International Journal of Engineering Research and Development
e-ISSN: 2278-067X, **p**-ISSN: 2278-800X, www.ijerd.com
Volume 7, Issue 4 (May 2013), PP.03-08

Search Log Publishing With Improved Utility Using Confess Algorithm

S. Belinsha[1], Mr.A.P.V.Raghavendra[2]
[1]Department of Computer Science and Engineering, VSB Engineering College,Karur
[2]Department of Computer Science and Engineering, VSB Engineering College,Karur

Abstract:- Search engines are being widely used by the web users. The search engine companies are concerned to produce best search results. Search logs are the records which records the interactions between the user and the search engine. Various search patterns, user's behaviors can be analyzed from these logs, which will help to enhance the search results. Publishing these search logs to third party for analysis is a privacy issue. Zealous algorithm of filtering the frequent search items in the search log looses its utility in the course of providing privacy. The proposed confess algorithm extends the work by qualifying the infrequent search items in the log which tends to increase the utility of the search log by preserving the privacy. Confess algorithm involves qualifying the infrequent keywords, URL clicks in the search log and publishing it along with the frequent items.

Keywords:- utility, information service, privacy, search logs, search behaviour, infrequent items, search items, threshold

I. INTRODUCTION

Web stores large amount of information. The information is retrieved by means of various techniques which termed as web mining. Privacy preservation is the hot topic in the world of web. Web mining involves web structure mining, web content mining and web usage mining. Analyzing and studying the search log falls under the category of web usage mining. Search logs are confined with the search engines. Search engines are the applications which support users to browse the web in an efficient way. Nowadays web users are more dependent on search engines to access the web. The search engine companies bend more to produce best search results to the users. Search logs are the record of interactions between the users and the search engine. It holds the data like the user id, search keywords, URL (Uniform Resource Locator) clicks, date and time of search. Publishing search log can be done in two ways: Providing the log to the third party and deploying the log for the search engine functions. The information in the log supports the analysis of the user's search behaviors and patterns which helps to enhance the search results. Analyzing the search log is performed by the research community. When these logs are provided to third party it should provide a privacy guarantee to the users of the search logs. When privacy is focused, the utility i.e. the number of items released in the log, is decreased as it involves elimination of more records.

When the AOL (American On Line) search log release is concerned, the log was released with replacing the user id with the random numbers[1]. The privacy factor compromised with more utility. Hence there is always a trade of between privacy and utility. So holding back the user's identity alone does not guarantee privacy. The user's identity can also be revealed by the formation of queries and the link followed by the user. The keywords also may involve the sensitive information like social security number, credit card number and also certain demographic information. Hence the focus has to be made on these items and strong strategies have to be followed to release the keywords formed by the users in the search log.

Earlier work involved the release of the logs with replacing the user identity with random numbers [2]. But this was not promising one because of less privacy concern and is prone to background linkage attack [1]. Also by the keywords formed by the users, the user's identity can be revealed. The AOL search log release stands as an example to this case. It released the search logs of several users by replacing with the random id but the public were able to identify certain users, by their formation of queries. Later the work was extended to anonymization [2], where the similar items were grouped and released. Achieving k-anonymity, l-diversity are some of the privacy preserving techniques used. The dilemma in those techniques was that it was prone to background knowledge attack [1]. The crucial effect produced as a result was that it lost the uniqueness of the user's search. The same effect was the case in generalization techniques also.

Zealous algorithm [1] was proposed to release the frequent items in the log by two threshold framework. The frequent queries are more privacy promising. A keyword may become frequent when it is a common public interest and when it is published it provides the less chance of identifying the user. Publishing

the frequent items alone will not contribute to the utility of the log further certain infrequent items also must be considered. In practical, the search log may contain less frequent items than several infrequent items. The infrequent items may have more probability of identifying an user. But there exists some infrequent queries which are of public interest and relevant to the frequent query. Hence the confess algorithm tries to find out such keywords and their corresponding URL click values and publishes it in the search log. To qualify the infrequent keywords and URL clicks in the log, separate qualifying strategies are needed to be formulated. Hence different qualifying constraints are set to qualify the keywords and the URL clicks.

The confess log obtained is applied to serve the search engine functions such as providing query suggestion, query substitution. With the results the performance is studied and evaluations are made. The confess log publishing strategy is also applied to the search engines and the effectiveness was studied in comparison with the zealous algorithm. The zealous and the confess log were compared in terms of the average number of items published in log.

II. ZEALOUS ALGORITHM

The Zealous algorithm uses a two phase framework to discover the frequent items in the log and finally publishes it. To discover the frequent items, the Zealous algorithm uses two threshold values. The first threshold value is set based on the number of user contributions in the log. The Laplacian noise is added to the first set threshold value and the items are filtered by the set values. The addition of noise is to divert the attackers and produce a non-exact statistics [4]. By this method of finding the frequent items, the result log achieves probabilistic differential privacy. The main objective of Zealous algorithm is to figure out the frequent items in the log. The Zealous algorithm is applied to a sample search log collected from a local search engine to the items in the log like keywords and URL values. The log contained more than 200 entries with 58 users. The Zealous algorithm was applied to the log with the threshold values in the table.

Keyword	Count
Sport stores 2009	31
Opera browser in mobiles	42
Laptop models	53
Antivirus software for windows	45
New theme music	63
Exam results	28
Projects	32

Table I: Keyword log of Zealous

The above are the keywords which have passed the filtration of the two phase framework. These keywords are identified as frequent keywords. Similarly it identifies the frequent URL clicks in the log by the two threshold values.

URL clicks	Count
http://esupport.trendmicro.com	17
https://blogs.oracle.com	21
http://en.wikipedia.org	24
http://www.entrance-exam.net	19
https://www.mcafeeasap.com	22
http://technet.microsoft.com	25
http://www.whatis.com	39

Table II : URL log of Zealous

However, Zealous algorithm leaves out the infrequent keywords in the log. However setting upon the threshold value is a challenging task. But in a search log, there will be several infrequent items. The infrequent item which has no possibility of revealing an user's identity has to be identified and it has to be published. Hence confess is proposed to qualify such infrequent items in the log.

III. CONFESS ALGORITHM

The confess algorithm follows the Zealous algorithm to trace out the frequent items. It isolates the frequent and the infrequent items and the further processing is done to qualify the infrequent items. The Zealous algorithm uses a two phase threshold framework to identify the frequent items. The infrequent items are then retrieved from the log and the following constraints are checked against the items like keyword and URL clicks. The two items considered to be qualified are the keywords and the URL click as they bind more user's information.

A. Qualifying the keyword

The keywords are the prime input of the user through which the user explores his needs in the web. The keywords formed by the user reveal more private information about the users. This will be a gold mine for the researchers to know the user's identity. So several strategies are formulated to qualify the keywords that are privacy promising [5].

*1)Profile information***:** The users are registered before performing the search. The users have to provide certain mandatory information for the registration. The infrequent queries are initially checked with the profile information to check whether it contain any sensitive data. If so, then they are not used for further processing.
Consider the keyword 07480433 of a user. This keyword contains the social security number, which is likely to reveal the identity of an user. This is identified by comparing the items with the profile information registered by the user. In case, the keyword contains the profile information given by the user, then the keyword is not qualified. In this way, if the keyword contains the information like name, date of birth, phone numbers, social security numbers, address information, they can be identified and prohibited from publishing.

*2) Sub keyword checking***:** The keywords formed by different users are different and holds user's uniqueness. The infrequent keyword is compared with the frequent keyword to find there is any sub keyword. If any such sub keyword is found in the infrequent keyword, then the keyword is qualified.
Consider the keyword "lecture notes about search logs" is the frequent keyword as discovered by the Zealous algorithm. The keyword "about search logs" is an infrequent keyword. But it is a sub keyword of the frequent item. If such infrequent item exists then those keywords are qualified to be published. This may improve the addition of useful entries in the log.

B. Qualifying the URL clicks

URL are the data which helps to know the location of a resource in the web. The URL clicks are the important item in the log, which points out the user's visiting of the web pages. The keywords and URL clicks together can lead to identifying an user. Hence certain constraints are set to qualify the URL clicks.

1) URL shortening: The URL(Uniform Resource Locator) reveals the location of a resource in the web environment. Normally an URL contains the fields like protocol, authority, filename, host, path, port. The complete URL of an user click is likely to reveal the user's identity and hence the attributes like filename, path are removed. This procedure would conceal the exact visit of the user.
Consider the URL click,

https://developer.cebv.in/search-appliance/document /50/help_mini/status_log, is shortened as "https://develepor.cbev.in. These shortening of the URL provide a less information about the page visited. Sometimes revealing the complete URL value would identify an user. This is done to preserve the privacy.

2)Multiple visit to same URL: A user obtains several search results for the keyword provided for searching. The user chooses the link appropriate to his search intension. The several links chosen by the user may point to the same URL. This reveals that the user finds the information in that page which satisfies their need.
Consider the keyword, exam results in the log. The URL clicked by the user from the search results are,
http://www.results.in/colleges/BEresults.html
http://www.results.in/colleges/MCAres.html
http://www.results.in/colleges/MEresutlts.html
http://www.results.in/colloges/MBAres.html

The above clicks of the user reveal that he finds the intended content on the web page http://www.chennairesults.in.The mentioned URL of the page is then qualified and is included in the published log. When multiple link pointing an URL is listed in the search engine showcase that it is a prevalent page which is offering more beneficial information regarding the input keyword and hence it can also be privacy promising.

*3)The URL with the keyword***:** The user searches by the keyword and obtains search results. Probably the URL chosen by the user may contain the keyword as its sub term. This denotes that it was a relevant click by the user. Such URLs can be included in the published log.
Consider the keyword, exam results is in the search log. The URL clicked by the user is http://www.examinfo.in then this URL is added in the published log. The URL containing the keywords which is chosen by the user, i.e. the entry in the log, showcase that the web page is of common interest. This highly depends on the user's way of providing the keyword and following the links in the result.

4)URL of top ranked pages: The selection of the link or the page of the user for a keyword from the search results may be due to various intensions. When the clicked page is one of the top ranked page, then the URL of the page can be published. The frequently visited page of an user is also considered to be published in the log. The top ranked pages are safe enough to be published in the log[1].

By the above constraints, the infrequent URL clicks and keywords of the users are qualified and published in the log which intends to improve the utility of the published log. The confess algorithm is applied to the keywords and the URL clicks of the several users in the search log.

IV. RESULTS

The following tables depicts the results produced by the confess algorithm on the search log which was used up by zealous algorithm.

Keyword	Count
Sport stores 2009	31
Opera browser in mobiles	42
Laptop models	53
Antivirus software for windows	45
New theme music	63
Exam results	28
Projects	32
Milk chocolates	33
Laptop models	33
Chocolates	12
Antivirus software	4
Antivirus software	20
Results	1
Theme music	7
Sports	2

Table III : Keyword log of Confess

The above is the keyword log produced as the result of applying confess algorithm of finding the infrequent items. It can be noted that the keywords which are qualified is the part of the frequent keyword. Releasing such keyword, would improve the utility as the log will contain more entries when published.

URL clicks	Count
http://esupport.trendmicro.com	17
https://blogs.oracle.com	21
http://en.sportstore.org	24
http://www.entrance-exam.net	19
https://www.mcafeeasap.com	22
http://technet.microsoft.com	25
http://www.whatis.com	39
https://www.docstoc.com	11
https://blogs.project.com	7
http://en.mcs-college.org	3
http://www.exam.net	4
https://www.webmasterworld.com	1
http://technet.puzzles.com	2
http://www.musics.net	6

Table IV: URL log of Confess

The above log produces the qualified infrequent URL clicks along with the frequent URLs. After qualifying the items in the search log i.e. keywords and URL clicks, they are compared with the entries in the search log. The entries with the qualified keyword, URL click, date and time of the users.

U	K	U	T
U42	Result	http://www.results.in	22/10/2012 2:25:00
U42	Result	http://www.exam.net	22/10/2012 2:27:36
U42	Exam result	http://www.webmasterworld.com	22/10/2012 2:30:09
U34	Sports	http://sportstore.org	22/10/2012 2:45:12

Table V : Portion of the search log after qualification of the items

The above log is the portion of the search log after qualification. The log contains User-id(U),Keyword(K),URL-click(U) and the Timestamp(T). The log retains the user's id to carry the uniqueness of the each users in the log. If user's id is eliminated it would loose various session information because the user's uniqueness will not be obvious

V. COMPARATIVE STUDY

The performance of the confess algorithm is analyzed through various parameters like response time, average number of items published in the log. Then the proposed confess algorithm is compared with the zealous algorithm to swot up the performance in terms of utility produced by the log.
The below statistics show the average number of keywords published in the zealous log and the confess log. The average number of keyword(N_k) is the ratio of the number of items released in the log to the total number of items in the original log. To perform this study various experimental search logs are considered.

Trails	Zealous log - N_k	Confess log - N_k
1	0.03	0.192
2	0.32	0.130
3	0.02	0.07
4	0.189	0.193
5	0.321	0.381

Table - 1.6 : Comparison with average number of keywords

With the above statistics the graph is generated as below.

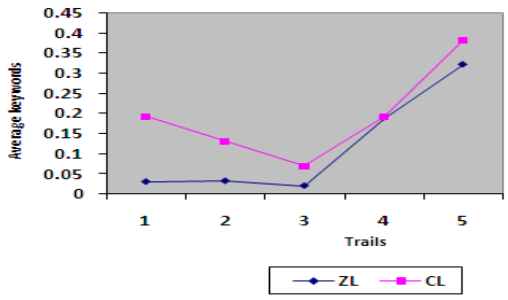

Figure - 1.1 : Comparison with average number of keywords

It can be inferred that the confess keyword log outputs more keywords when compared to zealous logs and at some instance, the average keywords produced is almost equal. This is highly probabilistic because it depends on the user's intention of forming keywords.

The below statistics show the average number of keywords published in the zealous log and the confess log. The average number of url-click(N_u) in the log is the ratio of the number of items in the published log to the number of items in the original unprocessed log. This metric considered for the study.

Trails	Zealous log - N_u	Confess log - N_u
1	0.116	0.35
2	0.0	0.26
3	0.112	0.3175
4	0.126	0.353
5	0.158	0.254

Table – 1.7 : Comparison with average number of URL clicks

With these statistical data a graph is generated below.

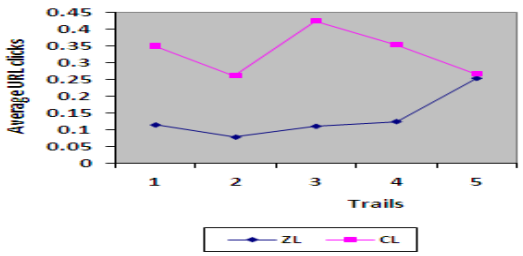

Figure - 1.2 : Comparison with average number of keywords

It can be inferred that the confess log also outputs more URL clicks than zealous log, that which are maintaining the privacy of the user. It can be noticed that the URL log produces more utility than the keyword log, as it qualifies more URL clicks.

From the above studies, it can be inferred that qualifying infrequent items in the log would enhance the utility of the published log. The resultant log can be deployed to support various search engine functions which would reduce the time complexity in the usage of the log when compared to the original unprocessed search log.

VI. APPLICATIONS

As confess log produces more utility in the published log, the log can be applied for various search engine functions like index caching, query substitution, query suggestions. These activities must be processed quickly to give better search experience for the users. The time consumption reduces when confess log is consumed rather than the original log. The utility of the log will be increased than that of the Zealous log, and helps to achieve privacy also.

VII. CONCLUSION

By the above studies, it can be inferred that the average number of items released is more. Hence the utility of the search log is improved by including the qualified infrequent items from the log. Also publishing those infrequent items will not disturb the privacy of the users as it has to satisfy various constraints which are privacy promising.

VIII. FUTURE ENHANCEMENT

Several better qualifying criteria can be set to qualifying the infrequent keywords and URL clicks. Also the work can be extended in setting constraints for the unregistered in the search engines whose However challenges still lies in discovering the frequent items in search logs. Efficient method to discover the frequent items can also be formulated.

REFERENCES

[1]. Michaela Gotz, Ashwin Machanavajjnala, Guozhang Wang, Xiaokui Xiao and Johannes Gehreke, ”*Publishing search logs – A comparative study of privacy guarentees*”, IEEE transactions on knowledge and data engineering,Vol.24, No.3, March 2012.

[2]. E. Adar, “*User 4xxxxx9: Anonymizing Query Logs*” Proc. World Wide Web (WWW) Workshop Query Log Analysis, 2007.

[3]. A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas, “Releasing *Search Queries and Clicks Privately*,” Proc. 18th Int’l Conf. World Wide Web (WWW), 2009

[4]. C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, “*Our Data Ourselves: Privacy via Distributed Noise Generation*” Proc. Ann. Int’l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT), 2006.

[5]. V. S. Iyengar, “*Transforming data to satisfy privacy constraints*” in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.