Secure Advanced Web Search Personalization

Mr. A. A. Patil
Dattakala Group Of Institution's
Faculty of Engineering
Swami-Chincholi,Daund, Pune-413 133
Email:amrishpatil1989@gmail.com

Prof. Amrit Priyadarshi
Dattakala Group Of Institution's
Faculty of Engineering
Swami-Chincholi,Daund, Pune-413 133
Email: amritpriyadarshi@gmail.com

ABSTRACT

General search engines are critical for recovering significant data from web. However these search engine take the "one for all" model which is not flexible to particular web users. In this paper we are try to enhance personalized web search as will as privacy of search. Client's Profile gives a critical information to performing personalized web search. We propose a PWS framework. where we are going to create Advanced user profile using user browsing history, and enriching it with domain knowledge, to achieve security and confidentiality we are going to encrypt the user profile. we also provide the security to the query which is requested from client side, by encrypting it at client side and decrypt at server side. result will also encrypted at server side and decrypted at client side. OWASP security guidelines are followed while designing the system. We also provide online system to decide which query session is typical and which one is atypical. That is system will automatically decide which query to be personalized.

Index Terms- Personalized Web Search, User Modelling, Domain Knowledge, Enhanced User Profile, Confidentiality, privacy.

I)INTRODUCTION

Now a days search engine became most powerful way for finding the right information in right way, but as the size of the net grows user needs more accurate search result as per their needs. Most popular Search engine like Google, yahoo are always ahead to improve their search algorithm and search engine to satisfy the user requirements. With amount of original content there are various noisy information are also present like spam and advertisement. Personalized Web Search (PWS) is the one of the category of web technique that aims for providing search result whatever required by user. A typical search engine do not properly handles the ambiguity of various search query for example if one user wishes to search sports of golf while another user may be inquiring about the automotive offering of a Volkswagen “Golf”. here both user may feel unhappy because of unwanted ads and search results because service providers revenue is depends on the ads. One way to solve this ambiguity is provide more personal information to server so that user will get the appropriate result as per the requirement. To perform Personalized Web search it is essential to model User's need/interest. Development of user profile is an essential part for customized web search. User profiles are built to model user's need focused around his/her web use information.

There are two main methods to structure the personal information handling of a personalized search service. The first method is keeping all user profile information at server side it’s called as server side personalized search, and second is keeping all user profile information at client side it’s called as client side personalized search. We are going to user server side profiling. One of the natural issues with personalized search is that clients are often frail about giving private or individual data in regards to themselves to search providers. Naturally, the more that a search provider thinks around a particular client, the more precise their search results can be custom-made for them, yet how are the clients to trust that the data that the search provider keeps up about them won't be misused, lost, or vindictively utilized?.We also aim to make user profile more precise and advance by using users browsing history and Domain knowledge and in order to enhance the privacy, this paper will look at philosophies and methods to optimize the privacy that users are given when using a typical personalized search service, User Profile Encryption schema is used to achieve confidentiality and privacy. Here we assumed possible security threats for example1) A2-Broken Authentication and Session Management 2)A3-Cross-Site Scripting (XsSS)3)A6-sensitive data exposer 4) A1-Injection and to cope with it we are using security guideline from OWASP top 10 .it also gives the protection from possible pollution attacks. And to differentiate typical session and atypical query session we are going to use query and user profile based on this system will decide which query should be personalized .

II) RELATED WORK

In this section, we overview the related works. We focus on the literature of various web search personalization and privacy techniques.

Previous work on personalization web search include [4]that uses a hybrid approach of personalized Web Information Retrieval that utilizes ontology for retrieval of user's context ,user profile that is temporarily updated according to users' browsing behavior and collaborative filtering for considering recommendation of similar users. It gives better result than normal user profile.

Several user in past has proposed search personalization by using various technique like[13]that collects the user information using search activity and build the profile based on it later results are re-rank based on user interest.[6]uses the location preference to personalize the query , it separate the location concept and content concept and organize then to create user profile. Protecting the user data is also important part of our framework, some authors like[3] arranging their individual data into a hierarchical client profile, where general terms are ranked to more elevated amounts than particular terms. Through this profile, clients control what bit of their private data is presented to the server by conforming the minDetail threshold. An extra protection measure, expRatio is proposed to estimate the measure of protection is uncovered with the defined minDetail value.

Several authors also proposed techniques to find out atypical and typical query session like[5],it uses session level features like session length, query length ,unique terms per sessions, clicks per query and profile based features like query term divergence, topic divergence to decide which query session is typical and which one is not.[1]have proposed a personalized web search display that consolidates community based and content built proofs situated in light of novel ranking procedure. These days, transferring information on web has turned into a day by day movement. A massive amount of information is transferred as web pages, news, and web journals and so forth all the time. In this way, it gets to be exceptionally troublesome for the client to search for significant substance. Not just for clients additionally for search engines like Google and Yahoo it gets to be troublesome. Data overload is the main reason behind this troublesome circumstance. Other than this present client's inclination is the second issue, which is not contemplated while delivering the outcomes. The author attempted to explain this issue through this model which deliver results on the premise of inclination and enthusiasm of the client. In this paper, authors proposed a special approach to discover the investment and inclination of the client. It's a two way approach, first it will discover out the exercises of client through his/her profile in social networking sites . Furthermore, it will figure out data from what the social networking site give to the client through friends and community. In light of the outcomes, client's interest also inclination will be prioritized by the web search or it is personalized

III) PROBLEM STATEMENT & IMPLEMENTATION

We propose a framework for secure personalized web search which considers singular's enthusiasm into mind and upgrades the customary web search by proposing the important pages of his/her interest. We have proposed a straightforward and productive model which guarantees great recommendations and in addition guarantees for successful and important information recovery. Our system consider user profile which is based on user browsing history and Domain knowledge. we use domain knowledge to specify information that belongs to different categories. And by analyzing user behavior i.e. browsing history and navigation it will learn user choices and it will manipulate user profile dynamically by analyzing most recent choices.

Proposed design additionally concentrates on safeguarding privacy of client profile by securing sensitive data of client profile . It uses browsing history to discover the client interest. Secrecy and integrity is accomplished by encrypting client profile at server side .security is likewise given to transportation of individual data to search provider, and in addition to the outcome which is exchange back from server to customer. We give a inexpensive mechanism to the customer to choose whether to personalize a query in PWS. Our framework also focuses on proper user interface, so that user can easily

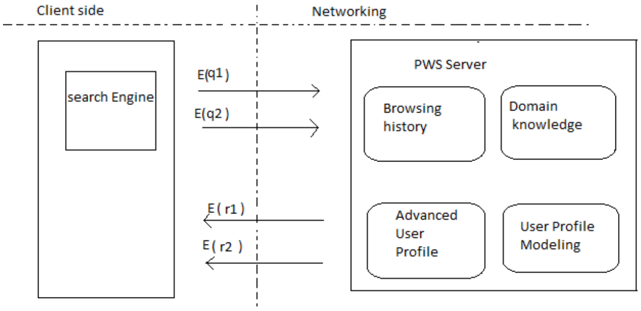

Fig1 . System architecture of Secure PWS

IMPLEMENTATION.

A) Domain Knowledge Modeling

Domain Knowledge Modeling information is the foundation learning that we used to upgrade the client profile. The source which we have utilized for preparing Domain Knowledge is DMOZ directory. to prepare Domain Knowledge, first we have crawled the website pages from DMOZ directory for some pointed out classifications, where every class is spoken to by gathering of URL's available in that classification. we have extricated the keywords from the crawled site pages. The accumulations of keywords, structure the vocabulary for the crawled pages. Presently we structure a term category matrix, which determines weight of each one term in each category. The weight may be spoken to by recurrence of the term in that category. In order to collect the domain knowledge we have crawled the dataset from DMOZ using Apache Nutch server. and Apache Solr has been used for indexing crawled pages

B) User Profile Modeling

To create the client profile, we have to group the website pages got to by a client into specific classification. Alchemy API will download the requested URL, extracting text and other important content from the HTML document structure, and perform document categorization alongside confidence (numerical quality) which demonstrates its likelihood of fitting in with that specific class. If the page is arranged with confidence above the indicated threshold level then we can consider that page. We using DMOZ later where this Alchemy categories are mapped in to DMOZ category. Learning agent learns the user interest and updates the user profile when web pages browsed by user cross the threshold level. User interest will thus be represented by fix number of categories weights. It can be denoted by U={$\mathrm{cw}_{1,}\mathrm{cw}_{2,}\mathrm{cw}_{3,}$.....$\mathrm{cw}_m$} Where, CW_i will be the number of web pages of category i visited by that user, normalized by maximum number of page visits among all categories .

C)Advanced user profile.

Advanced user profile is important part of our framework. with the help of domain knowledge it improves the overall search result. advanced user profile are created using url present in domain knowledge. We take url from user profile and find relevant url of this from domain knowledge using cosine similarity.

To create Advanced user profile first we take url(document) from user profile, add this url into advanced user profile find the cosine similarity of this URL with urls present into specific domain, retrieve first 20 url after arranging in descending order .then calculate average cosine similarity of this 20 url. find out the url which have value greater than average value and add it to the enhanced user profile.

To summarize this process , A cosine similarity measure is the angle between the web page in User Profile u and the document vector dj.

$$cosine(d_j, u) = \frac{< d_j * u >}{\|d_j\| \times \|u\|}$$

Cosine simlilarity gives good suggetion of relative urls.later we are attempting to improve the url suggetion by again re-ranking the number of url which we gor from cosine similarity. Aprori algorithm is used for this purpose.

D)Privacy and confidentiality of user profile

In our framework we are going to utilize server based user profiling, so we are all that much focused on security and privacy of client profile. Here server side encryption is used, each client profile is initially encrypted then it will stored at server side. For server side encryption RSA algorithm with 2048 bit key is used is used. bigger key size provide more security. At the point when customer fires the query then first it is encrypted then it will be send to the server, at the server side encoded query is decrypted, when server process that query re-ranked result is again encoded and send it to the customer.TLS/SSL is used form providing security for user data transportation. Through this we can accomplish client confidentiality and information integrity .To achieve security we are following the guideline from OWASP top 10 security cheat sheet. This framework provide the security from various Vulnerabilities for example Broken Authentication and Session Management,Cross-Site Scripting (XsSS),sensitive data exposer,and Injection. Our framework provides good protection against Pollution attack, in which user profiles are polluted by false seeds. so it will resist third party to change the rank of web pages.

IV) SYSTEM CONFIGURATION

TABLE1:Software Configuration

Operating System	*Windows*
Technology	Java and J2EE
Web Technologies	Html, JavaScript, CSS
Web Server	Tomcat
Database	My SQL
Java Version	J2SDK1.5

TABLE2:Hardware Configuration

Processor	PENTIUM – IV
Speed	2.0 GHz
Ram	512 MB (MIN)
Hard disk	20 GB
Keyboard	Standard windows keyboard
Mouse	Two or three button mouse
Monitor	SVGA

Before we get into the technical details of our framework lets see some benefits of this.

1)Different search results can be provided depending upon the choice and information needs of users.

2)Maintain Privacy of User

3)Provides security from inter and external threads

.

V) RESULT

Project demonstrate a framework for personalized search, we are going to use Apache Nutch for crowing the dataset from DMOZ. Apache Solr is used for indexing the crawled pages. Using the browsing history and domain knowledge we create advanced user profile. url’s are suggested based on user profile so here we compare the pages suggested by advanced user profile and normal user profile.

Result shows that advanced user profile gives better search result than normal user profile on same query.

VI) CONCLUSION

In this paper we have proposed a framework for secure personalized web search ,here we build the user profile by using domain knowledge, system keeps updating the user profile by suggesting relative url that make search more effective. We also proposed a system to maintain the privacy and confidentiality by encrypting the user profile at the server side. We performed some experiments that shows better search result when we use advanced user profile as compared with simple user profile on same queries. The result also confirms the effectiveness and efficiency of our solution. In future we try to enhance the search quality based on user search preference. We also aims to provide more security from the adversaries.

REFERENCES

[1] O. Shafiq, R. Alhajj and 1. G. Rokne, "Community Aware Personalized Web search", International Conference on Advances in Social Networks Analysis and Mining, pp. 3351 - 355,2010

[2] G. Chen, H. Bai, L. Shou, K. Chen, and Y. Gao, “Ups: Efficient Privacy Protection in Personalized Web Search,” Proc. 34th Int’lACM SIGIR Conf. Research and Development in Information, pp. 615-624, 2011.

[3] Yabo Xu, Benyu Zhang, Zheng Chen, KeWang,”Privacy-Enhancing Personalized Web Search”, WWW 2007, May 8–12, 2007

[4] Namita Mittal, Richi Nayak, MC Govil, KC Jain “A Hybrid Approach of Personalized Web Information Retrieval” 010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 978-0-7695-4191-4/10

[5] Carsten Eickhoff, Kevyn Collins-Thompson “Personalizing Atypical Web Search Sessions” *WSDM’13,* February 4–8, 2012.

[6] K.W.T. Leung, D.L. Lee and Wang-Chien Lee, "Personalized Web search with location preferences", IEEE 26th International Conference on Data Engineering, pp. 70 I - 712, 2010 .

[7] Zhicheng Dou, Ruihua Song, Ji-Rong Wen, and Xiaojie Yuan “Evaluating the Effectiveness of Personalized Web Search” Ieee Transactions On Knowledge And Data Engineering, Vol. 21, No. 8, August 2009

[8] Namita Mittal, Richi Nayak, MC Govil, KC Jain”A Hybrid Approach of Personalized Web Information Retrieval” 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[9]J.Teevan,S.T.Dumais,and D.J. Liebling, “To Personalize or Not to Personalize: Modeling Queries with Variation in User Intent,” Proc. 31st Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 163-170, 2008.

[10] 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.

[11] Wei Meng, Xinyu Xing, Anmol Sheth, Udi Weinsberg, Wenke Lee ”Your Online Interests – Pwned! A Pollution Attack Against Targeted Advertising” CCS’14, November 3–7, 2014, Scottsdale, Arizona, USA.

[12] Milad Shokouhi, Ryen W. White, Paul Bennett, Filip Radlinski,” Fighting Search Engine Amnesia: Reranking Repeated Results”, SIGIR’13, July 28–August 1, 2013

[13] M Speretta and S Gauch, "Personalized Search Based on User Search Histories", Proceeding Of International Conference on Web Intelligence,pp. 622-628, 2005

[14] G. Chen, H. Bai, L. Shou, K. Chen, and Y.Gao” Supporting Privacy Protection in Personalized Web Search” ieee transactions on knowledge and data engineering vol:26 no:2 year 2014.