Refinement in Personalize Web Search System with Privacy Protection

Anuradha K. Kudlikar
M.E.(CSE)
GECA, Aurangabad.
Maharashtra, India.

Meghana B. Nagori
Asst. Professor (CSE)
GECA, Aurangabad.
Maharashtra, India.

ABSTRACT
There are number of users searching for particular information with same topic. Personalized web search helps to improve the excellence of various searches on the Internet. But during searching the search engine may disclose or use user's personal information to improve search performance. We propose a fine tuning in Personalize Web Search system by generalizing user profiles. We suggest a technique to generate online profile with user's permission for query. Every time when user requests for certain information ,our system allows user to select profile information as per his or her requirement and risk of exposition of sensitive attributes such as name ,gender, contact number and many other different attributes. In addition our systems will also help to search accurate information based on user interests. Thus this system maintains stability between use of personalize information and the risk of exposing of personal profile by refining profile. This system is developed by GreedyIL algorithm which improves search quality and makes search computation fast

Keywords
Personalize web search, profile generalization, Greedy algorithm, discrimination power, information loss.

1. INTRODUCTION
In today's era most of the internet users use internet for searching all types of information like technical ,medical, educational, cultural entertainment, transportation, tourism information and many more. To search this information user takes the help of different search engines like Google.com, msn.com, yahoo.com, Alta vista etc. Search engine is the basic application in computer field which helps user to treasure desired information based on entered keyword. A perfect search engine is able to find out very specific information that satisfies user's requirements. But many times this won't possible, e.g. if user wish to search for the apple then different types of information like apple fruit, Apple company. Apple phones are being searched by the search engine. If in this case a user is a nutritionist, he or she may be interested in getting apple fruit information for its nutrition values. Or if the user is an electronic shop keeper he may be interested in finding new products of Apple company.

This leads to give birth to a concept called as Personalized Web Search (PWS). PWS is a technique that helps user to search appropriate information based on user demands requirements. For searching this type of information search engines like Google, search [1], yahoo search [2] uses user's information present in profile and previous search histories. This process is called as profile generalization. Profile generalization helps user to search expected information by selecting appropriate words from user profile. But the main drawback of this extraction is that without user's permission unnecessary as well as sometimes secret data may be used to search from the profile. This violates the privacy of the user.

To prevent from such unnecessary exposure of data we are proposing a solution to generate selected user profile every time, when user searches for the information. This reserves user privacy as well as helps user to get exact information as per his or her requirement.

2. RELATED WORK
Many Researchers and authors have contributed and put their efforts for Personalization Web Search .We are presenting researches and views of some of them.

According to J. Pitkow, H. Schuetze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel [3] there are mainly two techniques for search engines to search the information according to user's interest. First is contextualization. It searches according to information available on that topic, nature of search like web, images, pdf files etc. and applications that uses that search. Second technique is individualization. In this technique user's goal, previous search history is being considered and search results are given to the user.

Z. Dou, R. Song, and J.-R. Wen [4] describes two solutions to do Personalized Web search namely click log base and profile base. Click log base method selects search according to user's previous selected search. Profile based focuses on gathering the information from user's profile and generates the result according to that interest. They also proved that profile based search is more suitable as compared to click based searches.

Different types of Profile based methods are focused by many authors. From J. Teevan, S.T. Dumais, and E. Horvitz [5] focuses on user's history. From this history they suggest

re-ranking methods to get ranking of result. M. Spertta and S. Gach [6] , B. Tan, X. Shen, and C. Zhai [7] research for query history methods .This method collects user logs and based on that user profile is created and further used to access the information.

K. Sugiyama, K. Hatano, and M. Yoshikawa [8] and X. Shen, B. Tan, and C. Zhai [9] focus to create a profile based on browsing history. Browsing history is obtained through following ways -1). From user registration details which is filled up by user while profile creation. 2) From different queries that are submitted at the time of searching. 3) From results displayed by web search engines. To improve the

privacy different levels like pseudo identity, group identity are suggested.

X. Shen, B. Tan, and C. Zhai [10] and F. Qiu and J. Cho [11] have contributed to access user profile based on the click through data.

To collect all this type of information J.S. Breese, D. Heckerman, and C.M. Kadie [14] have suggested different reranking algorithm types as rank scoring, page rank, average rank algorithms. In rank scoring algorithm information to search is ranked. Page rank is based on number of times page clicked for information. Average ranking is nothing but calculating average of number that times the search is given.

Y. Xu, K. Wang, B. Zhang, and Z. Chen [12] contributes in enhancing privacy in personalize web search. To access the user profile full access to server is required. They show that many servers access this user information without permission of the particular user. This violates the privacy of user profile. Every time it is not necessary for server to go through the detailed user profile. For this it is required to gather the entire user profile in collective manner.

A. Krause and E. Horvitz [13] show that if user gets the good quality of search information, the user is willing to extract his or her personal and private data. In practical case if a small amount of secrete data which consists of less sensitive user profile is extracted then up to certain levels user's privacy is maintained.

To summaries the previous work we can say that click through methods are not that much sufficient to get the better search result. Profile based search are most suitable to get expected information for user. But the search quality and privacy protection both are proportional with each other. To improve search quality there is a need to access user's profile with accessing less sensitive user information and if possible with user's permission.

The existing system may have following drawbacks.

(1) Generally user profile is filled by the user at the time of registration for any services like different Google services, yahoo services. Every time that same profile is being used by the search engine's server to access information. Again for any service e.g. for Google this profile may be used for search, email, map and many more services provided by the Google. The server may try to expose unnecessary user profile to improve the search result. Instead of extracting offline information from user's profile, it is possible for a server to extract required user profile by taking online decision at the time of search. Server extracts only that much of information present in user's profile.

(2)While extracting the information some sort of data may be exposed beyond limits or it may be possible that some sort of data is much protected. This may generate adverse effects on search. More sensitive information from user profile is exposed for so many times. So while searching, everybody's privacy considerations and privacy levels of information should be taken in to account.

(3)Previous approaches are used iteratively to gather information for searching. This may disclose sensitive information many times and through many servers for searching to get better search results. In existing system all these sensitive information which is extracted is called as suprisal.

3. GREEDY ALGORITHM

Greedy technique is an algorithm design policy, built on configurations such as different choices, values to find objective .Greedy algorithms produce good solutions on mathematical problems. The important goal here is to discover some configurations that are either maximized or minimized. Greedy Algorithms provide a solution for optimization problems that has certain sequence of steps, with a set of choices for each step. Another solution for Greedy algorithm is dynamic programming. It is also used to conclude the best possibilities that can accepted. But greedy algorithm always makes the choice that is best at the moment to provide the optimal solution for the problem. A greedy algorithm for an optimization always provides the current sub solution. Basically greedy algorithm always gives an optimal solution to the MST (Minimum Spanning tree) problem. Some Examples that are solved by greedy algorithm are Dijkstra's shortest path algorithm and Prim/ Kruskal's algorithms.

• Configurations: It consists of different choices, values to apply on data.

• Objective: some configurations to be either maximized or minimized to get the predefined objective.

Each step in Greedy expands to construct sub solution until a complete solution to the problem is developed. Each solution for every step is feasible as it satisfies the problem's constraints. Once choice is made, it cannot be changed for subsequent steps of the algorithm. Greedy method works best when applied to problems with the greedy-choice property.

Greedy algorithms are categorised in three different types.

1) Pure greedy algorithms

2) Orthogonal greedy algorithms

3) Relaxed greedy algorithms

We propose privacy preservation technique based on user customizable privacy preserving search that can generalize profiles i.e. creates own user profile every time by using queries with respective user's privacy requirements. Generalized profile tries to maintain a balance between two things related to the personal. First is security of sensitive data present in personal profile of that user and second is the risk of exposing that sensitive information from the generalized profile. At the same time the search quality is also improved. The main aim here is to develop online profile. It can be developed by using two greedy algorithms. Those are-(1) GreedyDP (Greedy algorithm that reduces Discriminating Power) and (2) GreedyIL(Greedy algorithm that supports less information loss).

3.1 GreedyDP (Greedy Algorithm for reducing Discriminating Power)

In sociology, discrimination is nothing but an unexpected treatment to an individual based on his or her group or category [15]. P. Yuvasari, S. Boopathy focuses on direct and indirect discrimination in data mining algorithms. Discrimination is the process of denying opportunities and services to one group that are available to other groups. Discriminating power of an item distinguishes the items between two groups that are having higher ability from those having lower ability.

Data mining is a large way of creating discrimination, because it avoids certain data from data set by creating discriminatory models from data set through automated decisions. On the other hand it shows that data mining can become both a source of making discrimination and discovering discrimination items present in the dataset. Discrimination of data is categorized in two types as direct discrimination and indirect discrimination. Direct discrimination consists of data mining rules that inherently mention underprivileged groups

based on sensitive discriminatory parameters present in that data set. Indirect discrimination consists of data mining rules that will not explicitly mention the discriminatory items.

The direct greedy algorithm works in a bottom up manner. GreedyDP starts from each i^{th} iteration. It selects a leaf item for discrimination. At the same time, it tries to maximize the output from the current selected iteration [16]. During the iterations, a best profile is selected which shows the highest discriminating power by considering specific risk constraints. This process is continued till all the nodes from leaf to root is selected. From traversing all nodes the generalize profile is created. The main problem of GreedyDP is that it requires recalculation of profiles. This recalculation is of calculating sensitive items from lowest level i.e. leaf node up to the highest level .i.e. root. Also while mining for every node we have to consider their discriminating power and privacy risk. This algorithm also takes lot of memory requirements as compared to GreedyIL algorithm. It's computational cost is also more.

3.2 GreedyIL (Greedy Algorithm for Less Information Loss)

Exposing individual's data without disclosing sensitive information present in the profile is main important problem in privacy preservation. Information loss (IL) is difference in the measurement of the original database and the database after data mining that consists of sensitive information present in the data set. For example if the database consists of hospital dataset, the hospital prefers no information loss as well as protects patient privacy. This is achieved by not disclosing the name of the patient. For this re-identification of the information present in dataset is used. For better privacy preservation databases with low information loss is always preferable. The GreedyIL algorithm improves the efficiency of profile generalization by considering this unnecessary exposure of data.

If we select GreedyDP algorithm we have to work from leaf node to root. Any leaf operation reduces the discriminating power of the profile. Instead of that if we start from root to leaf which is the approach of information loss occurs in GreedyIL algorithm. It gives less information loss as well as less discriminating power. From above finding it suggests to use GreedyIL algorithm to develop candidate profile in descending order

By using GreedyIL algorithm reduces the total computational cost as compared with GreedyDP as it generates profile from root to leaf [17]. By using GreedyIL less data loss is achieved.

If we compare the both algorithm we realize that first maximizes the discriminating power (DP) with increasing computation cost. But GreedyIL minimizes the information loss (IL) with less computation cost required to compute profile. Performance time of GreedyIL is less as compared to GreedyDP.

4. EXISTING SYSTEM

For searching personalized web information mainly server side architecture is used. In server side architecture browser present at client side sends the query regarding search topic to the search engine's server. After receiving the query server goes through its database for searching related documents regarding it. Server sends the search result in the form of ranking result by rearranging. To get more efficiency in search sometimes server may take help of other related information regarding user. Server might use user's browsing history, cookies; last clicked pages as well as some times it may use user's profile. This user profile is filled once at the time of registration done by the user for use of different types of services by search engine like email, YouTube, Orkut ,map and many more.

The main advantage of this architecture is information in large volume which can be presented by server based on user's history and profile. But in order to satisfy the user requirement server sometimes illegally access the user's profile without his or her consent. This may violate user privacy and there by chances to expose the sensitive data present in user's profile if available.

Today most of the search engines use server architecture to search information. To increase search quality as well as protection against such unauthorized access, we propose the algorithm that will support user profile's personalization which is generated at client side. This client side architecture has following advantages over server side architecture – 1) User profile is generated at the client's side according to user's requirements. 2) It avoids unnecessary exposure of user's profile. 3) Profile is generated according to user's topics selection, so it helps to improve search quality.

5. PROPOSED SYSTEM

We propose greedy algorithm to generalize user profile at client side. This profile generalization depends on two metrics: a) information used from user's profile b) risk arrived due to exposure of that information. Our main aim here is that there should be less risk to disclose the sensitive information present in the profile as well as to improve better search results as per the user's expectations.

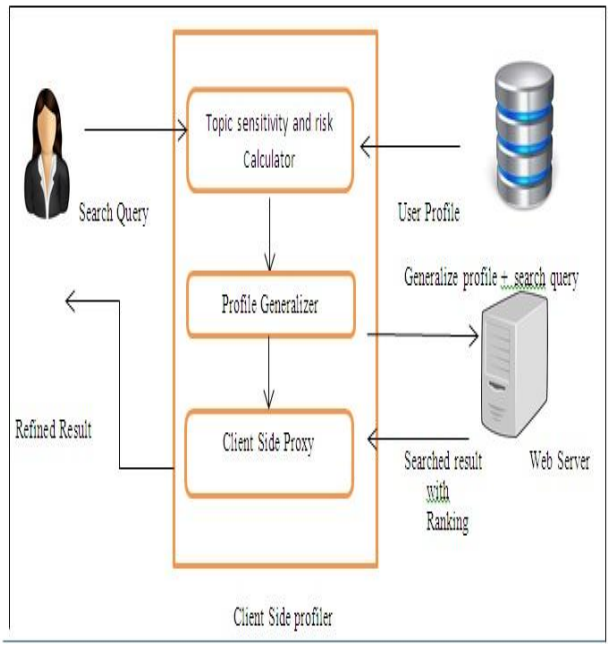

Figure 1. System Overview

Figure 1 shows the idea of complete system which is used to generate generalize profile at client side. Thus we can give freedom to user to select profile contents for searching required information.

5.1 Module Description

Refinement in Personalize Web Search System with Privacy Protection system mainly consists of four modules.

1. Offline profile generation.

2. Detection of offline privacy requirements.

3.Topic prioritization and generation

4. Online decision

1) Offline profile generation – User uses any kind of services provided by search engine like mail ,map ,You Tube, searching, chat provided by different search engines. For accessing these services for first time registration of user is completed by filling user's profile. This user profile includes user's interests for registration, his or her personal information, photos, address, contact numbers and much more. This detailed profile is generated only once. After profile filled up initial support for each topic or information present in the profile is calculated.

2) Detection of offline privacy requirements – For the support of information present in profile, threshold value of support is decided. If in present profile information any information's support is below of decided threshold then that respective topic is declared as sensitive topic. A set of all such sensitive topics is created. This is considered as risk of profile expositions which may be sensitive and interferes user's privacy.

3) Topic prioritization and generation – After generating offline profile only once, next step is to generalize that profile. It involves the creation of user profile every time according to user's requirement and sensitivity of information present in the topic. Whenever any query is generated by the user for any search, client profile generator computes a sub tree. For computing this sub tree .taxonomy based algorithm is used. This sub tree consists of topics present according to user's interest from the user's detailed profile. For this selected sub tree, difference between support and risk factor is calculated, on the basis of individual profile data present in the list. If this difference is above threshold value the risk of maximum profile exposure will be reduced. Thus from the third step we get the customized profile from user's detailed profile.

4) Online decision – After obtaining customized profile the decision is made by this step whether to send this profile to server for searching further information. If user is willingly ready to expose topics presented by generalized profile then this profile is finalized and profile is sent to server for search information accordingly.

5.2 System Requirements

For implementing this system dual core processor with 1.1 GHz speed or above configuration is required.

Software will able to execute with minimum 1GB RAM and 200 GB Hard disk.

This system is supported by many operating systems like Windows XP, Windows 07, and Windows 08 etc.

Application will be developed in Java. Java is used because of it's richness of many classes and packages present in it.

Front end will developed in Java. The main advantage in selection of Java is that it is platform independent with support of runtime environment by byte code generation and Application Programming interface. As program runs inside

virtual sand box environment Java application will be secured as compared with other development platforms. Moreover due to run time environment support Java is able to run on internet platform, that is distributed environment. For development of this software we will require Java Servlets. A Servlet is part of a Java web application. A Servlet has container that run multiple web applications at the same time, each having multiple servlets (server side program) running inside. The browser sends an HTTP request to the Java web server when request for web server is being typed in the title bar of browser window. The web server checks if the request is for a servlet. If it is for the servlet, the servlet container is passed the request. The servlet container will then find out appropriate servlet for that request, and activate that servlet. The servlet is activated by calling the Servlet.service() method. Once the servlet has been activated via the service () method, the servlet processes the request, and generates a response accordingly. The response is then sent back to the browser. There are many advantages of using Servlet as compare to CGI scripts (Common Gateway Interface). Servlet uses multithreading concepts. The web container for servlets creates threads for handling the multiple requests at a time to the servlet. Threads are nothing but the light weight processes that share a common code segmentation and user memory. So the cost of communication between the threads is low. The basic benefits of servlet are as follows:

a) Servlet can have better performance as it creates a thread for each request..

b) Servlets are managed by Java Virtual Machine so no chance of memory leak, garbage collection problems

The database is stored in Oracle. Oracle 10G express edition is the server side database used to store data. Oracle database net services allow to access data in remote application via various network sessions. As this will be used as back end to store profile records in database.

As this is based on client server architecture TOMCAT server will required to perform as server operations. Apache TOMCAT is open source web server supports to run distributed applications developed in JAVA.

Apache Tomcat is an open source software implementation of the Java Servlet and Java server Pages technologies. The Java Servlet and Java Server Pages specifications are developed under the Java Community Process. It is the web container that allows running Java Server pages and Java Servlets. As the Privacy refinement software is working in distributed environment it needs this freeware open source TOMCAT server.

5.3 Data Set

We are creating our own data set. The following table shows some of the records present in database. This database table profile consists of following fields which is filled once when user registers to the web site. Fields present in this table are as follows:

u_id, uname gender, address used to store user's identification number, user's name, user's gender and user's address respectively. User's contact number, email ids are stored in contact_no, email_id field. For storing user interest four different user interest fields are defined in the table. This may include different interest areas like literature, IT, Engineering, Medical, politics, travelling etc. This field consists of different interest areas of the user.

Tale 1. Data base used to profile generalization

u_id	uname	gender	bdate	address	contact_no	email_id	interest1	interest2	interest3	interest
1201	Anita Desai	F	11/2/1982	Auranagabad MH,India	9425631542	anitadesai@gmail.com	literature	IT	nutrition	fashion
1202	Megha Bhoge	F	2/2/1985	Auranagabad MH,India	9895689658	m_bhoge@yahoo.com	IT	bollywood	tourisam	
1203	Nilesh Wagh	M	5/8/1984	Pune,India	8087548945	nwagh@gmail.com	Civil	Politics	Newpaper	religion
1204	Mahesh Joshi	M	9/7/1987	Pune,India	8085794548	joshimahesh@rediff.com	IT	History	Maps	Politics
1205	Ashish Nene	M	20/7/1989	Pune,India	9456123568	ashishnene@gmail.com	Maths	bollywood	litreature	gadgets
1206	Sunit Sharma	M	23/9/1990	Mumbai,India	8589784501	ssharma@gmail.com	gadgets	mobiles	bollywood	religion
1207	Nikita Katariya	F	5/5/1991	Mumbai,India	9456120012		IT	fashion	diet	yoga
1208	Ketan Mohite	M	8/9/1980	Mumbai,India	8985478960	mohiteketan@gmail.com	IT	Spiritual	meditation	Politics
1209	Anil Nikam	M	7/10/1979	Nasik,India	8975480015	nikamanil@gmail.com	Medicine	bollywood	yoga	travellir
1210	Aditi Padhye	F	16/8/1990	Mumbai,India	8084794512	aditip@gmail.com	Fashion	yoga	travelling	

This one time filled user profile is accepted as input and every time depending on user's requirement from this profile table a generalized profile is created online. This generalized profile includes only certain filed which are of user's interest.

For generating user's generalize profile we take the help of GreedyIL [17] algorithm. Algorithm for calculating user profile will be as follows:

Algorithm: compute_genprofile(Pi,t,th)

Input: profile information (Pi),topic (t),threshold of risk(t_r),query item(qi),support for topic (s), threshold of support(t_s)

Output: Subtree (st), generalized profile(Gp)

1. Accept user profile Pi at the time of registration. Represent it in the form of topic t. Set threshold value for the profile. As shown in the above database table this profile consists of all the attributes which are necessary for storing user information.

2. For each t in Pi, accept support s and compute the sensitivity risk r by r=risk (t(s)). In this step we are calculating the relation between support and sensitivity which is used to identify the risk of exposition of specific attribute present in the table. e.g. for gender field risk is calculated by considering it's support s. The threshold is the value that specifies risk in entire profile is exposition.

3. Accept user's query for searching required information. For each query qi, generate a sub tree st of topic that can have maximum support and minimum risk (r).a for each t in Pi b. if ((s(t)>t_s) and (r(t)<t_r) then insert(t,st) If for each topic the risk of exposition is below threshold then it is add to leaf node.

4.Calculate overall support cost and overall risk cost.

a cost(S) = $s(t_0)+s(t_1)+s(t_2)+\ldots s(t_n)$

b cost(R) = $r(t_0)+r(t_1)+r(t_2)+\ldots r(t_n)$

5. Return Gi= (st, max(S), min(R)).

Here finally we get generalized profile from step 4 and 5 which has exposition profile risk is minimum than threshold. So there is no harm to use all the attributes present in this user's profile.

6. CONCLUSION

Communication by using networks and internet enables to reach a very large volume of information in a minimum amount of time. Search engines service us to provide the required information.This paper presents a client-side privacy protection for personalized web search technique. Providing privacy to one time user profile is based on generating a user profile by supporting customization. It allows us to generate online query. This allows users to protect the personal privacy without exposing sensitive data. At the same time search quality provided by search engine will remain constant. The above system uses greedy algorithm GreedyIL, for providing the online privacy protection.

7. ACKNOWLEDGMENTS

During the entire period of this preparation, it would not have been materialized without the help of many people, who made my work easier.

My special thanks to Prof. M. B. Nagori madam**,** my project guide, for kind guidance especially in finalizing this topic and for giving suggestions time to time. She has been constant source of inspiration**.**

I am thankful Computer Department of Government College of Engineering, Aurangabad, Maharashtra, India. for providing all technical facilities for development of software.

Last we would like to thank the anonymous reviewers for their valuable comments that helped to improve the quality of this paper.

7. REFERENCES

[1] Google personalized search: http://www.Google.com/psearch

[2] Yahoo! My Web 2.0: http://myweb2.search.yahoo.com/

[3] J. Pitkow, H. Schuetze, T. Cass, R. Cooley, D. Turnbull, A.Edmonds, E. Adar, and T. Breuel, 2002.Personalized search. Communications of the ACM, 45(9):50-55.

[4] Z.Dou, R. Song, and J.-R. Wen 2007. A Large-Scale Evaluation and Analysis of Personalized Search Strategies. Proc. Int'l Conf. World Wide Web (WWW), pp. 581-590.

[5] J. Teevan, S.T. Dumais, and E. Horvitz 2005. 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.

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

[7] B. Tan, X. Shen, and C. Zhai 2006. Mining Long-Term Search History to Improve Search Accuracy. Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD).

[8] K. Sugiyama, K. Hatano, and M. Yoshikawa 2004. Adaptive Web Search Based on User Profile Constructed without any Effort from Users. Proc.13th Int'l Conf. World Wide Web (WWW).

[9] X. Shen, B. Tan, and C. Zhai 2005. Implicit User Modeling for Personalized Search. Proc. 14th ACM Int'l Conf. Information and Knowledge Management (CIKM) .

[10] X. Shen, B. Tan, and C. Zhai 2005. Context-Sensitive Information Retrieval Using Implicit Feedback. Proc. 28th Ann. Int'l ACMSIGIR Conf. Research and Development Information Retrieval (SIGIR).

[11] F. Qiu and J. Cho 2006. Automatic Identification of User Interest for Personalized Search. Proc. 15th Int'l Conf. World Wide Web (WWW), pp. 727-736.

[12] Y. Xu, K. Wang, B. Zhang, and Z. Chen 2007. Privacy-Enhancing Personalized Web Search. Proc. 16th Int'l Conf. World Wide Web (WWW), pp. 591-600.

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

[13] J.S. Breese, D. Heckerman, and C.M. Kadie 1998. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. Proc. 14th Conference Uncertainty in Artificial Intelligence (UAI), pp. 43-52, 1998.

[14] P. Yuvasri S. Boopathy 2013. Performance Analysis on Direct and Indirect Discrimination in Data Mining. International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 11, ISSN: 2277 128X.

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