ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

Securing Personalised Web Search

Akshata Surana[1], Bhagyashree Ghodekar[2], Bhawana Purandare[3], A. G.Dongre[4]

U.G. Student, Department of Computer Engineering, PVG's College of Engineering and Technology, Pune, India[1, 2, 3]

Assistant Professor, Department of Computer Engineering, PVG's College of Engineering and Technology, Pune, India[4]

ABSTRACT: Personalized search refers to search experiences that are tailored specifically to an individual's interests by incorporating information about the individual beyond specific query provided. Personalized Web Search (PWS) has found many applications in day-to-day lives of users. With the increasing use of personalisation in web search, the risk of exposing users' personal information is also increasing. In this paper, text mining techniques are used for performing various operations on the data for improving the search quality of the application. On the other hand, hashing technique is used to preserve the privacy of the user's search query. SHA hash function is implemented to hide the content of users' search query at the server end.

KEYWORDS:personalised web search, text mining, hashing, privacy, profile.

I. INTRODUCTION

Web search engines play a major role in information exchange nowadays. Personalisation in web search has brought a revolution in the information retrieval process by building user profiles according to user interests and history. It is important to improve the search quality with the personalisation utility of the user profile as well as to hide the privacy contents existing in the user query to place the privacy risk under control [2]. There are two methods to implement personalisation in web search namely, click-log based and profile based. The limitation of the click-log based method is that it works only on the repeated queries from the same user [2]. Whereas, in contrary, in the profile based method, there is increasing usage of personal and behaviour information to profile its users, which is usually gathered implicitly from query history, browsing history, click-through data, bookmarks, user documents, and so forth.

While many search engines take advantage of information about people in general, or about specific groups of people, personalized search depends on a user profile that is unique to the individual. Research systems that personalize search results model their users in different ways. Some rely on users explicitly specifying their interests or characteristics. But user supplied information can be hard to collect and keep up to date. Others have built implicit user models based on content the user has read or their history of interaction with Web pages.

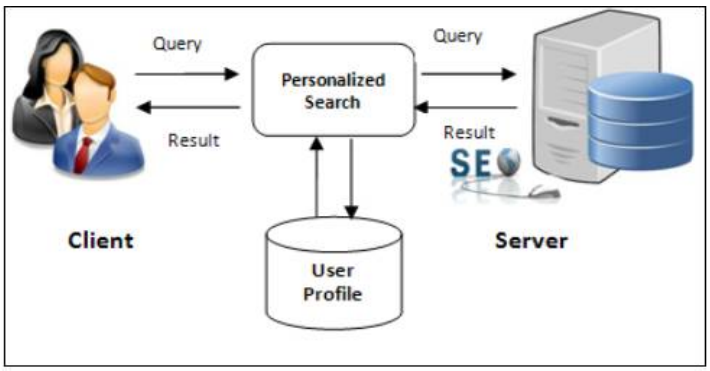

Figure 1: Personalisation in Web Search

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

Generally there are two classes of privacy protection problems for PWS. One class includes those treat privacy as the identification of an individual, as described in [5]. The other includes those considering the sensitivity of the data exposed to the PWS server.

II. RELATED WORK

A concept called generalised profile creation explains the process of extracting only a subtree of the user interest hierarchy and sending it to the server in order to get the results. Sensitivity and risk factors are calculated for each query to perform runtime generalisation[1]. Existing algorithms for subtree creation are as follows,

1. Brute-Force- The brute-force algorithm exhausts all possible rooted subtrees of a given seed profile to find the optimal generalization.The privacy requirements are respected during the exhaustion. The subtree with the optimal utility is chosen as the result.The exponential computational complexity of brute-force algorithm is still unacceptable.

2. GreedyDP- The first greedy algorithm GreedyDP works in a bottom-up manner. In every i^{th} iteration, GreedyDP chooses a leaf topic 't' for pruning, trying to maximize the utility of the output of the current iteration. During the iterations, we also maintain a best profile- so-far, which indicates the profile having the highest discriminating power while satisfying the -risk constraint. The iterative process terminates when the profile is generalized to a root-topic. The best-profile-so-far will be the final result (G) of the algorithm.

3. The GreedyIL-TheGreedyIL algorithm improves the efficiency of the generalization using heuristics based on several findings. In general, GreedyIL traces the information loss instead of the discriminating power. This saves a lot of computational cost. The heuristics are used to avoid unnecessary iterations, simplification of the computation of IL and reducing the need for IL-recomputation significantly.

III. PROPOSED SYSTEM

Text mining and hashing techniques along with profile based personalisation can be used in an efficient way to implement the classic client – server architecture placing the privacy risk under control.
In the proposed system, client-server architecture is maintained. On the server side, all the text mining operations are performed on the document while storing it in the database. These operations are performed on the data in order to accelerate the runtime results retrieval.

Text mining techniques:

1. *Parsing:*

Parsing or syntactic analysis is the process of analysing a string of symbols, either in natural language or in computer languages, conforming to the rules of a formal grammar. A parser is a software component that takes input data (frequently text) and builds a data structure – often some kind of parse tree, abstract syntax tree or other hierarchical structure – giving a structural representation of the input, checking for correct syntax in the process.

2. *Tokenizing:*

In lexical analysis, tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. It is the process of replacing sensitive data with unique identification symbols that retain all the essential information about the data without compromising its security.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

3. *Filtering:*

A Web filter is a program that can screen an incoming Web page to determine whether some or all of it should not be displayed to the user.The filter checks the origin or content of a Web page against a set of rules provided by company or person who has installed the Web filter.

4. *Stemming:*

Stemming is the term used in linguistic morphology and information retrieval to describe the process for reducing inflected (or sometimes derived) words to their word stem, base or root form-generally a written word form. e.g. A stemming algorithm reduces the words "fishing", "fished", and "fisher" to the root word, "fish".

5. *TF-IDF(Term Frequency-Inverse Document Frequency):*

TF-IDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining.

$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

With,

- N: total number of documents in the corpus $N=|D|$
- $|\{d \in D : t \in d\}|$ number of documents where the term t appears (i.e. $\mathrm{tf}(t,d) \neq 0$). If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to adjust the denominator to $1 + |\{d \in D : t \in d\}|$

Using all the above techniques and by re-ranking the words according to their importance, important keywords from all the documents are fetched and their hash value is calculated using SHA hash function and these values are again stored inside a different text document.

A cryptographic hash function is a hash function which is considered practically impossible to invert, that is, to recreate the input data from its hash value alone. The input data is often called the message, and the hash value is often called the message digest or simply the digest. The four main characteristics of the ideal cryptographic hash function are as follows,

1. It is quick to compute the hash value for any given message
2. It is infeasible to generate a message from its hash
3. It is infeasible to modify a message without changing the hash
4. It is infeasible to find two different messages with the same hash.

On the server side, all the users' profiles and their individual interests are also stored. These interests are derived from user logs and history.
Whenever a client fires a query, he will be able to choose between public and private modes according to the sensitivity of the query. If he chooses the public mode, the system works as a normal search engine. Similarity measures will be applied and according to users' interests result will be re-ranked and result will be sent back to the client. If the client chooses the private mode, the hash of the query will be calculated using SHA function and it will be sent to the server.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

At the server end, similarity between tags of database from the same category will be found and documents containing those tags will be fetched. These fetched documents will be re-ranked based on individual's interests. While sending these documents back to the client, encryption will be used.

IV. EXPERIMENTAL RESULTS

Text mining and hash algorithms can be used to preserve the privacy of the user as well as to maintain the performance of the search. Figures show the results of securing personalised web search application using hashing technique. These results are observed from the attacker's perspective after he tries to intrude client's privacy by stealing the user search logs.

USER	DATE & TIME	SEARCH DATA
user1	2016-02-29 17:25:44.0	imag
user1	2016-02-29 17:26:38.0	0CBAA46E44F363002E5FAFB016D5FADD2BA57213 2C2E6D263143CD98B0DBA8963400214CA538FB56
user1	2016-02-29 17:28:05.0	30CDC82B863DB3118F5F4F8D976B45430CEA8204
user1	2016-02-29 17:28:14.0	pixel
user1	2016-02-29 17:29:37.0	digit comput
null	2016-03-01 10:44:06.0	imag
null	2016-03-01 10:48:58.0	imag
user2	2016-03-01 11:00:16.0	imag
user2	2016-03-01 11:00:27.0	0CBAA46E44F363002E5FAFB016D5FADD2BA57213
user2	2016-03-01 11:23:29.0	indian tradit danc
user2	2016-03-01 11:24:28.0	stock market
user2	2016-03-01 11:24:57.0	sachin tendulkar
user2	2016-03-01 11:25:38.0	F60AA46E700712A6299BFECF5522E4AB2E22417D 579D0533A2C605F804F4E648B797E207290CE036
user2	2016-03-01 11:25:48.0	DE84E7E1786C57FA98BD10B192302FF0991BC1D7 C5F937FA3733F4EF53C006CE70E67544BDE0DC3A
user2	2016-03-01 11:26:10.0	3DCD02AAF0D3946931C8351AE8AB973B0F997E64 127E58FC82F4003B633CC2D7D32492F4B247F3FB 6B11FD1353E9A91BFB27F9...

Figure 2: Client search logs

In figure 2, the public queries are displayed as they are as the privacy techniques are not applied on them.
The private queries however, go through a series of processing. Text mining and hashing techniques are applied on them and a hash value is generated at client side and is sent to the server for retrieving results. In this way the client's privacy is preserved and only the irreversible hash value of the query is recorded in the user logs.

V. CONCLUSION AND FUTURE WORK

Web users increase because of the availability of large amount of information on the web. Providing relevant result to the user is based on their click logs, query histories, bookmarks. Text mining and hash algorithms can be used to preserve the privacy of the user as well as to maintain the performance of the search. User interests and user logs can be used to re-rank the data for relevant results. In future, more categories can be added to the database to increase the scope of the application.

REFERENCES

[1] LidanShou, He Bai, Ke Chen, and Gang Chen," Supporting Privacy Protection in Personalized Web Search",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL. 26, NO. 2,2014
[2] 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.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

[3] 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.
[4] 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.
[5] X. Shen, B. Tan, and C. Zhai, "Privacy Protection in Personalized Search," SIGIR Forum, vol. 41, no. 1, pp. 4-17, 2007.
[6] Y. Zhu, L. Xiong, and C. Verdery, "Anonymizing User Profiles for Personalized Web Search," Proc. 19th Int'l Conf. World Wide Web (WWW), pp. 1225-1226, 2010.
[7] 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.
[8] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2006.