I J C T A, 9(24), 2016, pp. 167-174 ISSN: 0974-5572
© International Science Press

A novel user profile mining framework for enhancing performance and preserving privacy with a comparative study on non Hadoop and Hadoop approaches

T. Anand*, S. Swamynathan and E. Kirubakaran*****

ABSTRACT

An investigation of mining client profiles from Web log records of Websites is examined with the site clients' distinguishing proof, their post and their change of enthusiasm after some time in non Hadoop and Hadoop environment. The nature of the mined profiles can be investigated with their flexibility despite advancing client conduct. The execution assessment of Non-Big information and Big information methodologies is done.

Index terms: User Profiling, Hadoop and non-Hadoop.

1. INTRODUCTION

Information digging is a stage for mining data from an extensive pool of information and perception. Web digging is utilized for mining the data spread crosswise over web. Client profile mining records client's profile which is utilized for simplicity of future skimming. The First class of web data is the Content information which is organized and exhibited to the end-client and they are basic content, pictures, or organized information. This data can be recovered from databases. The Second class is the Structure information which is the sorted out substance of information elements utilized inside of a Web page, for example, HTML or XML labels and the information elements which is utilized to assemble the Web webpage, for example, hyperlinks interfacing one page to another. The following class is the User profile information, which is utilized as a part of our proposed work, speaks to the demographic data name, age, nation, conjugal status, instruction, intrigues, perusing history and so forth. As needs be the clients' data can be acquired through enrollment by disconnected from the net means like predefined polls or might be gotten physically and computerized later and can be induced by investigating Web use logs. Such an induction from web logs is utilized as a part of the exploration work. The execution of client profile mining is broke down and it is further upgraded by applying Big information innovations like Hadoop and Map Reduce. The preparing time can be decreased to a vast degree and Hadoop Distributed File System(HDFS) is exceptionally a flaw tolerant one and information recuperation is likewise great as numerous duplicates of information are accessible and it overcomes equipment disappointments. Further, the security of the client's profile is saved and a novel methodology is distinguished for accomplishing high privacy.

2. MOTIVATION AND OBJECTIVES

2.1. Motivation

The determination of User profile mining system is a testing undertaking which can be utilized as a part of different applications like : CRM, Marketing and online overviews. Forecast of client conduct is a testing

* Research Scholar, Anna University, *Email: anandavcce@gmail.com.*

** Asso. Professor, Dept of IST, Anna University, Chennai, *Email: swamyns@annauniv.edu*

*** Senior AGM, BHEL, Trichy, India, *Email: e_kiru@yahoo.co.in*

undertaking which can be exceptionally helpful in the above applications. Improvements in Internet and e-trade sites requires the catching of the web clients conduct from the web log, made and kept up naturally Moreover the long range informal communication locales will be utilized prevalently and mining the profiles of the client will be a more beneficial and valuable assignment in which secrecy and security is a noteworthy issue. Removing and keeping up countless profiles will be a testing assignment. As needs be, for enhanced execution of client profile mining, Big information, the Latest Technology pattern, is utilized.

3. OBJECTIVES

In client profile mining, the skimming conduct ease, secrecy and quickly following the inclinations will be the principle issues and it is performed by a novel methodology which is utilized to (1) To facilitate the scanning conduct by foreseeing client inclinations (2) To guarantee and save protection of the clients by keeping client profiles classified (3) To break down the best system for enhancing execution by mining User profiles utilizing Big information procedures.

3.1. Methodology

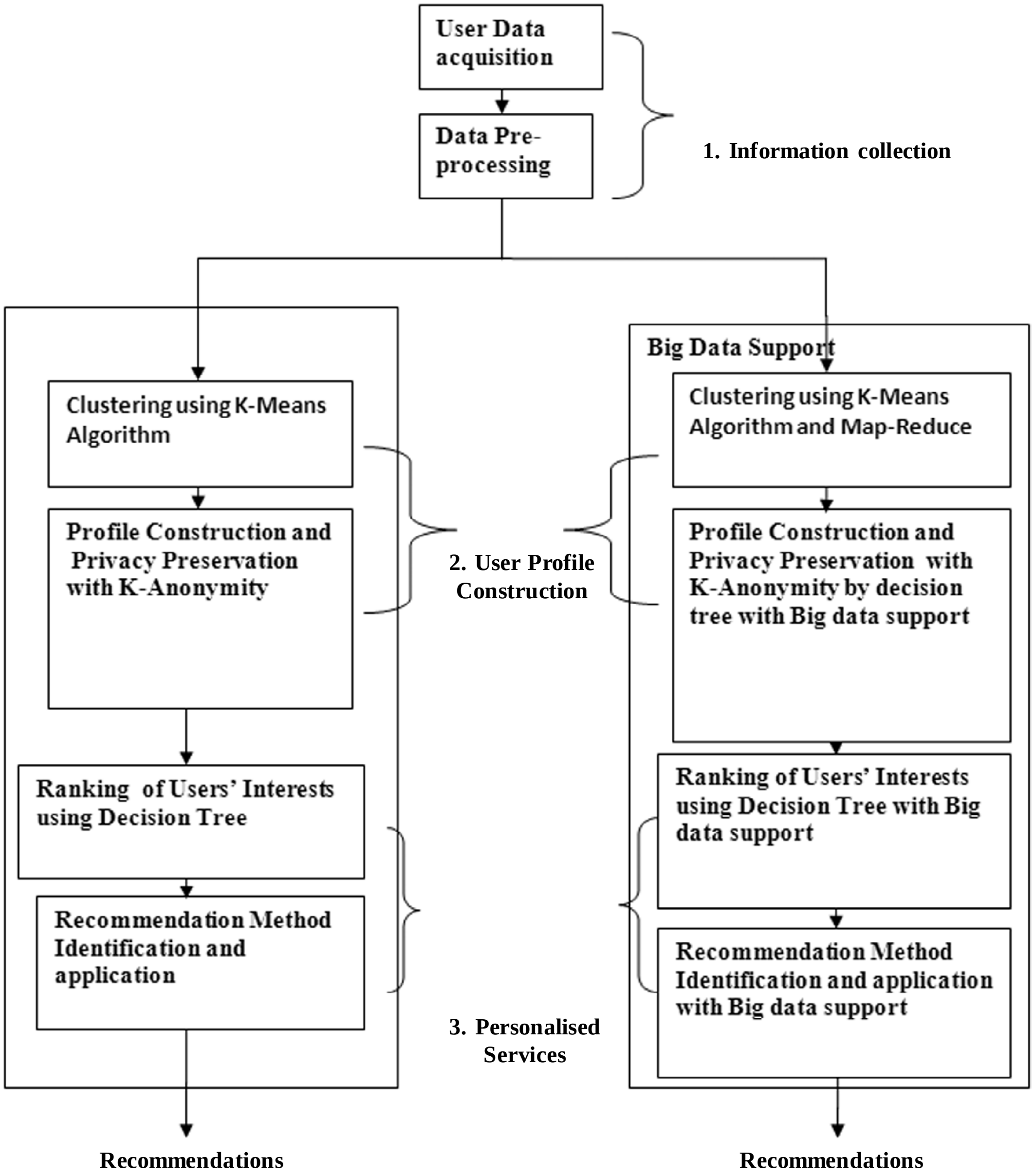
User Data acquisition
Data Pre-processing
1. Information collection
Clustering using K-Means Algorithm
Profile Construction and Privacy Preservation with K-Anonymity
Ranking of Users' Interests using Decision Tree
Recommendation Method Identification and application
Big Data Support
Clustering using K-Means Algorithm and Map-Reduce
Profile Construction and Privacy Preservation with K-Anonymity by decision tree with Big data support
Ranking of Users' Interests using Decision Tree with Big data support
Recommendation Method Identification and application with Big data support
2. User Profile Construction
3. Personalised Services
Recommendations
Recommendations

Figure 1: User Profile Mining Architecture

3.2. DESCRIPTION

A complete system and discoveries in mining client profiles from Web log records of Websites is examined in the proposed research work. Through the work, the site clients' recognizable proof, their post and their change of enthusiasm after some time is investigated. It is depicted, how the found client profiles can be advanced with express data require that is construed from hunt questions separated from Web log information. The nature of the mined profiles can be broke down with their flexibility despite developing client conduct. The execution assessment of Non-Bigdata and Big information methodologies is completed. The system and discoveries in mining Web use designs from Web log documents incorporates advancing client profiles and outer information portraying metaphysics of the Web content. The proposed work is utilized to comprehend the way of the clients, their pursuit, and their changed advantages. The work is done by three primary steps:

3.2.1. Information Collection

The user profile information is collected by:

3.2.1.1. Data Acquisition

The Data sets are obtained from the log document sources (nasa server logs, clarknet server logs, epa server logs and Amazon web administrations and test web shopping information).

For the proposed study, the NASA server web log contains the attributes like User name, Gender, Age, Host, Time stamp, Request method , URL, Status code and Bytes transferred is taken.

3.2.1.2. Data Preprocessing

The preprocessing can be done for the removal of Irrelevant and missing entries. Unuseful Error request's based on status code and Log entries with file extensions GIF, JPEG, JPG and CSS are also removed.

Table 1
Details of Acquired data

Data Source	*Number of Records*	*# of Attributes*	*Sample record*	*Size (Bytes)*
Source 1	100001	8	Jackson 133.43.96.45 01/Aug/1995:00:00:16 -0400] 40 GET /shuttle/missions/sts-69/mission-sts-69.html HTTP/1.0 200 10566	112,623,616
Source 2	47748	7	Neil wpbfl2-45.gate.net [29:23:55:46] 33 GET/information.html HTTP/1.0 617	4,733,650
Source 3	64753	8	Lia amber.RC.Arizona.EDU [28/Aug/1995:10:53:43 -400] 33 GET /theme/cgi-bin/serch.wrl HTTP/1.0 13379 200	2,386,581
Source 4	693	15	2010-03-2 07:00:01 ZZZZC941948879 RUFFLES 222.222.222.222 GET / - 80 - 220.181.7.113 HTTP/1.1 Baiduspider+(+http://www.baidu.com/search/spider.htm) - - www.mysite.com 304 0 0 322 299 299	324,658
Source 5	108	8	117.201.32.108 local host Windows XP Firefox 32.0 India 0:38:16 2014-09-17 17:10:29 user.php –4	40096 bytes

Table 2
Details of pre-processed data

Data Source	*Number of Records*	*Size (MB)*
Source 1	91028	77.4
Source 2	44211	7.1
Source 3	59998	4.0
Source 4	54	0.22
Source 5	10	0.02

3.2.2. User Profile Construction

The Profile of the user is constructed by

3.2.2.1. Clustering

Clustering is one of important technique in data mining process, whose main purpose is to group data of similar types into clusters and finding a structure among unlabelled data. Four different clustering algorithm i.e. K-Means algorithm, Hierarchical algorithm, Density based algorithm, EM algorithm were analyzed.

Table 3
Performance comparison of Clustering algorithms for NASA server logs

Name of the algorithm	# of Clusters	Cluster instances	Square error	Time Taken(s)	#of Iterations	Incorrect cluster instances(%)
Kmeans	2	9059	35431.14	0.27	3	80.28
Hierarchical	2	9059	-	0.28	-	84.02
EM	2	9059	-	132.2	-	86.05
Density Based	2	9059	35431.14	0.38	3	80.42

Based on the time taken and cluster instances, K-Means Clustering algorithm is used.

For Non big data based approach, K-means clustering algorithm is used for clustering the objects and the result shows an improved performance over the other algorithms. For Big data based approach ,Map reduce tasks are used along with K-means clustering algorithm for improving the success rate and reducing time taken using parallel processing and the time taken is 0.12 s in Hadoop Distributed File System(HDFS). Mapping is done for the identification of users and their sessions. While, Reduce is used for aggregating the identified results.

3.2.2.2. Construction of privacy preserved profiles

In the Non Big data based approach, the User profiles are constructed and privacy is maintained by using K-Anonymity algorithm. The same technique is applied with big data support for constructing the user profiles which preserves privacy. Each user is represented by a simple ordered tree, grouping all its queries. The similar user's query is treated as a single tupule for ensuring privacy.

Table 4
Constructed Profiles of various sources

Data Source	*Time taken(NS)*		*File Size(Bytes)*	
	NonBigdata	*Bigdata*	*NonBigdata*	*Bigdata*
Source 1	384370166	78397115	3488006	1744003
Source 2	16155349	3295086	146603	73301
Source 3	8145099	1661295	73914	36957
Source 4	1108016	225994	10055	5027
Source 5	136842	27911	1242	621

Microaggregation is done by grouping the data by means small divisions and the original data being replaced by the centroid value for preserving privacy by maintaining confidentiality.

3.2.3. Personalised Services

The personalisation services includes

3.2.3.1. Ranking

Ranking interests can be used for retrieving a particular one from a large collection. Thus, we choose to rank the concepts according to their importance in the profiles. The page ranking algorithm can be used for retrieving users' preferences from a large collection as it is popular and efficient. Further the user behavior, interest and time is taken for ranking. The ranking result is visualized performed by a Decision tree (J48).

Table 5
Decision Tree(J48) result

Data source	*Weighted TP Rate*	*Weighted FP Rate*	*Weighted Precision*	*Weighted Recall*
Source 1	0.844	0.016	0.751	0.844
Source 2	0.830	0.013	0.724	0.830
Source 3	0.670	0.0100	0.735	0.670
Source 4	0.540	0.011	0.743	0.530
Source 5	0.556	0.556	0.309	0.556

3.2.3.2. Recommendation modelling

The Hybrid approach which combines Content-based and Collaborative approach is used in our system using the combination of the k-Medoids and hierarchical clustering. User mobility and frequent user updates will be the objectives of hybrid systems and it is implemented in the work for personalising the web pages and provides preferences for the regular user. The recommendations is visualised by the Association rule mining. The minimal intra cluster parameter is set and at first there is only one cluster with all objects. The medoid value is calculated or selected randomly and average medoid is calculated next by tentative medoid and the other objects. The users' navigation sessions are divided into frames of navigation sessions based on a time interval as specified earlier. The Association rule is generated as a result of applying Apriori algorithm for various sessions and users which is used for recommendations.

Table 6
Association rule result for NASA server log cluster

Minimum Support	*Confidence*	*# of Cycles*	*Large item sets (L1, L2, L3)*
0.95	0.9	1	3, 3, 1

Accordingly the acquired data is pre processed, a valid user profile is constructed by the usage of clustering with privacy preservation , and ranking techniques for providing recommendations to the user.

4. RESULTS AND DISCUSSION

The overall performance of the user profiling on Big data and Non-Big Data is as follows:

Analysis I: Overall Performance Evaluation based on Time (Nano Seconds)

Table 7
Comparison of 2 approaches

Approach	*Time (Nanoseconds)*
Non-Bigdata	13376373
Bigdata	2728279

The result shows that the Big data based Approach is better than Non-big data approach for getting recommendations from the collective weblogs dataset

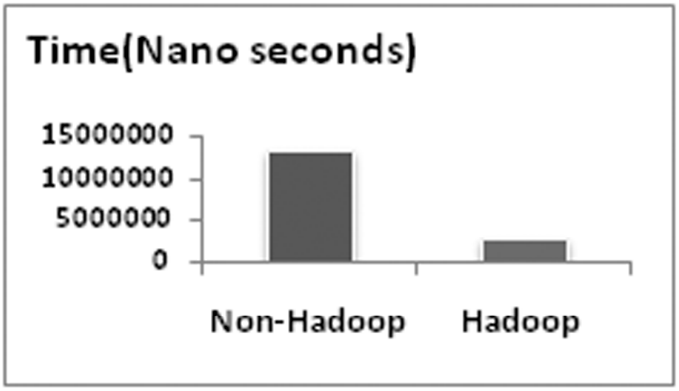

Figure 2: Comparison graph

Analysis II: Performance Evaluation based on Weighted TP Rate, Weighted FP Rate, Weighted Precision and Weighted Recall:

For the Table 6, the web logs of various sources are compared for big data based approach

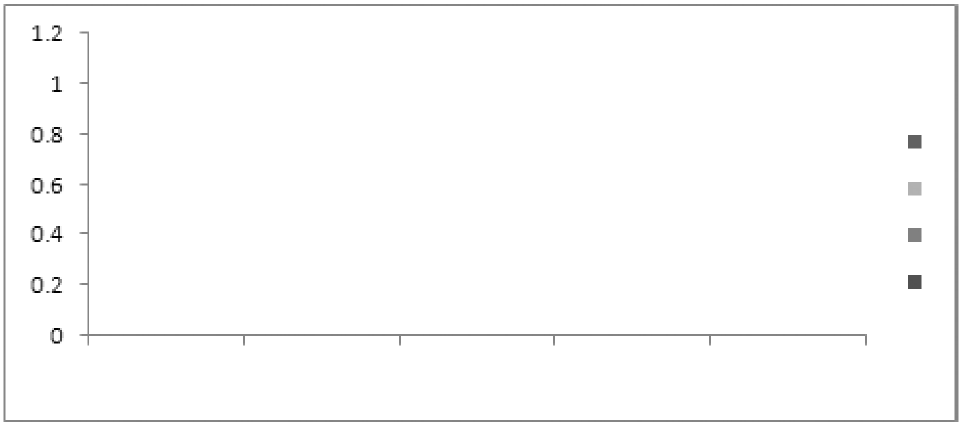

Figure 3: F measure for various data sources

Weighted TP rate is the number of items correctly labeled as belonging to the positive class and **Weighted FP rate** is items incorrectly labeled as belonging to the class

Precision: A summary profile's items are all correct or included in the original input data; that is, they include only the true data items.

$$\text{Precision} = TP / (TP+FP)$$

Coverage /Recall: A summary profile's items are complete compared to the data that is summarized; that is, they include all the data items.

$$\text{Recall} = TP / (TP+FN)$$

Analysis III: Analysis of Preserving privation

The privacy preservation is analysed from table 5, by evaluating the K-Anonymity algorithm on datasets with time taken and file size

5. CONCLUSION

A framework was presented for mining, tracking and validating evolving multifaceted user profiles on Web sites on non Big data and Big Data. User profile mining is performed for predicting user preferences and

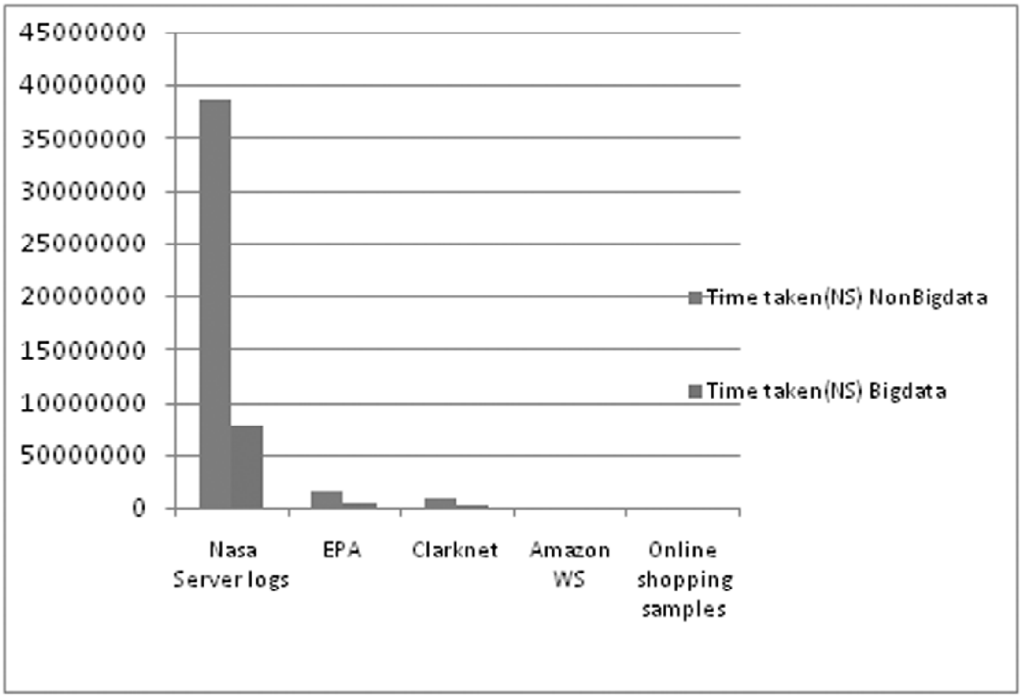

Figure 4: Time taken by various data sources for preserving privacy

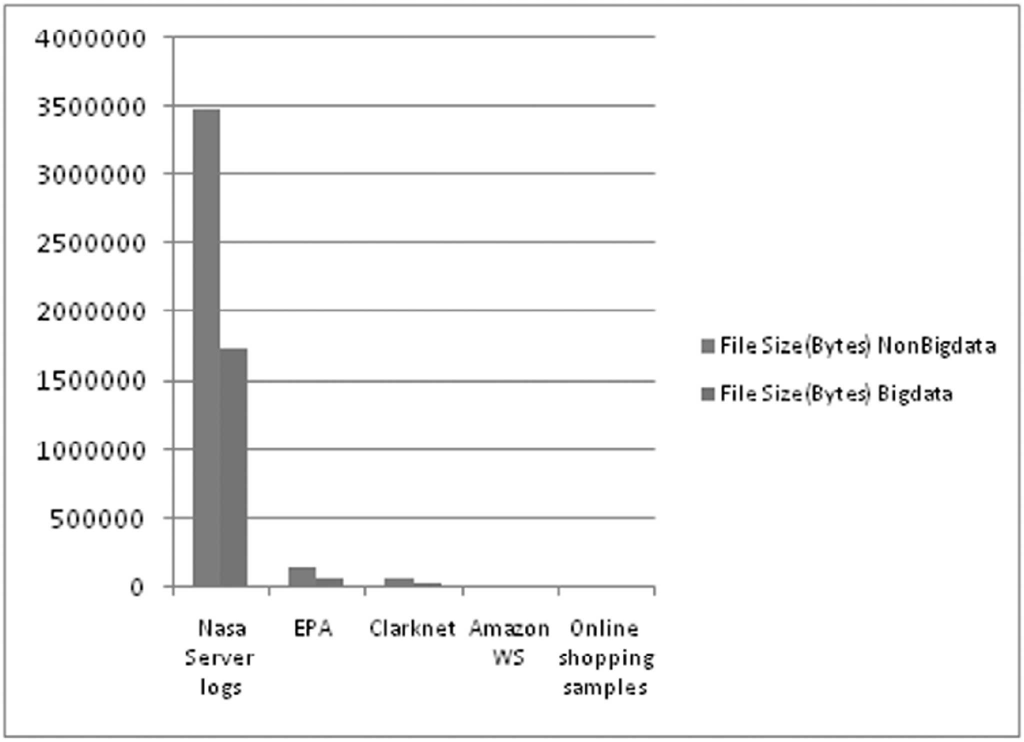

Figure 5: Bytes occupied by the confidential file for Preserving privacy

online surveys. From the evolving user profiles and access patterns, Web pages is presented according to users' preferences. Further the User Privacy is preserved.

6. FUTURE RESEARCH DIRECTION

Complete application of Big data for all the issues in Web mining including structural and content data which can be applied in a variety of applications .

The User profile mining is applied for the distributed cloud computing environment by allowing users to retrieve mining information from virtual area for reducing costs and using the existing infrastructure effectively. This can be possible with the usage of similarity distance measures and distributed algorithms.

REFERENCES

[1] Baglioni, M, Ferrara, U, Romei, A, Ruggieri, S & Turini, F 2003, "Preprocessing and Mining Web Log Data for Web Personalization", AI*IA 2003: Advances in Artificial Intelligence , Springer Berlin Heidelberg, pp. 237-249.

[2] Castellano, G, Fanelli, AM & Torsello, MA 2006, "Mining usage profiles from access data using fuzzy clustering", Proceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization, pp. 157-160.

[3] Evrim Acar & Bulent Yener, 2009, "Unsupervised Multiway Data Analysis: A Literature Survey", IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 1, pp. 6-20 .

[4] Fabrizio Lamberti, Andrea Sanna & Claudio Demartini, 2009, "A relation-based page rank algorithm for semantic web search engines" on IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 1, pp. 123-136.

[5] Feng-hsu Wang & Hsiu-Mei Shao, 2004, "Effective personalized recommendation based on time-framed navigation clustering and association mining" Elsevier Expert systems with applications vol. 27, no. 3, pp. 365–377.

[6] Jose Antonio Iglesias, Plamen Agelov & Ledezma, 2012, "Creating evolving user behavior profiles automatically", IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 5, pp. 854-867.

[7] Ke Zhou, Gui-rong xue, Qiang yang & Yong Yu 2010, "Learning with positive and unlabeled examples using topic-sensitive plsa", IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 1, pp. 46-58.

[8] Kenneth Wai-Ting Leung & Dik Lun Lee, 2010, "Deriving concept-based user profiles from search engine logs" IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 7, pp. 969-982 .

[9] Magdalini Eirinaki & Michalis Vazirgiannis, 2003, "Web mining for Web personalization", ACM Transactions on Internet Technology, vol. 3, no. 1, pp. 1–27.

[10] Michaela Go, Ashwin Machanavajjhala, Guozhang Wang, Xiaokui Xiao & Johannes Gehrke , 2012, "Publishing Search Logs—A Comparative Study of Privacy Guarantees", IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 3, pp. 520-532.

[11] Mohammed Kayed & Chia-hui Chang 2010, "Fivatech: Page-level Web data extraction from template pages", IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 2, pp. 249-263.

[12] Narayan . Bhamidipati & Sankar k Pal 2009, "Comparing scores intended for ranking", IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 1, pp. 21-34.

[13] Nizar R Mabroukeh & Christie I Ezeife, 2009, "Using domain ontology for semantic web usage mining and next page prediction" ACM conference CIKM'09, pp. 1677-1680.

[14] Omar Hasan, Benjamin habegger, Lionel Brunie, Nadia Bennani & Ernesto Damiani 2013, "A discussion of privacy challenges in user profiling with big data techniques: the eexcess use case" , IEEE International congress on big data, pp. 25-30.

[15] Panagiotis Symeonidis, Alexandros Nano poulos & Yannis Manolopoulos 2010, "A unified framework for providing recommendations in social tagging systems based on ternary semantic analysis", IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 2, pp. 179-192 .

[16] Silvia Schiaffino & Analía Amandi 2009, "Intelligent user profiling" Bramer (ed.): artificial intelligence, lnai 5640, pp. 193 - 216, Springer-Verlag Berlin Heidelberg.

[17] Tak-lam Wong & Wai Lam, 2010, "Learning to adapt web information extraction knowledge and discovering new attributes via a Bayesian approach" IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 4, pp. 523-536.