IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-3, Issue-9 E-ISSN: 2347-2693

Comparison User Customizable Privacy-preserving Search (UPS) with User Customizable Online Privacy-preserving Search with K-anonymity (UCOPSK)

V. Kavitha[1] and T. Uma Maheswari[2*]

[2] *Department of Computer Science,*
Sri Ramakrishna College of Arts and Science for Women, 95, Sarojini Naidu Road, New Sidhapudur,
Coimbatore, Tamil Nadu 641 044, India

www.ijcseonline.org

Received: Sep/02/2015 Revised: Sep/11/2015 Accepted: Sep /27/2015 Accepted: Sep/30/2015

***Abstract*—** Based on user interest and information requirement Personalized Web Search (PWS) delivers different search results for disguised users. Personalized web search have disguise characteristics while compared with common web search, as which deliver same set of search result for the same keyword search, by different kind of user have different needs. Really, these diligences have become one of the main hurdles for locating personalized search and how to do privacy-preserving personalization is a extensive challenge. Hence to overcome these difficulties privacy protection in Personalized Web Search provides a model hierarchical user profile, which have been built based on user preferences. Propose a PWS framework User Customizable Online Privacy-preserving Search with K-anonymity (UCOPSK) which generalizes profile as per the user specified privacy requirements in online and offline search. In this proposed work the profiles are constructed for each static and dynamic user in the websites. K-anonymity is applied to each user profile to manifest of sensitive information of user in privacy preservation, which can significantly prevent the sensational information leakage under attacks, and it is commonly used in discrete fields now a days. This paper describes the various approaches and techniques of preserving user data applied on personalized web search to build up a new algorithm & method to improve performance, utility and security of existing data and help to create the new predictions on the data. This paper describes the comparative study of clustering techniques used to improve privacy preservation on personalized web search.

***Keywords*—**Web pages, web search engines, personalized web search, web mining, privacy protection, risk, profile, generalization and k-anonymity.

I. INTRODUCTION

To get vital information on web, web search engine has become the important tool for common people. Sometimes users receive irrelevant information which do not meet their search criteria. This situation happens due to enormous amount of user's information and backgrounds and diffidence of texts. Personalized web search (PWS) is a special search technique which is used to provide valuable search results according to user individual needs. In the background, user information has been collected, analyzed, classified and categorized based on the user issued queries.

Each user have different goal while searching on web, but if users use general keyword, search result delivers common result set for all users. So user doesn't reach their specific search result by using general keyword search. Clear and without more user information it is difficult for search engine to find out the user's search context. So in order to overcome these issues and to optimize the search result, users need to provide more user information and personalized search result based on each user specification. To provide better search results for the user search query, Personalized Web Search collect and scrutinize the user information.

This paper explains the comparison metrics like Search Quality, Response Time, Scalability & Performance evaluation between GreedyDp, UPS & UCOPSK frameworks in Personalized Web Search.

II. BACKGROUND KNOWLEDGE

In this section, introduce some important fundamentals and basic terminology of PWS which is an active in research area of data mining, for which many algorithms have been discussed.

Speretta and Gauch [1], explained how user profiles and user interests have been used by search engine to provide better search result in Personalized Web Search. In this user information are collected through proxy servers or desktop bots and analyzed based on concepts and re-rank techniques have been applied for search results.

Chen et al [2] described automatically using genetic algorithm (GA) to retrieve accurate search results while the amount of data increased in a flash.

Halkidi et al [3] described the Recommender systems. Users submit their ratings about the items which have been view by them to the third party, that analyzed the rating and recommend the qualified items based on the users search category.

Xu et al [4] explained how to handle personalized web search, while the amount of data increased rapidly, as if the information grows continuously, raise the difficult for search engine, to find out accurate information based on user search criteria. To overcome these difficulties, this paper explains the scalable, way how to automatically construct the user hierarchical profile based on user interest and the privacy settings.

Bedi et al [5] described the proposed recommender system which uses ontologies to store user information. Recommender system uses databases to store information and for recommendation. But in this proposed system uses ontologies to store and recommend the system by building onthologies and maintain in numerous knowledge domains for the Semantic web applications.

M. Halkidi, I. Koutsopoulos [7] explained to protect privacy while a group of people sharing information across the net. Accordingly, privacy is measured as the similarity between the genuine profile of the group, and that has been observed from the outside. Y. Xu, K. Wang, B. Zhang, Z. Chen [8] explained how to quantify the privacy on profile based on observation weighted version. D. Rebollo-Monedero, J. Forné[9] described entropy technique to normalize the exposed profile.

Both [4] and [14] explained the concept to apply online anonymity for user profiles which have been generated by group of k users, that have been used to find the link between the query generated by multiple users.

Dou et al [17] explained the concept of large-scale framework, which used click based personalization instead of profile, based using the MSN query logs, as profile based is unstable through experiment. It also explained how to improve the search performance in profile based search criteria.

Sugiyama et al [18] described the proposed approach to normalization search results by constructing user profile without user efforts. This is novel method which absorbs the user activity to construct user profile without user effort to provide better search results.

Xu et al [19] proposed method introduced two parameters like detailed profile and building hierarchical profile, based on information hiding by user to set their privacy, which have been utilized by the search engine to improve the search quality, while compared with MSN raking algorithm. Based on the frequency level, this proposed system built the user hierarchical profile to improve the search results.

Teevan et al [20] explained the search algorithm based on users interest and interaction, while dig through the web search. Sun et al. [21] proposed a novel method CubeSVD which used to find out the correlation between the users search query and their click through information.

ODP [24], Wikipedia [23] described how to build the hierarchical user profile based on the frequency of user data and knowledge collections.

Online offerings [25] explained how to overcome the difficulties faced while searching data in news and e-commerce portal which consists of huge amount of data. This paper explained the concept of using user special knowledge and context.

Chaum [26] proposed system used anonymity network which consists of collection of routers which act as anonymizers to carry mail content using public key cryptography, to hide the information regarding the sender & the mail content. The disadvantage of this approach is very time consuming.

Brin [27] explained the concept of personalized page ranking algorithm to better personalized web search. Many web sites use this concept to link the web pages based on the ranking mechanism.

Qiu and Cho [28] proposed method which automatically update the user interest and page ranking mechanism to make the search faster and in personalized way. This page rank based on user frequency and the click through history stored to future search better results.

III. USER CUSTOMIZABLE PRIVACY-PRESERVING SEARCH (UPS) FRAMEWORK

User Customizable Privacy-preserving Search (UPS) framework which provide generalization profile based on user privacy requirement settings. Build on the demarcation of at odds (predicate) metrics, namely personalization software and privacy threaten, for hierarchic user profile. We formulate the trouble of privacy-protect personalized probe as hazard Profile Generalization, with its NP-austerity establishment. To support run time profile generalization two effective algorithm GreedyDP and GreedyIL have been developed. GreedyDP maximizes the discriminating power

and GreedyIL minimizes the information loss. By exploiting a number of heuristics, GreedyIL outperforms Greedy DP significantly. To provide an inexpensive method to the user and to provide stable user profile which used to prevent unnecessary exposure of the profile, each time before create runtime profile client need to decide whether to personalize the query in UPS or not .The drawback of this UPS methods are

- Did not surpass the exactly spare solitude of single use.
- Not obtain entire topics privacy preservation not obtained properly.
- The period entanglement proportion of the system is increased in this system.
- The accuracy rate is lower when compared with other system.

IV. PROPOSED USER CUSTOMIZABLE ONLINE PRIVACY-PRESERVING SEARCH WITH K-ANONYMITY (UCOPSK) METHODOLOGY

In this paper the proposed method User Customizable Online Privacy-preserving Search with K-anonymity (UCOPSK) assumes that the user queries might not contain sensational information, and it protects the privacy information without leaking in personalized web search and retains its design effectiveness. Figure.1 shows and illustrates the procedure of the entire system architecture of UCOPSK. The proposed UCOPSK framework contains two major concepts like online and offline phases for user profile design. In offline phase the hierarchical user profile is constructed based on user privacy customization.

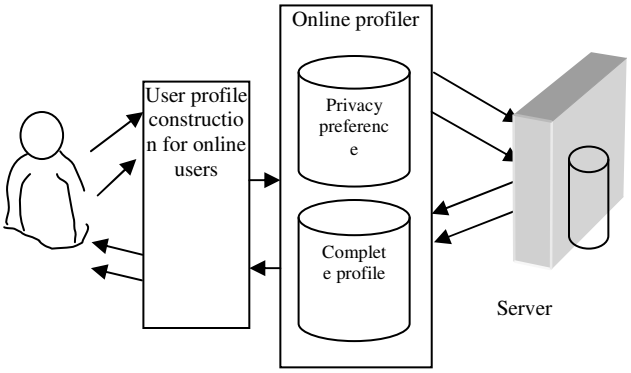

Figure.1. System Architecture of UCOPSK

During the online phase queries are submitted to the user and 4 majors steps are carried out as described below,

1) Once a query q has been submitted to the client, the proxy server automatically generates the generalized user run time profile G based on the two metric privacy settings and personalization convenience.

2) Then the user's submitted queries q and generalized user profile G is send to the PWS to investigate personalization.

3) Then the investigated personalized profile sent back to the proxy server.

4) Finally the proxy server presents the raw results to the user.

The proposed model UCOPSK aims to resolve privacy protection in personalized web search.

Knowledge bounded: In the classification repository the assailant background information is incomplete. In the tree H with user profile, privacy is determined based on classification repository R, Privacy risk of each user is determined based on the total probabilistic to each sensitive nodes, which are initially used to build run time profile.

Intentionally all users need to follow the below steps to solve the privacy protection in personalized web search.

- Profile construction in online and offline phases
- Privacy requirement customization
- Mapping online query-topics, and
- Online generalization

A. Profile construction in online and offline phases

Constructing hierarchical user profile based on user privacy setting playing a vital role in PWS. Constructing profile in offline is easy when compared with online, as in online, user hierarchical profile changes dynamically for each seconds. That is what this proposed method has been discussed for both online and offline profile buildings. Based on the public access point user hierarchical profile has been constructed in UCOPSK framework.

User profile construction in offline and online phases follow the below steps.

- Define the similarity among all users and the active user.
- Based on the similarity pick the related user from amount the group of users.
- Evaluate the prediction based on the weight of group of queries; the highest similarity is the user greatest threshold value.

In online mode sensitive value is calculated based on the K parameter assigned to the user query, which provide K anonymous that provide privacy to individual user for their query. The main focus of this K anonymity is to protect the privacy for the user and their sensitive information of search topics.

While constructing user profile verify the K anonymity depending upon checking each tuples of the search query of the individual user and assign weight and replace the weight

of each topic based on the iteration. If the weight is increased the privacy security level increased for the user search content sensitive information.

B. Privacy Requirement Customization

Users able to set their sensitive information while search their queries, those sensitive nodes are grouped into sensitive topics, which are protected by assigning cost for each sensitive node.

C. Query-Topic Mapping

Query-Topic Mapping is achieved through, identify the topics related to each user specific searched queries in the classification repository R, and obtain the value of the query q among the topics H, assign the value for iterated topics, and construct the root and leaf node, and the corresponding node values based on the hierarchical preferences assigned by the K- anonymity value.

In this proposed method the following metrics have been improved while compared with other methods.

Metric of Utility: The main focus of the utility loss of this method is used to improve the quality of the user search results based on user query q from the generalized profile G. This increases the performance of PWS of user hierarchical profiles.

Metric of Privacy: The main target of privacy loss is to protect the privacy information by analyze the sensitive information of each users queries on a generalized profile. The sensitive nodes have been collected from user during the offline phase and assign cost value for each sensitive node based on the iteration and assign k-anonymity to protect the user sensitive information in Personalized Web Search.

V. COMPARISON RESULT & DISCUSSION

Based on ODP web directory and AOL query log for the 3 month duration period more than 20 million clicks of 650k users the below result have been explained, the comparison between the UPS and proposed method UCOPSK. The format of log file for each user is as follows

<uid; query; time [rank; url]>

Based on the queries requested by each user in the specific time duration, the ranking is calculated based on the frequency of each query, the scalability, quality; response time and effectiveness have been explained below.

This below table explains the metrics evaluation based on no. of queries, clicks and used based between GreedyDp, UPS & UCOPSK methods.

Performance Metrics	***Users***	***Queries***	***Clicks***
Search Quality	600K	20 Millions	30 Millions
Response Time	600K	20 Millions	30 Millions
Scalability	600K	20 Millions	30 Millions
Privacy Threshold	600K	20 Millions	30 Millions

Table 1: User, Query, Clicks based on Performance Metrics

A. Evaluation result for search quality

The main objective of Search Quality displays the relevant search results for the user search queries from the constructed user hierarchical profiles.

Query Set	**Search Quality**		
	GreedyDp	***UPS***	***UCOPSK***
Q1	16.0	18	21
Q2	17.0	19	22
Q3	17.0	19	23
Q4	19.0	22	25

Table2: Search Quality Evaluation

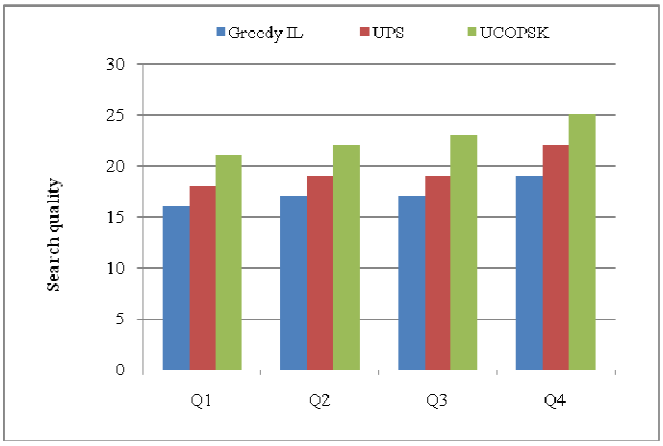

Figure.2. Search Quality Performance Comparison

Figure.2 shows the search quality comparison between the GreedyDP, UPS & UCOPSK, where queries are categories as Q1-Distinct Q2- Medium, Q3- Ambiguous Q4-Very ambiguous and those are denoted in X-axis and searching quality results are plotted in Y-axis. So the proposed method UCOPSK achieves 13% of improvement in search quality than other methods.

B. Effectiveness of personalization on varying threshold

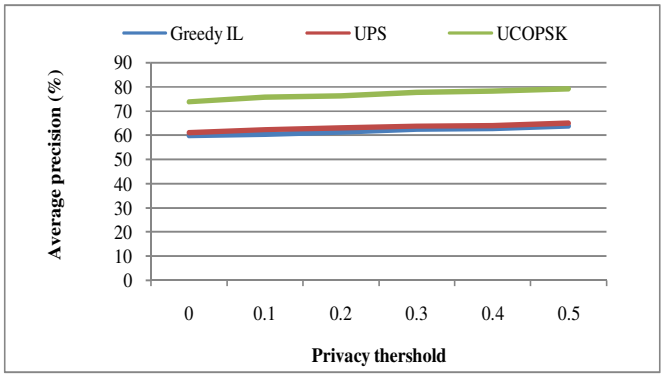

Figure.3. Effectiveness of Personalization on Varying Threshold

Figure.3 shows the performance comparison results of the various schemas by varying the privacy threshold. The Privacy threshold is plotted in X-axis and the average precision is plotted in Y-axis. Based on the privacy threshold value, the AVP varies through admiration to generalization. The UCOPSK achieves 15% of improvement in personalization than other methods.

C. Evaluation result for response time

Response time: Response time means time taken to generalize the user profile based on user search queries on the privacy requirements.

Query Set	Response Time (Sec.)		
	GreedyDP	*UPS*	*UCOPSK*
Q1	12	10	9
Q2	15	13	12
Q3	16	12	11
Q4	8	6	5

Table 3: Evaluation Results for Response Time

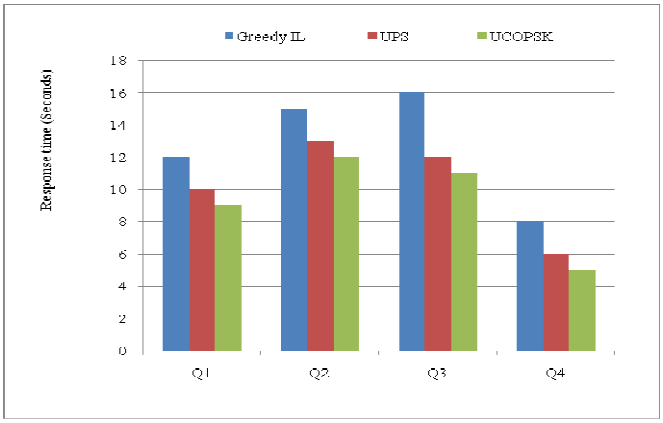

Figure.4. Response Time Performance Comparison

Figure.4 shows the response time between the methods GreedyDP, UPS and UCOPSK where queries are categories as Q1-Distinct Q2- Medium, Q3- Ambiguous Q4-Very ambiguous and those are denoted in X-axis and response time results are plotted in Y-axis. So the proposed method UCOPSK achieves 12% of improvement in response time than other methods.

D. Scalability Evaluation Result

Scalability: Scalability is defined as the system's capability to hold the rising profile size in a proficient manner or its capability to be distended to accommodate that growth.

Profile Size (No. of Nodes)	Average Time (Sec.)		
	GreedyDP	*UPS*	*UCOPSK*
10	5.4	4.85	3.25
20	6.83	6.23	4.26
30	13.5	11.45	10.11
40	14.65	13.14	11.85
50	15.68	14.12	12.58
60	18.94	16.68	14.21

Table 4: Scalability Evaluation Results based on Varying Profile Size

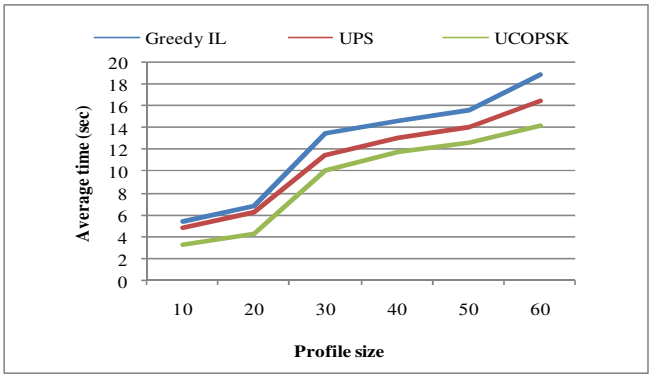

Figure. 5. Performance Comparison based on Profile Size

Figure.5 shows the scalability while varying profile size between GreedyDP, UPS and UCOPSK. The Profile Size is plotted in X-axis and the average time is plotted in Y-axis. The UCOPSK achieves 11% of improvement in scalability than other methods.

VI. CONCLUSION AND FUTURE WORK

This work proposed novel method which is used to provide most relevant information to the users query by providing privacy in constructing hierarchical profile for both static and dynamic users. To protect privacy in user profile the K-anonymity is calculated and applied to user query. It prevents the leakage of the sensitive information of user profile in the distributed environment and verify the authentication and authorization for the published data.

K-anonymity proposed way which is very effective to handle privacy data in a wide usage area of the network, distributed environment. This method used to construct the hierarchical profile based on user privacy setting customization in offline mode, and generalize the user profile based on privacy setting in online, to provide the data for users query without compromising privacy. When compared with online generalization of user profile with GreedyDP, the proposed method UCOPSK provide better quality privacy and scalable search result in personalized web search. The future work need to improve the drawback of this system with broader background knowledge to find out the prolific relationship amount topic and queries;. A most advanced method to build user profile to protect privacy and to improve the performance metric can be developed.

REFERENCES

[1] M. Speretta, S. Gauch "Personalized Search based on User Search Histories", IEEE/WIC/ACM International Conference on Web Intelligence (WI'05). Compiegne University of Technology, France, pp. **622-628** September **2005.**

[2] Y. S. Chen, C. Shahabi, "Automatically improving the accuracy of user profiles with genetic algorithm", Proceedings of IASTED International Conference on Artificial Intelligence and Soft Computing, Cancun, Mexico, pp. **283-288,** May **2001**.

[3] M. Halkidi, I. Koutsopoulos, "A game theoretic framework for data privacy preservation in recommender systems", Proc. European Mach. Learn., Prin. Pract. Knowl. Disc. Databases, ECML PKDD, Springer-Verlag, pp. **629–644, 2011**.

[4] Y. Xu, K. Wang, B. Zhang, Z. Chen, "Privacy-enhancing personalized Web search", Proc. Int.WWWConf., ACM, pp. **591–600, 2007**.

[5] Bedi, Punam, Harmeet Kaur, and Sudeep Marwaha. "Trust Based Recommender System for Semantic Web." In IJCAI, vol. 7, pp. **2677-2682, 2007**.

[6] B. Shapira, Y. Elovici, A. Meshiach, T. Kuflik, "The model for Private Web", J. Amer. Soc. Inform. Sci., Technol., pp. **159–172, 2005**.

[7] M. Halkidi, I. Koutsopoulos, "A game theoretic framework for data privacy preservation in recommender systems", Proc. European Mach. Learn., Prin. Pract. Knowl. Disc. Databases, ECML PKDD, Springer-Verlag, pp. **629–644,** , **2011**.

[8] Y. Xu, K. Wang, B. Zhang, Z. Chen, "Privacy-enhancing personalized Web search", Proc. Int.WWW Conf., ACM, pp. **591–600**, **2007**.

[9] D. Rebollo-Monedero, J. Forné, "Optimal query forgery for private information retrieval", IEEE Trans. Inform. Theory, pp. **4631–4642**, **2010.**

[10] G. Paliouras, "Discovery of web user communities and their role in personalization User Model. User Adapt. Interact", pp. **22(1–2), 151–175**, **2012**.

[11] P. Heymann, G.Koutrika, Garcia-Molina, "Can social bookmarking improve web search?", Proceedings of the international conference on web search and web data mining, WSDM '08, pp. **195–206**, **2008**.

[12] Ding, Shifei, Yanan Zhang, Xinzheng Xu, and Lina Bao. "A novel extreme learning machine based on hybrid kernel function." Journal of Computers 8, pp- **2110-2117**, **2013**.

[13] 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.**

[14] 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.**

[15] J. Castellı´-Roca, A. Viejo, and J. Herrera-Joancomartı´, "Preserving User's Privacy in Web Search Engines" Computer Comm., vol. 32, no. 13/14, pp. **1541-1551, 2009**.

[16] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), **2006**.

[17] Z. Dou, R. Song, & J.R. Wen, "A large-scale evaluation and analysis of personalized search strategies", In Proceedings of the 16th international conference on World Wide Web ACM, **pp.582-590**, May **2007**.

[18] K. Sugiyama, K.Hatano & M. Yoshikawa, "Adaptive web search based on user profile constructed without any effort from users", In Proceedings of the 13th international conference on World Wide Web, pp. **675-684**, May **2004**.

[19] Y. Xu, K. Wang, B. Zhang & Z. Chen, "Privacy-enhancing personalized web search", In Proceedings of the 16th international conference on World Wide Web ACM, pp. **591-600**, May **2007**.

[20] J. Teevan, S. T. Dumais & E. Horvitz, "Personalizing search via automated analysis of interests and activities", In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval ACM, pp.**449-456**, August **2005**.

[21] J. Sun, H. Zeng, H. Liu, Y. Lu, and Z. Chen, "CubeSVD: a novel approach to personalized web search", In Proc. 14th Int. World Wide Web Conference, pp. **382–390**, **2005**.

[22] B. Smyth, M. Coyle, O. Boydell, P. Briggs, E.Balfe, J. Freyne and K. Bradley, "A live-user evaluation of collaborative web search", In Proc. 19th Int. Joint Conf. on AI, **2005**.

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

[24] P. A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschu¨tter, "Using ODP metadata to personalize search", In Proc. 31st Annual Int. ACM SIGIR Conf. on Research and Development in InformationRetrieval, pp. **178–185**, **2005.**

[25] A. Krause & E. Horvitz, "A utility-theoretic approach to privacy in online services", Journal of Artificial Intelligence Research, pp. **633-662**, **2010**.

[26] D. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms", Commun. ACM , **1981**, pp. **84–90**

[27] L. Page, S.Brin, R.Motwani and T.Winograd, "The pagerank citation ranking: bringing order to the web", Technical report, Computer Science Department, Stanford University, **1998**.

[28] F. Liu and J.Cho, "Automatic identification of user interest for personalized search", In Proc. 15th Int. World Wide Web Conference, pp.**727–736**, **2006.**