of useful information they maintain. We first describe two negative results. We show that existing proposals to achieve *k*-anonymity in search details are insufficient in the light of attackers who can actively influence the search details. We then turn to differential privacy, a much stronger privacy guarantee; however, we show that it is impossible to achieve good utility with differential privacy. We then describe algorithm *zealous,* developed independently by Korolova et al. and for us with the goal to achieve relaxations of differential privacy. Korolova et al. showed how to set the parameters of *zealous* to guarantee indistinguishability and we here offer a new analysis that shows how to set the parameters of zealous to guarantee probabilistic differential privacy a much stronger privacy guarantee as our analytical comparison shows. Our paper concludes with an extensive experimental evaluation, where we compare the utility of various algorithms that guarantee anonymity or privacy in search details publishing. Our evaluation includes applications that use search details for improving both search experience and search performance, and our results show that zealous output is sufficient for these applications while achieving strong formal privacy guarantees. We believe that the results of this research enable search engine companies to make their search details available to researchers without disclosing their users' sensitive information. Search engine companies can apply our algorithm to generate statistics that are probabilistic differentially private while retaining good utility for the two applications we have tested. Beyond publishing search details we believe that our findings are of interest when publishing frequent item sets, as zealous protects privacy against much stronger attackers than those considered in existing work on privacy preserving publishing of frequent items or item sets.

3. Frequent Itemsets

We introduce the problem of publishing of a search details. Queries, clicks, and other items of a search details. Search engines such as ***Bing, Google, or Yahoo*** details interactions with their users. When a user submits a query and clicks on one or more results, a new entry is added to the search details. Without loss of generality, we assume that a search details has the following syntax format:

(User-Id, Query, Time Clicks)

Where a *User-Id* identifies a user, a *Query* is a collection of keywords, and *Clicks* is a list of *url (*universal resource locator) that the user clicked on. The user-id can be identified in various ways like through Cookies, IP addresses, or User Accounts. A user history or search history contains of all search entries from a single user. Such a history is usually classified into various modules containing similar related queries. A paired query includes of two subsequent queries from the same user within the same session.

We say that a user history contains a keyword *k*, if there exists a search details entry such that *k* is a keyword in the query of the search details. A keyword histogram of a search details *S*, records for each keyword *k* the number of users c_k whose search history in *S* contains *k*. A keyword histogram, the query pairs histogram and the click histogram. We classify a keyword, query, consecutive query, and click in a histogram to be frequent if its count exceeds some predefined threshold *T*; when we do not want to specify whether we count keywords, queries, etc., we also refer to these objects as items.

want to specify whether we count keywords, queries, etc., we also refer to these objects as items.

With this term, we can define our aim is to publishing frequent items (utility) without disclosing sensitive information about the users (privacy). We will make both the notion of utility and privacy more formal in the next sections.

4. Experiment & Result

We introduce a search details publishing algorithm called *Zealous* that has been independently developed by korolova et al. and us. *Zealous* ensures probabilistic differential privacy, and it follows a simple With this terminology, we can define our goal as publishing frequent items (utility) without disclosing sensitive information about the users (privacy). We will make both the notion of utility and privacy more formal. In the first phase, *Zealous* generates a histogram of items in the input search details, and then removes from the histogram the items with frequencies below a threshold. In the second phase, *Zealous* adds noise to the histogram counts, and eliminates the items whose noisy frequencies are smaller than another threshold. The resulting histogram (referred to as the sanitized histogram) is then returned as the output. Fig. 1 depicts the steps of *zealous* and this algorithm will accept to add &

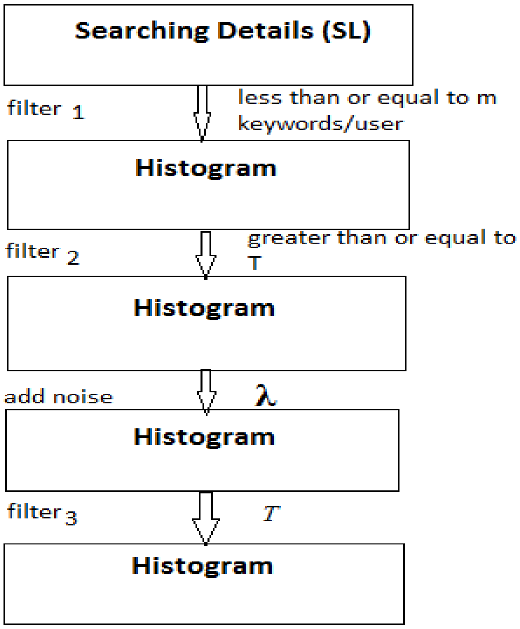

Fig. 1. Privacy preserving algorithm.

Algorithm ZEALOUS for publishing frequent items of a search details.

Input: Search details S, positive numbers m, λ, *T*, *T*′

1. For each user u select a set s_u of up to m distinct items from u's search history in S[3].
2. Based on the selected items, create a histogram consisting of pairs (k, c_k), where k denotes an item and c_k denotes the number of users u that have k in their search history s_u. We call this histogram the original histogram.
3. Delete from the histogram the pairs (k, c_k) with count c_k smaller than T.
4. For each pair (k, c_k) in the histogram, sample a random number n_k from the Laplace distribution Lap(λ),[4] and add n_k to the count c_k, resulting in a noisy count:
5. Delete from the histogram the pairs (k, c_k) with noisy counts
6. Publish the remaining items and their noisy counts.

To understand the purpose of the various steps one has to keep in mind the privacy guarantee we would like to achieve. From the above Steps 1, 2 and 4 of the algorithm are fairly standard. It is known that adding noise to histogram counts achieves e-differential privacy. However, the previous section explained that these steps alone result in poor utility because for large domains many infrequent items will have high noisy counts. To deal better with large domains, we restrict the histogram to items with counts at least T in Step 2. This restriction release information and thus the output after Step 4 is not e-differentially private. One can show that it is not even probabilistic differentially private. Step 5 discusses the information leaked in Step 3 in order to achieve probabilistic differential privacy.

In what follows, we will investigate the theoretical performance of *Zealous* in terms of both privacy and utility. Section 4.1 and 4.2 discuss the privacy guarantees of *Zealous* with respect to indistinguishabiligy and probabilistic differential privacy, respectively. Section 4.3 presents a quantitative analysis of the privacy protection offered by *Zealous*. Section 4.4 and 4.5 analyze the utility guarantees of *Zealous*.

4.1 Indistinguishability Analysis

Theorem 2 states how the parameters of *Zealous* can be set to obtain a sanitized histogram that provides indistinguishability.

Indistinguishability Analysis states how the parameters of *Zealous* can be set to obtain a sanitized histogram that provides indistinguishability.

To publish not only frequent queries but also their clicks. We suggesting to first determine the frequent queries and then publish noisy counts of the clicks to their top-100 ranked documents. In particular, if we use *Zealous* to publish frequent queries in a manner that achieves indistinguishability, we can also publish the noisy click distributions of the top-100 ranked documents for each of the frequent queries, by simply adding laplacian noise to the click counts with scale. Together the sanitized query and click histogram achieves indistinguisability.

4.2 Probabilistic Differential Privacy Analysis

The following theorem tells us how to set the parameters to ensure that *Zealous* achieves probabilistic differential privacy.

4.3 Quantitative comparison of probabilistic differential privacy and indistinguishability for *zealous*

We illustrate the levels of indistinguishability and probabilistic differential privacy achieved by *Zealous* for various noise and threshold parameters. We fix the number of users to (U), and the maximum number of items from a user to m=5, which is a typical setting that will be explored in our experiments. The tradeoff between utility and privacy: A larger results in a greater amount of noise in the sanitized search. The details of similar when the sanitized search details provides less utility (since fewer items are published) but a higher level of privacy protection.

Interestingly, we always have these data. This is due to the fact that probabilistic differential privacy is a stronger privacy guarantee than indistinguishabiligy, as well be discussion later.

4.4 Utility Analysis

Next, we analyze the utility guarantee of *Zealous* in terms of its accuracy. It is easy to see that *Zealous* provides perfect accuracy of filtering out infrequent items. Moreover, the probability of outputting a very frequent item is at least in which is the probability that the Lap distributed noise that is added to the count is at least –ve so that a very frequent item with count at least +ve remains in the output of the algorithm. This probability is at least ½. All in all, it has higher accuracy than the baseline algorithm on all inputs with at least one very frequent item.

4.5 Separation Result

Combining the analysis, we obtain the following separation result between e-differential privacy and probabilistic differential privacy.

Our probabilistic differentially private algorithm *Zealous* is able to retain frequent items with probability at least ½ while filtering out all infrequent items. On the other hand, for any differentially private algorithm that can retain frequent items with nonzero probability (independent of the input. The database and its inaccuracy for large item domains are larger than an algorithm that always outputs an empty set.

TABLE:-

(A) Distinct item counts with different *m*.

M	1	4	8	20	40
Keywords	8956	7845	6554	5423	3251
Queries	4556	3251	2532	1523	742
Clicks	2135	1645	1052	752	532
Query pairs	425	185	121	60	32

(B) Total item counts x 10^3 with different *m*.

M	1	4	8	20	40
Keywords	456	1252	1945	3456	3987
Queries	245	354	456	468	456
Clicks	155	265	284	345	298
Query pairs	10	15	17	11	8

Average number of items per user in the original search details

	keywords	queries	click	Query pairs
avg items/user	63	21	18	9

5. Conclusion

This paper contains a comparative study about publishing frequent keywords, queries, and clicks in search details. We compare the disclosure limitation guarantees and the theoretical and practical utility of various approaches. Our comparison includes earlier work on anonymity and indistinguishability and our proposed solution to achieve probabilistic differential privacy in search details. In our comparison, we revealed interesting, relationships between indistinguishability and probabilistic differential privacy which might be of independent interest. Our results (positive as well as negative) can be applied more generally to the problem of publishing frequent items or item sets.

A topic of future work is the development of algorithms to release useful information about infrequent keywords, queries, and clicks in a search details while preserving user privacy.

References

[1] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar, *"Privacy Accuracy and Consistency Too: A Holistic Solution to Contingency Table Release,"* Proc.ACM SIGMOD-SIGACT-SIGART *symp. Principles of Database Systems (PODS), 2007.*

[2] E. Adar, "User 4xxxxxx9: Anonymizing Query Details," *Proc. World Wide Web (WWW) Workshop Query Details Analysis, 2007*

[2] R. Baeza-Yates, "*Web Usage Mining in Search Engines," Web Minning: Application and Techniques,* Idea Group, 2004.

[3] M. Gotz, A. Machanavajjhala, G. Wang, X. Xiao, and J. Gehrke, "*Privacy in Search Detailss," CoRR, abs/0904.0682v2, 2009.*

[4] J. Han and M. Kamber, *Data Minning: Concepts and Techniques,* first ed. Morgan Kaufmann, Sept. 2000.

[5] R. Jones, B. Rey, O. Madani, and W.Greiner, "*Generating Query Substitutions," Proc. 15th Int'l Conf. World Wide Web (WWW), 2006.*

[6] R. Motwani and S. Nabar, *"Anonymizing Unstructured Data," Corr, abs/0810.5582, 2008.*

[7] K. Nissim, S. Raskhodnikova, and A. Smith, *"Smooth Sensitivity and Sampling in Private Data Analysis," Proc. Ann. ACM Symp. Theory of Computing (STOC), 2007.*