Extended Search Model Enhancing Personalization with Privacy Using User Identity

Nidhi Saxena[1], Bineet Gupta[2]

[1]*Information Technology, SRMGPC ,Lucknow, UP,India*
[2]*Computer Application, SRMU,Lucknow,UP, India*
Corresponding Author: Nidhi Saxena

Abstract: Internet is huge source of data and information and every one depends on it for attaining each type of knowledge. With the existence of large data and variety of information it becomes difficult for end user to extract required information. Therefore there comes the necessity to personalize each user search in order to obtain selected and relevant result. This paper is an extension to current personalized search model being used in search engine for increasing the efficiency of search engine and to provide user with more relevant and desired results. The proposed search model distinguishes each user search by uniquely identifying each user, providing fast and efficient results. It is compared with existing technology in search engine. Experimental test conducted on the user browsing indicate that the proposed scheme efficiently improves the search results with minimal structural changes. The proposed design is more useful where more than one user access single device for searching. The system provides user friendly behavior as searching will now an easy task. The results will be more accurate as well as secure. Therefore the system acts as an improvement in the area of web mining.
Keywords: -Explicit,Implicit,Page ranking, Personalization, Web Search log

---

I. INTRODUCTION

Our world revolves around internet and we are becoming dependent on World Wide Web for attaining any type of knowledge .Search engine is accessed by every age of people, for gaining variety of information. On the other hand millions of pages are added every day therefore the size of information available on the internet is increasing day by day. This results in rising size of data on web. This growing size of data provides better service to user but they have to spend more time on the web or searching relevant result. Sometimes user fails in getting required and interested information. Such problem can be minimized to a great extent with the use of web personalization. Personalization is the methodology of foreseeing the user needs by considering users' behavior and interest explicitly or implicitly.

Explicit collection of user feedback is an extra effort from user side as it requires certain answers to queries before searching which is not practically possible all the time. On the other side the implicit approach for understanding user interest seems to be more user's friendly. Various methods in personalization developed which are based on analyzing user search history on server or on client. These methods provide relevant results but problem arises when different users keep different interest related to identical query. For example two users making similar query "RAM". First user is a religious person wish to read some article on Lord Ram but the second user is interested in computer "Random Access Memory" and interested in attainting some knowledge about computer RAM memory. If both the users uses short query result will not be ranked according to user need and will get results related to all possible meanings of RAM. In order to reduce such problems there is a necessity to provide thorough description of query each time while searching. The searching time as well as precision of results both affect.

To identify different interest and to reduce searching problems there is a necessity to personalized result based on each user requirement.

Our system is an extension to personalization approach being used. It allow user to distinguish his query and thereby provides unique results with more user friendly approach. The system reduces human efforts in writing lengthy queries and analysis the user interest to provide relevant results.

II. RELATED WORK

Since 1995 with the introduction of Page Rank algorithm, ranking method wasintroduced, in order to improve the efficiency of web search process. Many different personalization model developed so far and being in use which follow explicit and implicit profiling techniques. Dynamic ranking [1] takes feedback of user in order to have user involvement in ranking process. Another algorithm Query extension [2] uses tags and web

document as a social means for personalization. Probabilistic profile RLT profile [3] considers key entries of web search and develops a user profile for ranking. Many other profile developing algorithms and methods for searching came into scenario. Data from social means like semantic extraction from news article [4] shown a great impact on personalization.

Many techniques createsweb log by collecting information explicitly for understanding user interest and search optimization. Some of them consider documents presents on system [5] [6]. Such algorithms considers that if the document is on device it reflects the interest of user in those documents and can be used for ranking searched pages.

On the other side implicit personalization technique are very much popular and are in great demand now a days. Among them we have technique which considers three conditions [7] Session, Historic and Aggregate. Session record all earlier work in the current session, historic keep an eye on the work other than current session and aggregate look towards all work by user. This technique provides all possible ways in which we can find the user interest .Implicit profiling can be seen in algorithms [8] [9] [10] which records every page visited by user and time spend on them. These algorithms provide efficient way of tracking user search activities and providing search result based on their previous activities. In order to make the searching more optimized and user oriented click event was considered later on [11]. Click event shows the interest of user in that page which means that user has gone through that page after searching. As the number of factors increases the result improves and personalization becomes more advanced and effective. Further in 2013, Ratio Rank [12] model was introduced which uses in link and out link weights along with number of visit count which was an added feature in personalization. For implicit personalization three major factors were came into light-
• Visiting frequency
• Time spend
• Click event

These factors were considered by New Enhanced page rank [13] along with mouse track. The user behavior is thoroughly understood by considering these factors.

Implicit and explicit profile creation together is being use in some algorithms which keep some personal information along with user behavior while searching. Email is also considered as an interest factor in some algorithm [14].Technique like Embedded [15], Automated Analysis of Interests and Activities [17], Measuring and Predicting Search Engine Users' Satisfaction [18] had a great impact on personalization.

Many advanced techniques developed recently in few years taking web usage mining [19]. User navigation behavior is predicted using Hybrid approach [20] and Clustering algorithm is adopted [21] [22] for finding user preference.

III.   IDENTIFICATION OF PROBLEM

Various methodologies for personalization are in use as discussed above. New researches are conducting to improving and making the search process more efficient. Both explicit and implicit method of personalization uses user data whether taken explicitly or implicitly and create web search log on the server. This results to heavy load on server and also leads to privacy issues. A list of problems identified in current search engine are listed below-

• As the search history increases on server or PC computational complexity increases
• Users face problems while specifying the information they want which results to wastage of time in searching by making useless hits.
• Some words have more than one meaning then user need to give long query in order to save time.
• If there exist more than one user searching from same system then results will be same for each user.
• Lack of privacy seems to be major effect of personalization. Web search log on server or clients contains lot of information of user which can be disclosed.

In order to overcome the given problem there is a necessity of a system which can provide relevant search results in efficient way.

IV.   EXTENDED PERSONALIZED SEARCHED MODEL

The proposed model is an extended approach to personalization using implicit method. The system records user search data in web log. Beside this it keeps record of user identity in client log history in order to uniquely identity the user search.
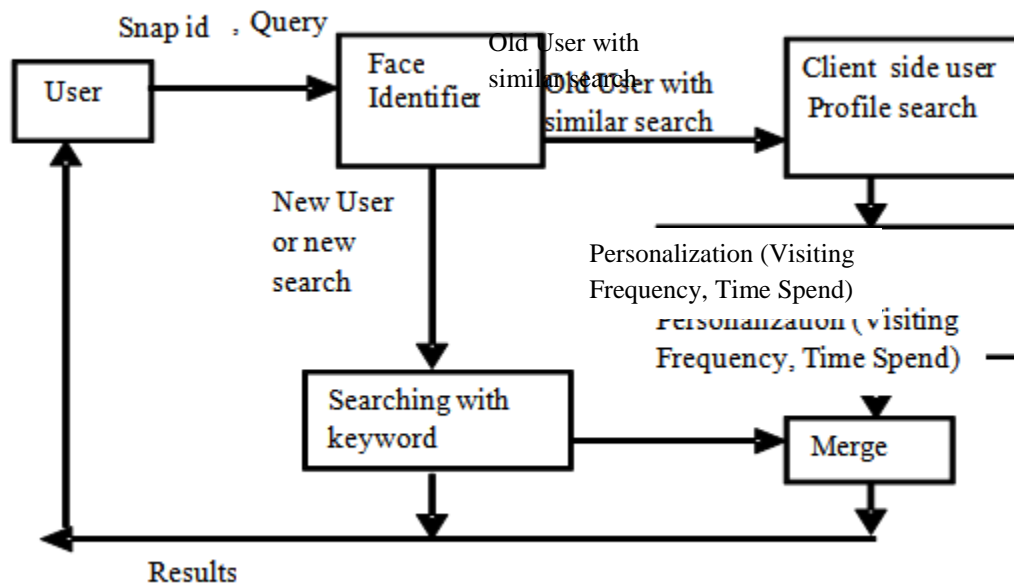
Fig. 1 Extended search model

The given model illustrates the personalization task in a unique way. It combines the personalization based on user interest as discussed above with a new concept of keeping user identity and using it for searching and ranking.
The model encompasses four phases-

4.1 Face Identification:
The system identifies user at the time of login in order to maintain unique search record in web log. This identification is performed by taking user snap for the first time and using it further identification. This will allow separate search record and personalize the search.

4.2 Web log creation:
The system maintains search log based on user identification. Log history will be created on each client machine. Number of log history depends on the number of users using the same system. Each log contains identity record along with page visited frequency and time spent.
The web log history consists of following records-
• User Identity
• Query
• Searched URL
• Time Spend

4.3 Searching and Ranking:
User enters the keyword which need not be very descriptive .The keyword will be matched in the user search log. If exist the ranking will be based on frequency of search and time spend. But if the keyword is missing then it proves that the user is making such query for the first time and results will be displayed without ranking. The ranking will start by the second time with the same keyword.
The system considers two ranking factors-
4.3.1 Frequency – Frequency defines the number of times the page is visited by user in the earlier searches. As the frequency of a web page visited increases its probability of top rank also increases. It is measured by increasing the count by 1 each time the page is visited by the same user.
4.3.2 Time Spend- Frequency along with time spend provides proper interest of user in a web page. Sometimes user visits a page but do not show much interest in that page and close it therefore if user selects a page that does not reflect that user is interested in that page . Therefore if user opens a web page and spend sometimes on it then we can consider page relevancy. It is estimated with the help of calculating time duration for which the web page remains open.

4.4 Merging- After searching and ranking based on user identity the ranked results will be merged with other searched data based on keyword. The ranked results will be displayed at the top for giving user friendly approach along with other searched pages.

The approach described above is keeping in record the interest of each user by user's identification which results to make all search history personalized on separate search log created on client machine and maintains security by distinctly maintaining search record. The complexity and time is reduced to greater extend as ranking depends on separate log on client not on server.

Algorithm for Extended Personalized Model-
If New user
 Then
Crawled (Matching with keyword)
Results displayed
Add data in client web log
 Else
If (query match not present in the web log)
Then
Crawled (Matching keyword)
Results displayed
  Add data in client web log
 Else
Crawled (Matching keyword (MK))
Calculate max. Frequency(MF), max. Time spend (MT)
Ranking on interest (MK+MF+MT)
Results displayed
Add data in client web log

V. EVALUATION METHOD

For the purpose of evaluation two queries "kingfisher" and "mouse" are taken having more than one meaning and search differs with person to person. For the purpose of evaluation a comparison is performed between two search logs. One is of standard transaction log of web search engine and other is proposed system web log on client side having user search record. The system considers two different persons having different interest making above two queries from single machine. The results will be saved in both the search log shown below-

Table1. Web Search Engine Standard Transaction Log

Uid	Date	Time	Searched URL
ce00160c04c4158087704275d69fbecd	25/Sept./2018	04:08:50	http://www.flykingfisher.com
ce00160c04c4158087704275d69fbecd	25/Sept./2018	04:08:50	http://www.kingfisherworld.com
ce00160c04c4158087704275d69fbecd	25/Sept./2018	06:08:15	http://a-z-ni als.com/animals/kingfisher
ce00160c04c4158087704275d69fbecd	25/Sept./2018	07:08:15	http://en.wikipedia.org/wiki/Mouse

TABLE 1 show the data logged on standard transaction log for both searched queries. In standard transaction log user is identified by IP address. This is used as a user code address taken by the search engine server. This system is also keeping data regarding time spend by user which is taken as a personalization factor but it does not distinguishes between people. There persist a problem that more than one user may be using the same device for searching with different interest therefore here the personalization is not user specific but machine specific.

Table 1 . Client Machine Proposed Web Search   Profile

User ID	Query	Searched URL	Time Spend
User1.jpg	Kingfisher	http://www.flykingfisher.com	10
		http://www.kingfisherworld.com	8
	Mouse	http://en.wikipedia.org/wiki/Mouse...	12
		https://www.apple.com/in/magicmou	11
User2.jpg	Mouse	http://en.wikipedia.org/wiki/Mouse	12
		http://www.ncbi.nlm.nih.gov/genome	8
	Kingfisher	http://a-z-Nimals.com/animals/kingfish	9
		http://www.birdlife.org/datazone	8

TABLE 2 is keeping User identification record in the form of snapfor unique identification of different users on client side. Web search log records data on the basis of user identity. Two users perform search on two different words "King fisher" and "mouse".These words have more than one meaning. User1 wish to attain some information about Kingfisher flight and computer input device mouse while other User2 was searching for a bird named kingfisher and an animal mouse. Both users are using same machine for searching. Earlier standard transaction log was keeping records without distinction and therefore provide personalized result based on machine.

On the other hand the proposed web log keeps distinct record. Ranking will become easier by having unique identification. Relevancy of the model can be analysed on the basis of returned web pages. User oriented ranking is performed in the proposed system which is actually being required by the user not the machine oriented.The system maintains accounts on client side of each user who accesses the system and any unknown person is not allowed to make search from particular system even all the users are not allowed to look into the searched results of each other which lead to increased privacy.

A comparison is made between the existing search engine and the extended personalized search based personalization and privacy.

Table3. Comparison of Model with Existing Search Engines

	Google	Bing	Extended Personalized model
Personalization	Search log on client and location based	Search log on client and location based	Based on individual search log on client with identification
Privacy and security	Through encryption	Secured private datatransfer	Very High Due to user identification and separate search log,

VI.   CONCLUSIONS

Personalized search improves the search quality that lead to highly advance search engine with more user satisfaction. It is a method to provide relevant results in short time with less effort and therefore considered to be user friendly approach. . However, this approach requires full access to personal information on Internet, which disrupts users' privacy.

In this paper, number of existing techniques have been discussed and analysed. It is obvious that for personalizing a search, user profiling is necessary. However, all the methods that are used uses client IP address for creating search log. This creates a problem of interest conflict as more than one user can make search from same machine and create privacy issues.

The extended search model is an effort to reduce these problems by keeping search web log history on client machine with distinct facts related to users searching on it search engine. For the purpose of distinction in search log face recognition method is used.

The result shows that user profile is helpful in improving search quality when united with the original page ranking. This paper is emphasizing on personalization with   protection making search method to be more user friendly.

Yet, this paper is an exploratory work on small scale. There is a necessity to adopt such features while maintaining search log in home or in office where more than one user is using the same system for search. The system can act as a basis for higher scale search engine.

REFERENCES

[1]. C. Brandt , T. Joachims, Y. Yue, J. Bank ,Dynamic Ranked Retrieval,*Proc.4th ACM Conf. on Web search and data mining*,Hong Kong, China ,2011.

[2]. S. Lawless S, V. Wade, D. Zhou, Improving Search via Personalized Query Expansion using Social Media. Information Retrieval For Social media, *Springer,15(3)*,2012,218-242.

[3]. P. N. Bennett,S. T.Dumais, J. Y. Kim JY, Thompson KC, Characterizing Web Content, User Interests, and Search Behavior by Reading Level and Topic,*WSDM'11 ACM*, 2012, 9–12.

[4]. F. Abel, Q. Gao, G. J.Houben, K. Tao, Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web,*ESWC'11, Vol. Part II, Springer2*, 2011, 375-389.

[5]. C. Srinvas, Explicit User Profiles for Semantic web search using XML, *IJERA, 2(6)*, 2012, 234-241.

[6]. R. S. Bhadoria ,R.Moriwal, D. Sain , Data Mining Algorithm for personalizing user's profile on Web, *JCTEE,1(2)*,2011.

[7]. P. Bailey ,N. P. Bennett, F.Borisyuk ,W. Chu, X. Cui, S. T.Dumais , R. W. White, Modeling the Impact of Short- and Long-Term Behavior on Search Personalization., *Proc. 35th Annual ACM SIGIR Conference,*2012.

[8]. V. Keselj, H. Liu, Combined Mining of Web Server logs and web contents for classifying user navigation patterns and predicting user's future requests, *Data and Knowledgeengineering Elsevier, 61(2)*,2007, 30 330.

[9]. R. Khanchan, M. Punithavalli, An Efficient WebPage Prediction based on Access Time Length and Frequency.*Proc. IEEE 3rd International Conference on Electronics Computer Technology*,2011.

[10]. N. Matthijs, F. Radlinski, Personalizing WebSearch using Long Term Browsing History, *ACM WSM'11* 2011.

[11]. R. Agarwal, K. V. Arya, S. Shekhar,An Efficient Weighted Algorithm for Web Information Retrieval System.,*Proc. IEEE InternationalConference on CICN,* 2011.

[12]. D. K. Sharma ,R. Singh ,Ratio Rank :EnhancingImpact of Inlinks and Outlinks., *Proc. IEEE International Advance Computing Conference*,2013.

[13]. D.K.Sharma,R. Singh,Enhanced Ratio Rank:Enhancing Impact of Inlinks and Outlinks.,*Proc. IEEEInternational Advance Computing Conference*, 2013.

[14]. D. E. Horvitz, T. Susan, J. Teevan , Personalizing Search via Automated Analysis of Interests and Activities,*Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 51(3)*, 2017.

[15]. M. Johnson, D. Q. Nguyen, D. Song, T. Vu, A. Willis,Search Personalization with Embed dings ,*Proc. ACM 39th European Conference on Information Retrieval, ECIR* ,2017.

[16]. S. Lawless, J. Liu ,X. W. Zhao , D. Zhou , Query Expansion with Enriched User Profiles for Personalized Search Utilizing Folksonomy Data,*Proc..IEEE transaction on Knowledge and Data Engineering, 29(7)*, 2017.

[17]. D. E. Horvitz,J. Teevan ,T. Susan , Personalizing Search via Automated Analysis of Interests and Activities, *ACM SIGIR Forum , 51(3)*, 2017.

[18]. O. Dan, B. D. Davison, Measuring and Predicting Search Engine Users' Satisfaction, *Proc. ACM Computing Surveys, 49(1)*,2016.

[19]. R. Khatri, D. Gupta, An Efficient Periodic Web Content Recommendation Based on Web Usage Mining, *Proc. IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*,,2015,132-137.

[20]. M. Narvekar, S. Banu, Predicting Users Web Navigation Behavior Using Hybrid Approach,*Proc. International Conference on Advanced Computing Technologies and Applications (ICACTA-2015)*,2015,3-12.

[21]. R. Thiyagarajan, K. Thangavel, R. Rathipriya, Recommendation of Web Pages using Weighted K-Means Clustering, *International Journal of Computer Applications (09758887), 86(14),* 2014, 44-48, 2014.

[22]. X. Li, Y. Cheng, The Improved Clustering Algorithm for Mining Users Preferred Browsing Paths *Springer International Publishing Switzerland*,2016.