An Improved Approach for Enhancing the Quality of Web Contents by using Web Access Information and Big Data Mining Techniques

Swapna Sahu*, Prof. Anuj Kumar Pal**, Dr. Smita Shandilya*** and Dr. Shishir K.Shandilya****
*-****Bansal Institute of Research and Technology, Bhopal

Abstract: A User personalization is for the purpose of delivering the content information that is related to individual user or groups of individuals in a specified time interval. Where it means to gather user information and improve user experience of support. The present work aims to create an application to deliver quality web contents by using web access log data of the consumer. This aim of this proposed strategy is to enhance the standard of web contents by assessing consumer's behavioural patterns from log associated information. Experimental results reveal that the strategy is more efficient than conventional methods regarding delivery of quality web info.

Keywords: Web Data Mining, Big Data Mining, Web content mining.

Introduction

Interest at the analysis of user behaviour on the Web has been increasing rapidly. This growth comes from the realization that added value for Web site visitors aren't gained only through larger amounts of information on a site, however through easier access to the essential information at the perfect time and at the most suitable form. Web Personalization is simply defined as the task of earning Web-based information systems flexible to the needs and interests of individual users. Typically a personalized Web site recognizes its customers, gathers information about their tastes and adapts its services, in order to match the customers' needs. Web personalization boosts the Web experience of a visitor by introducing the information that the visitor wishes to find in the proper way and at the appropriate time [1].

Web Mining is that place of Data Mining, which deals with the extraction of concealed and interesting knowledge from the large volume of internet documents and documents [2]. It's a comprehensively integrated technique, including Internet, Artificial intelligence, Computer language, informatics, statistics etc.. Web Mining can be broadly separated into three classes [3]: content mining, construction mining and utilization mining. Web content mining is that part of Web Mining, which concentrates on the raw information that can be found on Web pages or the searched results(e.g. words); Web Construction Mining is that part of Web Mining, which concentrates on the construction of Web site including intra-page structural info and inter-page structural information presented on Web pages(e.g., links to other webpages). Web Usage Mining is that part of Web mining, which addresses the extraction of information on users' access patterns and consumer behavior from information collected from the chief sources: Web servers, proxy servers, Web clients (including registration data and user profile data) using some kind of data mining techniques. In Internet usage mining, the focus is on data describing the usage pattern of internet pages, including: Web server side access log files, proxy side log files, client side log records, user registration info, user tips and user request information etc., which may be used to track the behavior, the target and the motivation of users generating these data. Exploiting these usage data can largely help government to identify the taxpayer's or the business' needs, prerequisites, requirements and behaviours etc. and make corresponding policies.

Big Data mining refers to the activity of going through large data sets to Start looking for relevant information. Substantial data samples are available in astronomy, atmospheric science, social networking websites, life sciences, medical science, government data, natural disaster and source management, web logs, mobile phones, sensor networks, scientific research, telecommunications [8]. Two main aims of high dimensional data evaluation are to develop effective procedures that can accurately forecast the future observations and at exactly the same time to gain insight into the connection between the features and answer for scientific purposes. Big data have applications in several fields such as Business, Technology, Health, Smart cities etc.. Substantial data are characterized by 3 V's: Volume, Velocity, and Variety [9].

- **Volume:** The huge scale and rise of dimensions makes it difficult to store and analyse with traditional tools.
- **Velocity:**Big data should be employed to mine great deal of data within a pre-defined time period. The conventional procedures of mining can take huge time to mine such a volume of data.
- **Variety:**Big data comes from a Variety of resources which contains both unstructured and structured data. Standard database systems were created to tackle smaller volumes of Structured and consistent information whereas Big Data

is geospatial information, 3D data, This heterogeneity of unstructured data creates problems for mining, storage and analyzing the data.

Related Work

In a method [4] the KDD stages for the association rules mining the ESOG database is presented which contains educational data. This process produced 127 association rules that could help and guide Greek Educators and School Managers to make educational decisions, design learning activities according their student's interests and efficiently manage the classroom (divide class into groups of students with similar interests, adapt course's content etc.

Conventional web usage mining approaches does not use the semantic information of the web page for pattern generation process. Semantic Web aims to make the web contents more understandable for both humans and computers. An lot of investigation has been done in order to annotate web contents on the basis of semantic information by using ontologies [5].An approach is presented to generate navigation patterns on the basis of Semantic information of the web pages. Sequence association rules are used to generate navigation pattern structure. The quality of generated patterns is then evaluated involving web page recommendation.

Analysis of web site regularities and patterns in user navigation is getting more attention from business and research community a web browsing becomes an everyday task for more people around the world [6]. This extremely large-scaled data called big data are in terms of quantity, complexity, semantics, distribution, and processing costs in computer science, cognitive informatics, web-based computing, cloud computing, and computational intelligence. The size of the collected data about the Web and mobile device users is even greater. Apache Hadoop and other technologies are emerging to support back-end concerns such as storage and processing, visualization based data discovery tools focus on the front end of big data on helping businesses explore the data more easily and understand it more efficiently.

Traditionally a web page is extracted from single web document using linkage method and some regular expressions and matrix model calculation were implemented [7]. The investigation uses searched user content from multiple data item-set using customization linkage processes in order to perform ranking method. By using user personalization technique the content is delivered to individual user based on the characteristics of user, such as (interest, social category, and context) features. In the suggested approach user data retrieval is done using association rule mining, filtered using hybrid filter and classification is done by A-priori fuzzy logic.

Problem Identification

Big data analysis is the process of applying advanced analytics and visualization techniques to large data sets to uncover hidden patterns and unknown correlations for effective decision making. The analysis of Big Data involves multiple distinct phases which include data acquisition and recording, information extraction and cleaning, data integration, aggregation and representation, query processing, data modelling and analysis and Interpretation. Some of major challenges traditional web mining techniques are:

Unstructured Data

Data can be both structured and unstructured. 80% of the data generated by associations are unstructured. They are exceptionally dynamic and does not have particular format. Transforming this information to structured arrangement for later investigation is a major challenge in large data mining.

Incompleteness

Incomplete data generates doubts during data analysis and it must be managed during data evaluation. Doing this properly is also a challenge.

Scalability and Complexity

Traditional software tools are not enough for managing the increasing volumes of data. Data analysis, organization, recovery and modelling are also challenges due to scalability and sophistication of information That needs to be analysed.

The current approach tries to reduce the effect of the above issues stated by making use of web access logs in order to study the user behaviour during web access. The contents are then analysed and efficiently structured using Big Data mining techniques.

Proposed Methodology

The methodology is divided into three major steps as mentioned below:

1. Design and develop the front end for data collection and preparation.
2. Conversion of semi-structured data to structured data.

3. Analysing data using big data analysis approach.

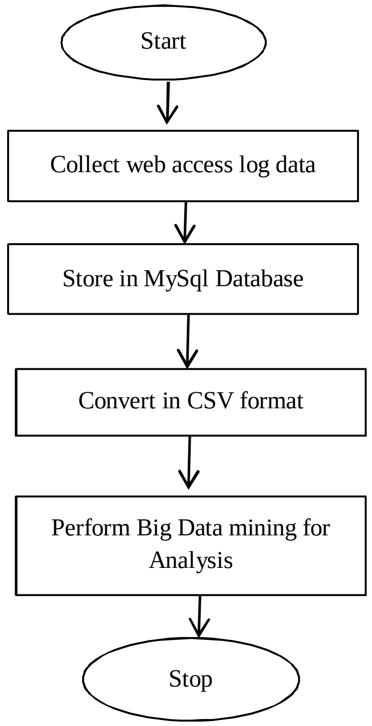

Start
Collect web access log data
Store in MySql Database
Convert in CSV format
Perform Big Data mining for Analysis
Stop

Fig 1: Proposed Methodology

Design and develop the front end for data collection and preparation

A number of sample pages developed like index.jsp, c.jsp ,cpp.jspetcin order to carry out the proposed research. The web pages pages where first designed using Html, CSS and Javascript technologies and logic implementation was carried out by using JSP as server side programming. MySQL is used as a database back end.

Conversion of semi-structured data to structured data

After accessing this web application by different client side through different geographic location the data is stored in mysql data base the structure of our database is given as userId , visitDate, pageId, clientInfo, client_Ip, page_od, page_brw, page_country. The stored data is then converted into CSV format in order to proceed with further analysis.

Analysing data using big data analysis approach

The data stored in CSV format is analysed and mined using Big Data mining approaches. The implementation of mining approaches is carried out using Hadoop technology.

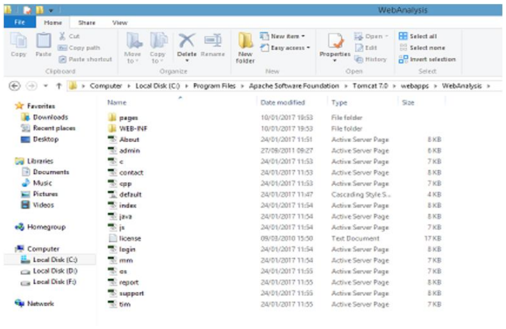
Experimental Results

For the research work platforms used where:

- Notepad++ V6 for writing java program.
- JDK 1.7 for java development environment.
- Windows based Hadoop version 2.3 for Big data mining environment.
- Windows 8.1 operating system environment.

For checking pattern statistics we used news records mashable.com public API as it was available free for development purposes.

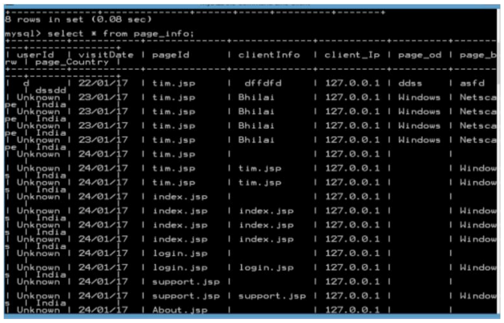
The sample screenshots are as given below:



Name	Date modified	Type	Size
pages	10/01/2017 19:53	File folder	
WEB-INF	10/01/2017 19:53	File folder	
About	24/01/2017 11:51	Active Server Page	8 KB
admin	27/09/2011 09:27	Active Server Page	6 KB
c	24/01/2017 11:53	Active Server Page	7 KB
contact	24/01/2017 11:53	Active Server Page	8 KB
cpp	24/01/2017 11:53	Active Server Page	7 KB
default	24/01/2017 11:47	Cascading Style S...	4 KB
index	24/01/2017 11:54	Active Server Page	8 KB
java	24/01/2017 11:54	Active Server Page	8 KB
js	24/01/2017 11:54	Active Server Page	7 KB
license	09/03/2010 15:50	Text Document	17 KB
login	24/01/2017 11:54	Active Server Page	8 KB
mm	24/01/2017 11:54	Active Server Page	7 KB
os	24/01/2017 11:55	Active Server Page	7 KB
report	24/01/2017 11:55	Active Server Page	8 KB
support	24/01/2017 11:55	Active Server Page	8 KB
tim	24/01/2017 11:55	Active Server Page	7 KB

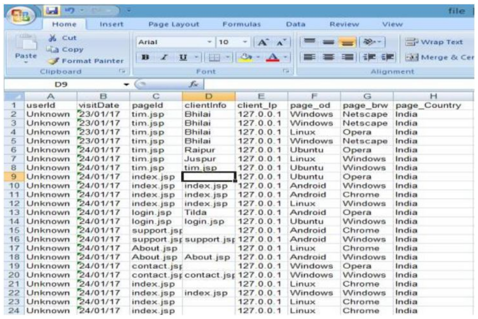
Fig 2: Snapshot of our web application

Fig 3: Output of index.jsp



```
+------------+-------------+---------+-----------+------------+---------+
8 rows in set (0.08 sec)

mysql> select * from page_info;
+------------+-------------+--------------+--------------+-------------+------------+-------
---+--------------+
| userId     | visitDate | pageId       | clientInfo   | client_Ip   | page_od    | page_b
rw | page_Country |
+------------+-------------+--------------+--------------+-------------+------------+-------
---+--------------+
|   d        | 22/01/17  | tim.jsp      |  dffdfd      | 127.0.0.1 | ddss       | asfd
    | dssdd        |
| Unknown | 23/01/17  | tim.jsp      | Bhilai       | 127.0.0.1 | Windows    | Netsca
pe | India        |
| Unknown | 23/01/17  | tim.jsp      | Bhilai       | 127.0.0.1 | Windows    | Netsca
pe | India        |
| Unknown | 23/01/17  | tim.jsp      | Bhilai       | 127.0.0.1 | Windows    | Netsca
pe | India        |
| Unknown | 23/01/17  | tim.jsp      | Bhilai       | 127.0.0.1 | Windows    | Netsca
pe | India        |
| Unknown | 24/01/17  | tim.jsp      |              | 127.0.0.1 |            |
   |              |
| Unknown | 24/01/17  | tim.jsp      | tim.jsp      | 127.0.0.1 |            | Window
s  | India        |
| Unknown | 24/01/17  | tim.jsp      | tim.jsp      | 127.0.0.1 |            | Window
s  | India        |
| Unknown | 24/01/17  | index.jsp    |              | 127.0.0.1 |            |
   |              |
| Unknown | 24/01/17  | index.jsp    | index.jsp    | 127.0.0.1 |            | Window
s  | India        |
| Unknown | 24/01/17  | index.jsp    | index.jsp    | 127.0.0.1 |            | Window
s  | India        |
| Unknown | 24/01/17  | index.jsp    | index.jsp    | 127.0.0.1 |            | Window
s  | India        |
| Unknown | 24/01/17  | login.jsp    |              | 127.0.0.1 |            |
   |              |
| Unknown | 24/01/17  | login.jsp    | login.jsp    | 127.0.0.1 |            | Window
s  | India        |
| Unknown | 24/01/17  | support.jsp |              | 127.0.0.1 |            |
   |              |
| Unknown | 24/01/17  | support.jsp | support.jsp | 127.0.0.1 |            | Window
s  | India        |
| Unknown | 24/01/17  | About.jsp    |              | 127.0.0.1 |            |
```



	A	B	C	D	E	F	G	H
1	userId	visitDate	pageId	clientInfo	client_Ip	page_od	page_brw	page_Country
2	Unknown	23/01/17	tim.jsp	Bhilai	127.0.0.1	Windows	Netscape	India
3	Unknown	23/01/17	tim.jsp	Bhilai	127.0.0.1	Windows	Netscape	India
4	Unknown	23/01/17	tim.jsp	Bhilai	127.0.0.1	Linux	Opera	India
5	Unknown	23/01/17	tim.jsp	Bhilai	127.0.0.1	Windows	Netscape	India
6	Unknown	24/01/17	tim.jsp	Raipur	127.0.0.1	Ubuntu	Opera	India
7	Unknown	24/01/17	tim.jsp	Juspur	127.0.0.1	Linux	Windows	India
8	Unknown	24/01/17	tim.jsp	tim.jsp	127.0.0.1	Ubuntu	Windows	India
9	Unknown	24/01/17	index.jsp		127.0.0.1	Ubuntu	Opera	India
10	Unknown	24/01/17	index.jsp	index.jsp	127.0.0.1	Android	Windows	India
11	Unknown	24/01/17	index.jsp	index.jsp	127.0.0.1	Android	Chrome	India
12	Unknown	24/01/17	index.jsp	index.jsp	127.0.0.1	Linux	Windows	India
13	Unknown	24/01/17	login.jsp	Tilda	127.0.0.1	Android	Opera	India
14	Unknown	24/01/17	login.jsp	login.jsp	127.0.0.1	Ubuntu	Windows	India
15	Unknown	24/01/17	support.jsp		127.0.0.1	Android	Chrome	India
16	Unknown	24/01/17	support.jsp	support.jsp	127.0.0.1	Android	Windows	India
17	Unknown	24/01/17	About.jsp		127.0.0.1	Linux	Chrome	India
18	Unknown	24/01/17	About.jsp	About.jsp	127.0.0.1	Android	Windows	India
19	Unknown	24/01/17	contact.jsp		127.0.0.1	Windows	Opera	India
20	Unknown	24/01/17	contact.jsp	contact.jsp	127.0.0.1	Windows	Windows	India
21	Unknown	24/01/17	index.jsp		127.0.0.1	Linux	Chrome	India
22	Unknown	24/01/17	index.jsp	index.jsp	127.0.0.1	Windows	Windows	India
23	Unknown	24/01/17	index.jsp		127.0.0.1	Linux	Chrome	India
24	Unknown	24/01/17	index.jsp		127.0.0.1	Linux	Chrome	India

Fig 4: Database filled with web usage content

Fig 5: Output of conversion Program

More than 100 records in our database where used for testing and analysis purpose.Analysis was carried out on the basis of Page ID, Client Info and Operating Systems. The results of which are as discussed below:

Analysis on Page ID

In our web application we have 15 jsp pages for accessing our clients. Pages are about.jsp, Admin.jsp, c.jsp, contact.jsp, cpp.jsp, default.jsp, index.jsp, java.jsp, js.jsp, login.jsp, mm.jsp, os.jsp, report.jsp, support, tim.jsp.

Table 1: No of hits for each page

Page ID	Hits
tim.jsp	21
About.jsp	34
Admin.jsp	11
Index.jsp	67
c.jsp	32
contact.jsp	11
cpp.jsp	56
Os.jsp	44
Default.jsp	5
Java.jsp	52
Js.jsp	21
Login.jsp	66
mm.jsp	44
Support.jsp	3
Report.jsp	22

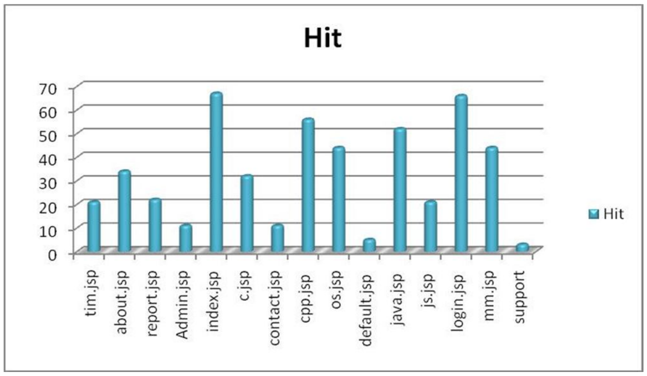


Fig 6: Analysis using Page id

The above analysis shows thatindex.jsp and login.jsp and cpp.jsp are most frequent pages that our client access.

Analysis on Client location Info

Table 2: Client access records location wise

City	Hit	City	Hit
Bhilai	23	Bilaspur	43
Durg	54	Champa	12
Raipur	33	Dongarhgadh	32
Raigarh	45	Tilda	21
Jagdalpur	24	Bhatapara	11

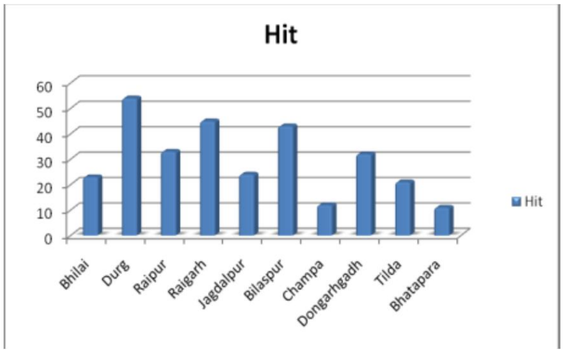


Fig 7: Location wise analysis

Analysis based on Operating System

Table 3: Records OS Wise

OS	HIT
Windows	70
Linux	3
Ubuntu	32
Sun	12
MacOS	11

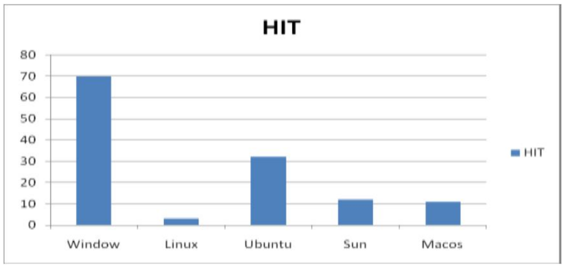


Fig 8: Analysis OS wise

Analysis based on Browsers

Table 4: Records Browser Wise

Browsers	Hits
Firefox	54
Chrome	56
IE	23
Safari	44
Opera	32
Edge	2
Chromium	43
Maxthon	0
SeaMonkey	5
Midori	21
Slim Browser	7
Gnome	1
Rockmelt	4
Surf	1

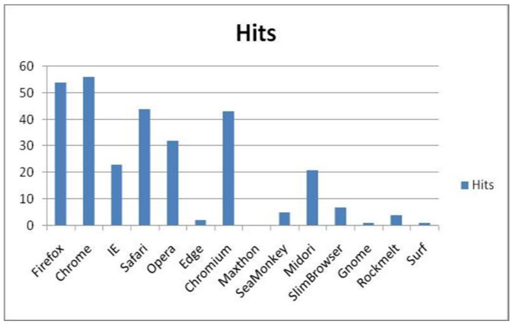


Fig 9: Analysis Browser wise

Conclusion

A lot of research work is being carried out in order to personalize web contents on the user's web usage behaviour. This paper presented an efficient approach for improving the quality of web contents by mining web access log information using Big Data mining approaches. Experimental results show that this approach performs better as compared to traditional web usage mining approaches.

References

[1] Jiawei Han and Michelinekamber. *Data Mining Concepts and Techniques{sQCond*editionX China Machine Press, Bei Jing, 2006).

[2] Federico Michele Facca, Pier Luca Lanzi, Mining Interesting Knowledge from Weblogs: a survey, *Data & Knowledge Engineering* f53), 225 - 241(2005).

[3] K.D. Fenstermacher, M. Ginsburg, Mining Client-side Activity for Personalization, *in: Fourth IEEE International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems* (WECWIS_02), 205 - 212(2002).

[4] StefanosOugiaroglou and Giorgos Paschalis, Association Rules Mining from the Educational Data of ESOG Web-Based Application,Springer-Verlag Berlin Heidelberg 2012.

[5] Pinar Senkul and SuleymanSalin ,Improving pattern Quality in web usage mining by using Semantic Information,Springer-Verlag London Limited 2011.

[6] Pranit B. Mohata and Prof. SheetalDhande, Web Data Mining Techniques and Implementation for Handling Big Data, IJCSMC, Vol. 4, Issue. 4, April 2015.

[7] Santhinisha.E and Dr.P.S.K.Patra, User Personalization In Big Data, International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE), March 2015.

[8] Priya P. Sharma, Chandrakant P. Navdeti, (2014), "Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution", IJCSIT, 5(2), pp2126-2131 .

[9] Richa Gupta, Sunny Gupta, AnuradhaSinghal, (2014), "Big Data:Overview", IJCTT, 9 (5).

[10] Joshua L. Moore, Shuo Chen, Thorsten Joachims, and Douglas Turnbull. Taste Over Time: The Temporal Dynamics of User Preferences. In Proc. ISMIR, 2013.

[11] C. Ding and J. Zhou, ―Log Based Indexing to Improve Web Site Search,‖ Proceedings of the ACM Symposium on Applied Computing, Seoul, Korea, 2007, Mar 11-15, pp. 829 -833

[12] B. Mobasher, R. Cooley and J. Srivastava, ―Automatic Personalization Based on Web Usage Mining,‖ Communications of the ACM, 2000, Vol. 43, pp. 142-151.

[13] A Framework for Web Usage Mining in Electronic Government Ping Zhou, ZhongjianLe School of Information Management, JiangXi University of Finance and Economic, NanChang ,China 330013 Zpjx@126.com, Zhou, P., Le, Z., 2007, in IFIP International Federation for Information Processing, Volume 252, Integration and Innovation Orient to E-Society Volume 2, eds. Wang, W., (Boston: Springer), pp. 487-496.

[14] Data Preparation for Mining World Wide Web browsing Patterns Robert Cooley*, BamshadMobasher, and J aideepSrivastava Department of Computer Science and Engineering University of Minnesota 4-192 EECS Bldg., 200 Union St. SE Minneapolis, MN 55455, USA

[15] Effectual Web Content Mining using Noise Removal from Web Pages. Sivakumar1 Published online: 24 April 2015 _ Springer Science+Business Media New York 2015, Wireless PersCommun (2015) 84:99–121 DOI 10.1007/s11277-015-2596-7

[16] Improving pattern quality in web usage mining by using semantic information Pinar Senkul · SuleymanSalin, KnowlInfSyst (2012) 30:527–541 DOI 10.1007/s10115-011-0386-4 , Received: 19 April 2010 / Revised: 5 January 2011 / Accepted: 6 February 2011 /Published online: 24 February 2011 © Springer-Verlag London Limited 2011

[17] Leung CW, Chan SC, Chung F (2006) A collaborative filtering framework based on fuzzy association rules and multiple-level similarity. KnowlInfSyst 10(3):357–381

[18] Missaoui R, Valtchev P, Djeraba C, AddaM (2007) Toward recommendation based on ontology-powered web-usage mining. IEEE Internet Comput 11(4):45–52

[19] Mobasher B, Cooley R, Srivastava J (2000) Automatic personalization based on web usage mining. Commun ACM 43(8):142–151

[20] A.P. Sheth, C. Ramakrishnan, and C. Thomas, ―Semantics for the Semantic Web: The Implicit, the Formal and the Powerful,‖ Int'l J. Semantic Web Information Systems, vol. 1, no. 1, pp. 1-18, 2005.

[21] Mabroukeh NR, Ezeife CI (2009) Using domain ontology for semantic web usage mining and next page prediction. In: Proceedings of conference on information and knowledge management (CIKM), pp 1677–1680 .

[22] Shahabi, C., Banaei-Kashani, F. and Faruque, J.: 2001,AReliable,E⁄cient,and Scalable System for Web Usage Data Acquisition,In : WebKDD'01Workshop in conjunction with the ACMSIGKDD 2001,Sa n Francisco, CA, August .