Privacy Protection in Personalized Web Search using Homomorphic Encryption

Ghilby Varghese Jaison, Charlse M Varghese

Department of CSE, KMPCE

Abstract- Personalized web search (PWS) has showed its effectiveness in improving the quality of search services on the Internet. However, evidences show that users' fear to disclose their private information during search has become a major barrier for the wide proliferation of PWS. Here we are introducing the concept of Homomorphic encryption to encrypt the server such a way that an eavesdropper nor an untrusty admin could access the search words and the profile. It is less complex and more efficient than the UPS framework.

Index Terms- Privacy protection, personalized web search, utility, risk, profile.

I. Introduction

The web search engine has long turn into the most vital source for individuals searching for helpful data on the web. However, clients may encounter disappointment when web search return unimportant or unwanted results that don't meet their requirements. Such immateriality is generally because of enormous variety of users' contexts and backgrounds, and additionally the equivocalness of writings. Personalized Web Search (PWS) is a general class of pursuit procedures going for giving better indexed lists, which are custom-made for individual client needs. As the cost, client data must be gathered and broke down to make sense of the client expectation behind the issued inquiry.

The PWS can be divided into two types, to be specific click-log-based systems and profile-based systems. The click-log-based strategies are clear they basically force inclination to clicked pages in the client's search history. Despite the fact that this system has been exhibited to perform reliably and extensively well[2], it can just chip away at new domain inquiries from the same client, which is an in blow to the keeping of its appropriateness. Interestingly, profile-based routines enhance the inquiry involvement with convoluted client interest models created from client profiling strategies. Profile-based systems can be possibly successful for a wide range of questions, however are accounted for to be temperamental under some circumstances[2].

Although there are advantages and disadvantages for both types of PWS techniques, the profile-based PWS has proved more effective in improving the quality of web search recently, with increasing usage of personal and behavior information to profile its users, which is usually gathered from query history [3], [4], [5], browsing history[6],[7], click-through data[8],[9], bookmarks[10] , user documents[3],[11] , and so forth. Unfortunately, such certainly collected personal data can easily reveal a matters of interest of user's private life.

Privacy issues rising from the lack of protection for such data, for instance the AOL query logs scandal, not only raise panic among individual users, but also hinders the enthusiasm of client in using personalized search service. In fact, privacy concerns have become the major barrier for wide success of PWS services. The major concern being that of an eavesdropper and that of an untrusting admin at the server. From fig 1 shows the attack model where an eavesdropper can eavesdrop both the search query and client profile in transit, and that of an untrusting admin at search engine server. The admin as well as the eave can have full access to data send from client.

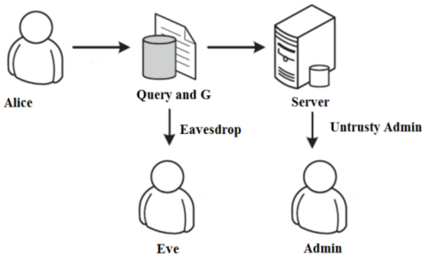


Fig 1: Attack Model of PWS

A. Motivations

To ensure client's security in personalized web search the analysts need to consider two variables amid the pursuit process. The primary component being the thought for expanding the proficiency of customized web search. Also, the second variable being the need of security of the client profile to place protection chance under control. By incorporating the concept of Homomorphic Encryption into the existing system can improve privacy. Thus, user privacy can be protected without compromising the personalized search quality. The problems with the existing methods are explained in the following observations:

1) At the client level user needs to repeatedly select the word or terms that need to be hidden from the eavesdropper
2) The system is complex and takes much time for computation.
3) The system is away from PWS.

II. THE EXISTING SYSTEM

L. Shou et al. proposed "Supporting Privacy Protection in Personalized Web Search" to protect user privacy in profile-based PWS, specialists need to consider two repudiating impacts that affects the search process. From one perspective, they endeavor to enhance the inquiry quality with the personalization utility of the client profile. Then again, they have to conceal the security substance existing in the client profile to put the protection hazard under control. The existing system known as UPS has following components :

A. User Profile

Consistent with numerous past works in personalized web search, each user profile in UPS adopts a hierarchical structure. A diagram of a sample user profile is illustrated in Fig. 2(a), which is constructed based on the sample taxonomy repository in Fig. 2(b). As the owner of this profile is mainly interested in Computer Science and Dance, because the major portion of this profile is made up of fragments from taxonomies of these two topics in the sample repository. Some other taxonomy also serves in comprising the profile, for example, Sports and gender.

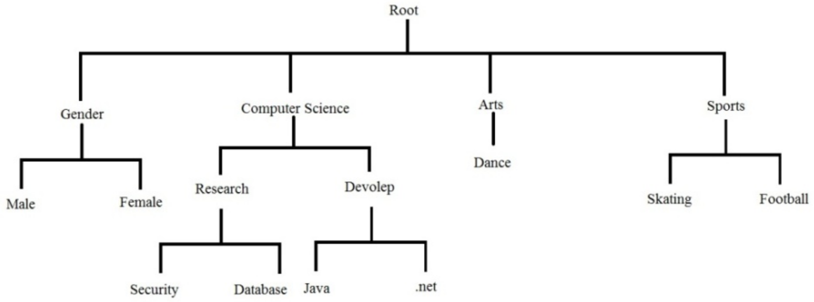


Fig 2(a): Sample User Profile

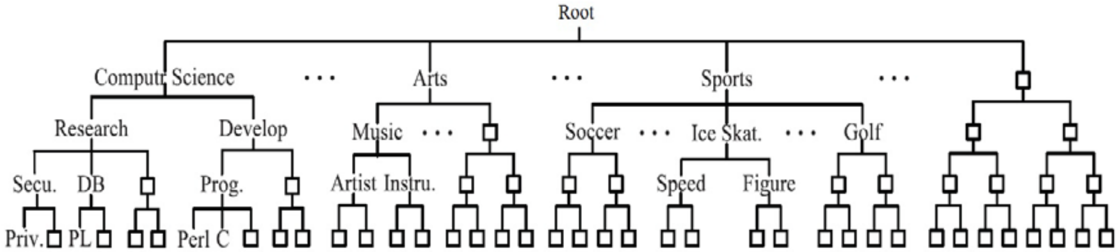


Fig 2 (b) Sample Taxonomy Repository

B. Customized Privacy Requirements

Customized privacy requirements can be specified with a number of sensitive-nodes (topics) in the user profile, whose disclosure (to the server) introduces compromising he privacy of the user. In the sample profile shown in Fig. 3.1 a, the sensitive nodes S = {Adults; Security; Skating}; are shaded in gray color in H. It must be noted that user's privacy concern may differ from one sensitive topic to another. In the above example, the user may hesitate to share her personal interests only to avoid various advertisements. Thus, the user might still tolerate the exposure of such interests to trade for better personalization utility. However, the user may never allow another interest in topic gender to be disclosed.

As the sensitivity values expressly show the client's security concerns, the most clear protection saving strategy is to uproot sub-trees established at all delicate hubs whose sensitivity value are more prominent than a limit. Such system is alluded to as forbidding. In any case, forbidding is a long way from enough against a more refined foe.

C. Generalizing User Profile

To address the problem with forbidding, the existing system proposed a technique, which detects and removes a set of nodes X, that are sensitive, from H, such that the privacy risk introduced by exposing G is always under control. For clarity of description, it is assumed that all the sub-trees of H rooted at the nodes in X do not overlap each other. This process is called generalization, and the output G is a generalized profile.

The generalization technique can seemingly be conducted during offline processing without involving user queries. However, it is impractical to perform offline generalization due to two reasons:

1. The output from offline generalization may contain many topic branches, which are irrelevant to a query. A more flexible solution requires online generalization, which depends on the queries. Online generalization not only avoids unnecessary privacy disclosure, but also removes noisy topics that are irrelevant to the current query.
2. It is important to monitor the personalization utility during the generalization. However, overgeneralization may cause ambiguity in the personalization, and eventually lead to poor search results. Monitoring the utility would be possible only if generalization is done at runtime.

III. THE HOMOMORPHIC ENCRYPTION SCHEME

Parameters: The construction below has many parameters to control the number of integers in the public key and the bit-length of the various integers. As basic,we use the following four parameters for the algorithm :
γ is the bit-length of the integers in the public key,
η is the bit-length of the secret key ,
ρ is the bit-length of the noise, and
τ is the number of integers in the public key.

These parameters are set under the following constraints:

• ρ = ω(log λ), to protect against brute-force attacks on the noise;

• η ≥ ρ ・ Θ(λ log2 λ), in order to support homomorphism for deep enough circuits to evaluate the "squashed decryption circuit";

• γ = ω(η2 log λ), to thwart various lattice-based attacks on the underlying approximate-gcd problem;

• τ ≥ γ + ω(log λ), in order to use the left over hash lemma in the reduction to approximate gcd.

We also use a secondary noise parameter ρ′ = ρ + ω(log λ). A convenient parameter set to keep in mind is ρ = λ, ρ′ = 2λ, η = Õ (λ^2), γ = Õ(λ^5) and τ = γ + λ. (This setting results in a scheme with complexity ˜O(λ^{10}).) For a specific (η-bit) odd positive integer p, we use the following distribution over γ-bit integers:
Dγ,ρ(p) ={ choose q $\overset{\$}{\leftarrow}$ Z ∩ [0, 2^{γ}/p), r $\overset{\$}{\leftarrow}$ Z ∩ (−2ρ, 2ρ) : output x = pq + r }.

A. The Construction

KeyGen (λ).

The secret key is an odd η-bit integer: p ← (2Z + 1) ∩ [2η−1, 2η).

For the public key, sample x_i← Dγ,ρ(p) for i = 0, . . . , τ . Relabel so that x_0 is the largest. Restart unless x_0 is odd and rp(x_0) is even. The public key is pk = <x_0, x_1, . . . , x_τ >.

Encrypt (pk,m ∈ {0, 1}):

Choose a random subset S ⊆ {1, 2, . . . , τ} and a random integer r in $(-2^{\rho'}, 2^{\rho'})$, and output c ←[m + 2r + 2P_i∈S x_i]$_{x0}$.

Evaluate (pk,C, c1, . . . , ct):

Given the (binary) circuit C_ε with t inputs, and t ciphertexts c_i, apply the (integer) addition and multiplication gates of C_ε to the ciphertexts, performing all the operations over the integers, and return the resulting integer

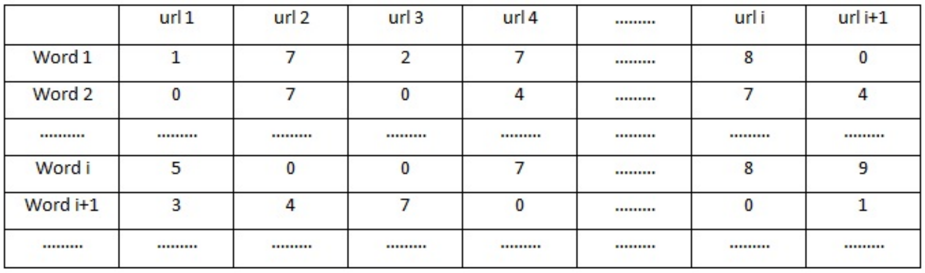
Decrypt (sk, c):

Output m′ ← (c mod p) mod 2.

IV. THE PROPOSED SYSTEM

The new scheme called "Privacy Protection in Personalized Web Search using Homomorphic Encryption" is proposed here to support privacy preservation in personalized web search using an index server which is encrypted using Homomorphic encryption. It provides privacy for personalized web searches at server.

In the system proposed attains personalized web search using click-log based technique. Our main consideration is for privacy of searched terms at server and that of an eavesdropper. For that we send all the data's ,i.e. search terms, using index number for the search word which will be encrypted homomorphically before sending it to the server. We assume of an index matrix as shown below, it consists of 3 components word, url and rank which denotes words to be searched for, unique resource locator of sites and rank for a word with respect to the url it can vary from 0 to n, respectively.

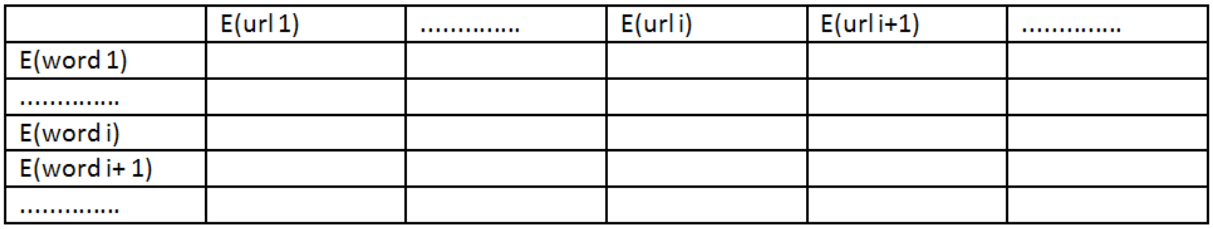


	url 1	url 2	url 3	url 4	....	url i	url i+1
Word 1	1	7	2	7	....	8	0
Word 2	0	7	0	4	....	7	4
....	....	....	....	....	....	....	....
Word i	5	0	0	7	....	8	9
Word i+1	3	4	7	0	....	0	1
....	....	....	....	....	....	....	....

Fig 3: Index Matrix

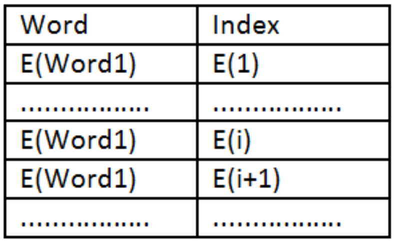
We have a similar index table in our index server but with a difference that the index matrix is homomorphically encrypted in the proposed system and instead of words we store index for each word as shown below where E() shows the Homomorphic encryption function.

The index for each word is created or determined as an continues process at web crawler and is simultaneously updated at each client. This index will send from client to server rather than words as such, moreover this index will also be homomorphically encrypted.



	E(url 1)	....	E(url i)	E(url i+1)	....
E(word 1)					
....					
E(word i)					
E(word i+ 1)					
....					

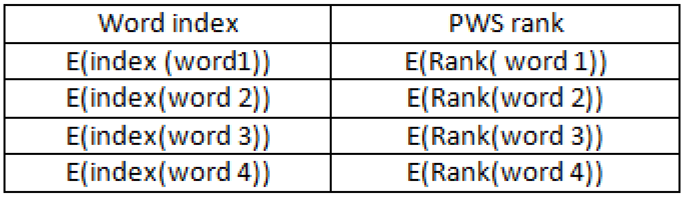
Fig 4: Homomorphic encrypted index matrix



Word	Index
E(Word1)	E(1)
....	....
E(Word1)	E(i)
E(Word1)	E(i+1)
....	....

Fig 5: Word Indexer

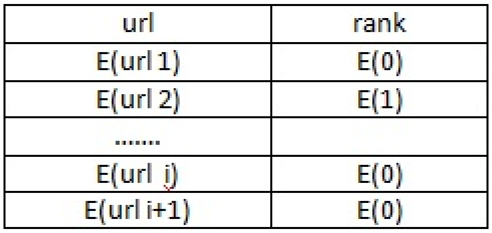
When user searches for a word, a matrix called query matrix is send to the index server form of query matrix is as shown below, it consists of two columns word index and rank. It is an n*2 matrix where n stands for each n search words. Each word index is homomorphically encrypted and has a value attached to it which is the rank of word for the client. The figure 4.4 below shows query matrix for 4 search words say word 1, word 2, word 3 and word 4.



Word index	PWS rank
E(index (word1))	E(Rank(word 1))
E(index(word 2))	E(Rank(word 2))
E(index(word 3))	E(Rank(word 3))
E(index(word 4))	E(Rank(word 4))

Fig 6: Query matrix

This value column is multiplied with corresponding column in the index matrix and sum of each column is taken. What will be send back to the server will be the list of urls with corresponding obtained ranks and key for decryption of data send by index server.



url	rank
E(url 1)	E(0)
E(url 2)	E(1)
....	
E(url j)	E(0)
E(url i+1)	E(0)

Fig 7: Results

On receiving the url matrix it will be decrypted using the decryption key send from the index server. Then according to the rank value attached url's will arranged in order with url with highest rank displayed first and so on

The whole functioning of the system can be summarized as show below at client side user gives up the word for search, then corresponding word index, rank and profile is send to the server, remember all data send to and from server are homomorphically encrypted. The server then processes the received data to process result for the corresponding search indexes. The server then sends the results matrix ,which includes url and rank of each url, back to the client. The web crawler continuously builds the search engine and the word index matrix at client and send to the client along with encryption key.

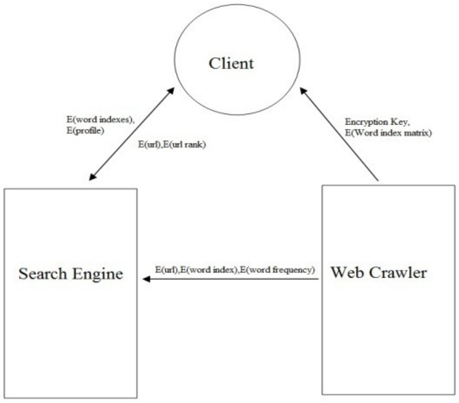


Fig 8: The Proposed System.

V. CONCLUSION

This paper presented a client-side privacy protection framework based on Homomorphic encryption for personalized web search. The system could potentially be adopted by any PWS that can in-co-operate an homomorphically encrypted server. The framework allowed users to have full privacy over the PWS. Our experimental results revealed that the system could achieve quality search results without affecting the privacy of client/end user. The results also confirmed the effectiveness and efficiency of our solution.

References

[1] Lidan Shou, 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 Year 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

[14] 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.

[15] 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.

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

[17] E. Gabrilovich and S. Markovich, "Overcoming the Brittleness Bottleneck Using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge," Proc. 21st Nat'l Conf. Artificial Intelligence (AAAI), 2006.

[18] K. Ramanathan, J. Giraudi, and A. Gupta, "Creating Hierarchical User Profiles Using Wikipedia," HP Labs, 2008.

[19] K. Ja¨rvelin and J. Keka¨la¨inen, "IR Evaluation Methods for Retrieving Highly Relevant Documents," Proc. 23rd Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), pp. 41-48, 2000.

[20] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley Longman, 1999.

[21] X. Shen, B. Tan, and C. Zhai, "Privacy Protection in Personalized Search," SIGIR Forum, vol. 41, no. 1, pp. 4-17, 2007.

[22] Y. Xu, K. Wang, G. Yang, and A.W.-C. Fu, "Online Anonymity for Personalized Web Services," Proc. 18th ACM Conf. Information and Knowledge Management (CIKM), pp. 1497-1500, 2009.

Authors

First Author : Ghilby Varghese Jaison , B-Tech, CEH, KMPCE, ghilbz1991@gmail.com
Second Author : Charlse M Varghese, BE, ME, Asst. Professor-KMPCE, charlse.varghese.1@gmail.com