Privacy Preservation and Enhanced Utility in Search Log Publishing

S. BELINSHA
Department of Computer Science and Engineering
VSB Engineering College
Karur
India
s.belinsha@gmail.com

A. P. V. RAGHAVENDRA
Department of Computer Science and Engineering
VSB Engineering College
Karur
India
raghu221084@gmail.com

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 behaviours 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.

Key-Words: - information service, privacy, infrequent items, threshold

1 Introduction

Web stores large amount of information. The information is retrieved by means of various techniques which is termed as web mining. Web mining is the application of data mining that helps to analyze pattern from the web. Web mining involves web structure mining, web content mining and web usage mining according to analysis targets. Web content mining is the process of extracting the data of interest from the pool of documents in web. Web structure mining is the process of discovery and analyzing the structure of web documents. Analyzing, studying and extracting the user history falls under the category of web usage mining [1]. User history in search engine is maintained in the search logs. 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, and other useful information. However, the user-id, keywords, URL, timestamp are considered in the study. As the web users are increasing in today's world, privacy preservation has become the hot topic. Presence of the user identity in a search log, may reveal a user's private information in the search results produced. Also, preserving privacy is more important in publishing the search log. Privacy becomes an issue when one has to use the data which contain sensitive information of other users. Search log publishing is not periodically done in search engine companies because of the privacy issues. It is the responsibility of search engine authority to hold their customer's private information. Publishing search log makes two senses: One is providing the log to the third party and the other is deploying the log for the search engine functions. The information in the log supports the analysis of the user's search behaviours and patterns which helps to enhance the search results. Also the amount of successful and unsuccessful searches can be identified which helps to improve the performance of the search. 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. There are more ways of providing the privacy to the data like attribute disclosure, data swapping, generalization, anonymization, and diversification. Among these, anonymization, diversification, attribute disclosure leads to loss of data in the log [3][5]. When privacy is focused more, the utility i.e. the number of items released in the log, is decreased as it involves elimination of more records.

In 2006, AOL (American On Line) took an attempt to release the search log of thousands of its users to the mass. When the AOL search log release is concerned, the log was released with replacing the user identity with the random number[2]. But just few minutes after the release the identity of the users are revealed. So the authority who have released the log went for an apology. So, replacing of user identity results in the release of all the entries leading to utility. But the privacy factor is compromised. Hence there is always a trade off 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 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. So the contribution is to set various privacy satisfying constraints to qualify the items in the log specifically keywords and URL clicks while publishing.

2 Related work

Earlier work involved the release of the logs with replacing the user identity with random numbers [3]. But this was not promising one because of less privacy concern and is prone to linkage attack [6]. Linkage attack involves finding the user from the random id, by linking with another database. Also by the keywords formed by the users, the user's identity can be revealed. Later the work was extended to anonymization [1][8], where the similar records were grouped together 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 [6] . 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 [4] also. Even though these techniques brought about privacy but they greatly reduced in utility, since there are more dissimilar entries in the log. Zealous algorithm [2] was proposed to release the frequent items in the log by two threshold framework. The frequent queries are more privacy promising. A keyword will become frequent when it is of common public interest. Releasing a frequent keyword 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. There are various methods to find the frequent items like counter based algorithm, quantile algorithm, and sketch algorithm. Among these algorithm, for search log based data, the two phase threshold framework works better as it tends to preserve privacy. 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.

3 Utility in search logs

The utility in search log refers to the degree to which a search log retains the useful information in the search log, which is privacy preserving. In a search log more entries would be privacy preserving. So certain criteria must be set to find out the privacy promising items like keyword, URL clicks and it has to be released. Privacy preserving method of generalization does not provide utility as it involves more loss of data in the published log. The same is the case identified with several anonymization techniques [9]. Zealous algorithm focuses on finding such items, but however only frequent items are released. The infrequent items are

not considered in zealous algorithm. The proposed confess algorithm tries to figure out those infrequent items and based on certain qualifying constraint, the infrequent items are qualified. Mathematically, utility can be measured by the average number of items released in the log. It is the ratio between the number of items released to the number of items in the original log. Hence by qualifying the infrequent items in the log, the utility of the published log can be improved by confess algorithm.

4 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 [4]. The addition of noise is to divert the attackers and produce a non-exact statistics [6]. 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. The input log is provided to the Zealous algorithm. For each user, the top distinct items are chosen from the search logs. Based on the selected items histogram is constructed with the pair of items and the number of occurrences of the items. This is stored as the original histogram. A threshold value is set based on each user's contribution [8]. User contribution denotes the number of keywords and the URL followed by an user. The histogram below this count is deleted. A noise value is determined from the Laplacian distribution from the first threshold and user contribution. The noise is added to the first set threshold value and the second threshold is obtained. Delete from the histogram for which the count value is smaller than the second threshold value. Those items are the frequent items. Hence these frequent items are published. This method of finding the frequent item results in the establishment of differential privacy.

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 1: 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 2: 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.

5 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 users' information.

5.1 Qualifying the keyword

The keywords are the prime input of the user through which the user explores his needs in the web. Generally, the keywords can be informational, navigational and transactional. But this strategy is not considered for the qualification of the keywords. 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.[11]

5.1.1 Profile Information

The users are registered before performing the search. The users have to provide certain mandatory information for the registration. The possible data gathered can be some of the personal details like name, date of birth, zip code. The infrequent queries are initially checked with the profile information to check whether it contains any sensitive data mentioned in the profile. The user identifying data like name, social security number is mostly considered sensitive. If any sensitive information is found in the log, then they are not qualified and also hence it is not used for further processing. However, the user id is randomly assigned with another id and stored in the log.[11]

5.1.2 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 sub keyword of the infrequent keyword is found in the frequent keyword , then that infrequent 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 "search logs" is the sub keyword of the frequent keyword mentioned above. So "about search logs" is qualified to be published in the log. If such infrequent item exists then those keywords are qualified to be published. The keyword log and the URL log of confess is combined to form a complete confess log. This method of sub keyword checking will improve the addition of useful entries in the log related to similar keywords.

5.2 Qualifying the URL clicks

URL (Uniform Resource Locator) are the data which helps to identify the location of a resource in the web. The documents or the data in the web are accessed through the URL. So, the URL clicks are the important item in the log, which points out the user's visiting of the web pages. The URL helps to know the user's real intention of performing a search in the search engine. The keywords and URL clicks together can lead to identifying an user, by matching up with several keywords and clicks. Hence certain constraints are set to qualify the URL clicks.

5.2.1 URL shortening

The URL 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.html

In the above URL value, https is the protocol value, developer.cebv.in is the authority and host, file name is status log.html. It is shortened as "https://develepor.cbev.in", considering the protocol, host and the authority of the URL value. These shortening of the URL provide a less information about the page visited. Sometimes revealing the complete URL value would identify an user as the log is made open. So when publishing the log, all the entries are replaced with the shortened URL value to preserve privacy.

5.2.2 Frequent visit to a page

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.[10]

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
http://www.results.in/colloges/BTechres.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.

5.2.3 URL with keyword

The user searches by the keyword and obtains the 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 [10]. This highly depends on the user's way of providing the keyword and following the links in the result.

5.2.4 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 pages, then the URL of the page can be published. The top ranks are calculated from the zealous log. Also, the count of the frequent URL clicks with respect to an single user is also calculated. With these both counts, and by the position of the URL values in the search results the rank is found. We consider the top three ranked pages from the calculated rank. 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. 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.

6 Results

By using the above strategies the confess keyword log and the confess URL log are obtained. The following tables depicts the results produced by the confess algorithm on the search log which was used up by zealous algorithm. The confess algorithm consumes zealous algorithm to find out the frequent and infrequent items. The infrequent items are sustained in the log, instead of being deleted.

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 3: 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 inferred that some of the infrequent items are present in the log. The keywords which contained the profile information of the users are removed from the log first and then confess steps are applied. It can be noted that the infrequent 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 4: 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 are with the user id 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 2:45:2

Table 5: 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 user in the log. The user id which the user provides at the time of registration is not used while publishing a log. The user id is assigned with a random id, to be published in the log. If user's id is eliminated it would loose various session information because the user's uniqueness will not be obvious. So the user identity in the log is sustained but it is replaced with a random identity number.

7 Comparative study

The performance of the confess algorithm is analyzed through the parameter called the 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.

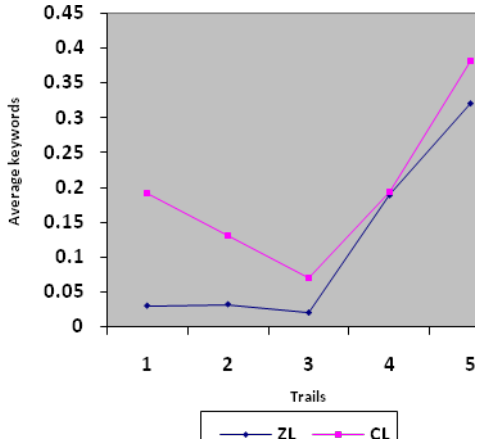
The below statistics show the average number of keywords published in the zealous log and the

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 calculated parameters and statistics the graph is generated as below.



ZL – Zealous log

CL – Confess log

Fig. 1: Comparison with average number of keywords

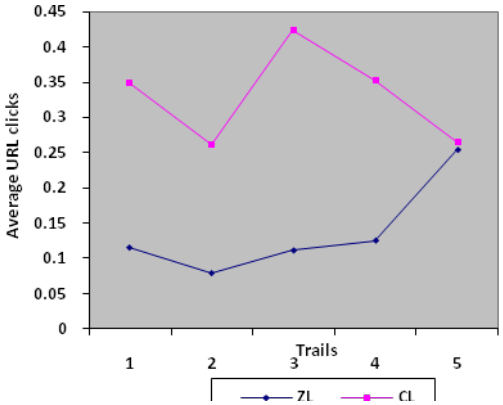
It can be inferred that the confess keyword log outputs more keywords when compared to zealous keyword log and at some instance, the average keywords produced is almost equal. The result obtained is highly probabilistic because it depends on the user's intention of forming keywords. Frequent and infrequent keywords depend on the user's need to from keywords to perform their search. So the degree up to which utility can be enhance through confess depends on the search log of the user.

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 7 : Comparison with the average number of URL clicks

With these statistical data a graph is generated below.



ZL – Zealous log

CL – Confess log

Fig. 2: Comparison with the average number of keywords

It can be inferred that the confess log also outputs more URL clicks than zealous log, which are maintaining the privacy of the user. The occurrence of the frequent and the infrequent URL clicks highly depends on the user's intention to follow the URL for a keyword provided for searching. So the degree to which the URL log of confess is enhanced is probabilistic. However, it can be noticed that the URL log produces more utility than the keyword log, as it qualifies more URL clicks. An user may follow various URL click for a keyword in one session itself. So the URL entries will be more. There is more probability that more URL will be qualified than the keywords in the log. From the above studies, it can be inferred that qualifying infrequent items in the log would enhance the utility of the published log. The enhancement of utility is also supported with privacy preservation of the users. 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 where in the original log does not provide any privacy to the user performing the search.

8 Enhanced Utility

Search logs contain large repositories of user's search data. The processed search log must filter the privacy preserving data while publishing. In the course of providing privacy the log may lose it's utility. So focus is made to include more search information to improve the items released in the log. The proposed mechanism of qualifying the infrequent keyword and URL clicks preserves the privacy as well as increases the number of items released in the log. There is a increase in the average number of keywords and URL clicks released in the log which have satisfied the privacy constraint. Also it does not disclose any attributes while the final log is published. The degree to which the utility is enhanced is highly probabilistic as the frequency of the item depends on the user's search intention. Further, the resultant log is deployed in the search engine to study the performance of the published log to serve various search engine functions.

9 Application

Confess algorithm can be used in the search websites concerned with medical and banking applications. When a hospital website is considered, the users may look for the doctor's appointments, and to know other details. It is observed that the occurrences of the infrequent item are more in this application. The similar inference is obtained from the banking website, where an user searches about various bank information. Further, privacy is an important feature in the banking applications, where the user's private information is more sensitive. 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.

9.1 Query suggestion

Query suggestion is the common function in search engines, where when a keyword is provided by the user, the search engine finds out various related keyword from its log and provides it to the user[7] .The amount of related keywords retrieved from the original log is more but the time taken to perform this query suggestion is more. Hence the confess log can be utilized, which performs the query suggestion more similar to that of the original log but with less time. An advantage is that the keywords which are privacy preserving alone is made open to all the users.

9.2 Query substitution

Query substitution is the process of substituting the query with the relevant keyword when the keyword provided by the user has incorrect spelling, unfound keywords. The formation of keywords from the set of sub keywords is important here. The possible combination of the various sub keywords is more when compared with the zealous log. Hence it provides better query substitutions to the users. The confess log provides less time in doing this function when compared with the deploying of the original log, also preserving privacy. Further the confess log can be used in query expansion to enhance the search results.

9.3 Index caching

The index caching is an activity performed in the search engine, to improve the speed of the retrieval of the search results. When a keyword is provided by the user, the search results are obtained. The user clicks a link to obtain the necessary information. Some set of index value of the URL clicks are

cached, to provide a better search experience to the users. The URL clicks found are less when compared to the original log and thus it can be properly indexed. The confess algorithm provides the way to cache the items that are privacy promising.

10 Conclusion

Search log contains the up-to date information about the user's search behaviour while interacting with the search engine through web browser. Search logs helps in improving the various scenarios like web search ranking, personalize web search and correct search query spellings[13]. The search log published is undergone various analysis to improve the above services in the search engine. Search logs are huge and increase rapidly with time. The search log data is considered to publish within a period of time. The work involved various steps to enhance the utility in the search log publishing. First, the frequent items are discovered by the zealous algorithm which follows a two phase threshold framework in which the frequent items are found by filtering with two threshold values. Second, the infrequent items which are deleted by zealous algorithm are recovered and it is qualified against various privacy constraints to meet the privacy requirements. The privacy constraints are set separately for the keywords and the URL clicks. By the above studies, it can be inferred that the average number of items released is more in confess log. The utility of the URL log is increased more than that of the keyword log of confess. 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. The privacy constraint which is set also provides an effective means of identifying the riskless search items from the log. Further, stronger constraint can be set and the constraint can be reduced in order to achieve efficiency.

11 Future Enhancement

The simple notion of the privacy preservation technique is that when data provided to the third party, it should not reveal any sensitive information of the user. This work involved various efforts to improve the utility in the search logs by preserving the privacy. So few qualifying constraints are set to achieve it. When the qualifying constraints are considered, they are more specific for the users who have registered so that their identity becomes obvious. For the users who have not registered the identity will be difficult to spot out. IP address, cookie can be logged, but the criteria of confess algorithm does not work with it. So the work can be modified in setting constraints for the users unregistered in the search engines. So, several better qualifying criteria can be set to qualifying the infrequent keywords and URL clicks. To improve efficiency in terms of response time, the constraints can be reduced, and made stronger to achieve privacy. An algorithm can be formulated to enhance utility in data publishing other than search log data as the future work. However challenges still lies in discovering the frequent items in search logs[12]. Methods can be framed to set a proper threshold value to accurately distinguish between the frequent and the infrequent items in the log. Threshold value can be calculated by more strong parameters, because setting of threshold value is tougher in the environment with more number of users. The processed log can be deployed in various search engines instead of the unprocessed log, so that the user's privacy is maintained. Further, the log can be made available to other users to obtain beneficial information from the log during their search[14].

References

[1]Singh,B.; Singh,H.K. Web Data Mining research survey, *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2010 IEEE*

[2] Kagal, L., & Pato, J. (2010). Preserving privacy based on semantic policy tools. *Security & Privacy, IEEE, 25-30, 2010*

[3] A.Korovola, K.Kenthapadi, N.Mishra, & 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(2007). Data Privacy, Our Data, Ourselves: Privacy via Distributed Noise Generation *Proc. Ann. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT),* 2006

[5] E. Adar. User 4xxxxx9: Anonymizing Query Logs. *Proc. World Wide Web (WWW) Query Analysis, 2007*

[6] R. Agrawal and R. Srikant, Privacy-Preserving Data Mining. *Proc. ACM SIGMOD Conf. Management of Data, ACM Press, 2000, pp. 439–450*

[7] Liu, J., Zhu, L., & Wang, C.(2011). Query log mining on query suggestion. *2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011, PP. 2268 -227.*

[8] Michaela Gotz, Ashwin Machanavajjnala, Guozhang Wang, Xiaokui Xiao and Johannes Gehreke, Publishing search logs – A comparative study of privacy guarantees: *IEEE transactions on knowledge and data engineering,* 2012, 520 – 532.

[9] Mimi, A., & Bahloul, S., Protect user anonymity in Query log. *Machine and Web Intelligence (ICMWI), 2010 International conference*, 2010, *(pp. 421 – 425).*

[10] Park, K., Lee, T., Jee, H., Chang, J., Jung,S., & Lim,H .*Mining User Intention from Web search Query log*. *APIC_IST & ICONNI,* 2009, (pp.229-236).

[11]Belinsha, S., Raghavendra, A.P.V. Privacy Preservation and Enhanced Utility in Search Log Publishing using Improved Zealous Algorithm. *ICGHPC 2013,* (pp.1-5), *2013 IEEE*

[12]Sharma,K. ; Shrivastava,G. ; Kumar,V.*Web mi ning: Today and tomorrow*. Electronics Computer Technology (ICECT), 2011 3rd International Conference

[13] Qingtian Han ; Xiaoyan Gao ; Wenguo Wu, Study on Web Mining Algorithm based on Usage Mining *Computer-Aided Industrial Design and Conceptual Design,9th International Conference on 2008 , Page(s): 1121 - 1124*

[14] Ramakrishna, M.T. ; Gowdar, L.K. ; Havanur, M.S. ;Swamy, B.P.M. *Web Mining: Key Accomplishments, Applications and Future Directions Data Storage and Data Engineering (DSDE), 2010 International Conference*