Using Domain Ontology and Sequential Rule Mining for Extracting Behavior Patterns from Web Navigation Logs

C. Ramesh[1], Dr. K. V. Chalapati Rao[2] Dr. A. Govardhan[3]
[1]Department of CSE, CVR College of Engineering Hyderabad, India
Email: hmcr.ramesh@gmail.com
[2]Department of CSE, CVR College of Engineering, Hyderabad, India
Email: chalapatiraokv@gmail.com
[3] Department of CSE, School of Information Technology, JNTUH, Hyderabad, India
Email: govardhan_cse@yahoo.co.in

Abstract: **Due to unprecedented growth of information on the Web and lack of structure in many Web sites, it became real challenge to the Web users to find relevant information. To solve this problem, Personalization becomes a popular solution to customize the World Wide Web environment toward the user's preferences. Recent studies show that Web Usage Mining plays an important role in designing recommendation systems. Classical Web Usage Mining does not take Semantics Knowledge into pattern discovery and recommendation process. Recent studies show that Ontology as domain knowledge can improve pattern's quality. Our work aims to incorporate semantics knowledge into Web Usage Mining process. ERMiner, a state-of-the art algorithm for Sequential rule mining is applied over the Semantic space to generate frequent Sequential rules. Experimental results shown are promising and proved that incorporating Semantic Knowledge into Web Usage Mining process can provide us with more quality patterns which consequently make the recommendation system more functional, smarter and comprehensive. The experimental results of our Web recommendation system show a significant improvement on the quality of the recommendations.**

***Index Terms*—Web Usage Mining, Semantic Web, Sequential Rule Mining, Ontology, Semantic Web Usage Mining, Web Personalization, Recommendation System.**

I. INTRODUCTION

With the explosive growth of information on World Wide Web, it has become a real challenge for Web users to access relevant information. One possible approach to solve this problem is Web Personalization [1]. Web personalization [2] is the process of customizing a Web site to the needs of each specific user or set of users, taking advantage of the knowledge acquired through the analysis of the user's navigational behavior.

Web recommender system is a specialized personalization system. Understanding the information needs of users has become a crucial task for Web site owners on the Web. A key requirement in developing successful personalized Web applications is to build user models that can accurately represent users interests and preferences. In addition to the above feature, it has to be machine-understandable and machine- processable.

Though user needs may be elicited in many ways, Usage mining of Web logs is a widely used alternative for understanding usage patterns. However conventional Web usage based recommender systems are limited in their ability to use the domain knowledge of the Web application and their focus is only on Web usage data. As a consequence, the quality of the discovered patterns is low. These patterns do not provide explicit insight into the user's underlying interests and preferences, thus limiting the effectiveness of recommendations as well as the ability of the system to interpret and explain the recommendations [3].

Recent studies[4] hint that Ontology which is an explicit representation of the domain knowledge of the application, if integrated with Web Usage Mining, can enhance the quality of generated usage patterns and help in developing effective Recommendation system.

The combination of Web Usage Mining and Semantic Web has created a new and fast emerging research theme – Semantic Web Usage Mining [5].

The key contributions of our work can be summarized as follows:

1) Feeding domain Ontology into Web Usage Mining Process to extract Sequential navigational patterns.
2) A state-of-the art algorithm ERMiner, is used in the Sequential rule mining process to generate frequent Sequential rules.
3) Generated Sequential rules have antecedent and consequent as sequence of ontological instances instead of mere page views.

The rest of the paper is organized as follows: in section II we review recent advances in Semantic Web Usage Mining research. In section III proposed model and architecture is discussed. Experimental set up and Performance evaluation of the proposed model is presented in section IV. Finally section V provides the concluding remarks and sheds light on future enhancements.

II. RELATED WORK

Web Usage Mining is the process of extracting navigational patterns by applying data mining techniques on Web log file. In recent years, Web Usage Mining techniques such as Clustering, Association rule mining, Sequential pattern mining were employed in extracting

 DOI:10.32377/cvrjst0702

navigational patterns and using those patterns in recommendations [4,5]. An extensive literature on Web Usage Mining based recommendation systems is available in [6,7,8]. However the amount of work presenting the combination of Web Usage Mining and Semantic Web is very limited. Work presented by Stumme et al. [9] and Oberle et al. [10] is regarded as first contribution towards the Semantic Web Usage Mining. The authors have sketched out the benefits of combining Semantic Web and Web Mining. The first part of the work is on extracting semantics from Web page. The second part is on the improvement of Web Usage Mining by using Semantics structures in the form of Ontology. Bamshad Mobasher et al. [11] proposed a unified framework based on Probabilistic Latent Semantic analysis to create user models taking into account both usage data and web site contents. The work presented by stumme et al.[12] sketeched different possibilities of combining Semantic web and Web Mining. In a recent work [13], Nasraoui et al, proposed a Web Usage Mining Framework for mining evolving User Profiles of dynamic Web sites by exploiting the external ontology, used for mapping and relating dynamic Web pages. Eirinaki et al [3], presented a system, SEWeP which integrates the Web usage logs with the semantics of Web site's content to improve the personalization. The innovative feature of the architecture was C-logs, an extended form of Web usage log which encapsulates the site semantics. But the framework was limited only to concept hierarchy. Amit Bose et al, [15] proposed a framework for personalization combining usage information and domain knowledge based on ideas from bioinformatics and information retrieval. Vanzin et al.[16] present ontology – based filtering mechanisms for retrieval of Web Usage patterns and the studies presented by Mehdi et al. [17] proposed a framework XPMiner, which mines frequent patterns over ontology based pattern space. The studies assume that meta data of the web page contents can be typically organized into domain ontology and can be used in frequent pattern mining task. The authors have emphasized the importance of semantic relations in the Mining task.

In summary all the above studies attempted to improve the quality of the navigational patterns and subsequently the recommendations by integrating Semantics into Mining tasks. But the content domain ontologies concerned in the above studies share a common limitation. They invariably represent concept taxonomies. Recent approaches which use Semantics knowledge in the form of ontology for extracting behavior patterns from Web navigation logs are presented by Mabroukeh et al. [18] and Yilmaz et al. [19]. Julia Hoxha et al. [20] presented an approach for the Semantic Formalization of Web browsing behavior across multiple sites. The Usage logs are mapped to comprehensive events from the application domain.

III. PROPOSED SYSTEM

The proposed system has extended the classical Web Usage Mining system. It includes the basic steps such as data acquisition, data preprocessing, extraction of Semantic frequent Sequential rules and Web page recommendation. Web Log file forms the main basis of input to the Web Usage Mining process.

A. *Data Preprocessing:*

Data preprocessing phase includes data cleaning, user identification, and session identification. This task along with Sequential Rule Mining task is implemented as offline phase. Generally several preprocessing tasks need to be performed on the Web access database before user navigational patterns are extracted. Due to large amount of irrelevant information in the Web log file, raw usage data need to preprocessed by applying preprocessing techniques and converted into sequential database.

Initially the usage data logs are centrally stored in raw form as produced upon user interaction. We regard each log record as a browsing event. Log files are pruned to remove the non-responded Web requests and also the requests made by software agents such as Web crawlers, and bots are eliminated. The browsing events are grouped into sessions based on user‘s IP address. Then the browsing events are formalized into Semantic form by mapping the URLs into respective RDF form. The formalized browsing events obtained constitute a semantic rich user model and form the basis of the Semantic Web Usage Analysis.

B. Ontology Construction:

An Ontology is defined as "an explicit specification of a conceptualization"[21]. A conceptualization consists of a set of entities (such as objects and concepts) that may be used to express knowledge and relationships. Protégé [22] tool is used for constructing and editing the ontology of the Web application. The Semantic Web Dog Food (SWDF) [23] and DBpedia Ontology [24] are available publicly. OWL and RDF are the popular Semantic Web technologies used in representing Ontology.

C. Sequential Rule Mining:

After preprocessing step, Web access sequence database, consisting of a sequence of page views is obtained. Since the RDF representation of the web resources for SWDF and DBpedia datasets are available, the ontological instances of the objects of the browsing events are obtained. We employed the procedure presented in [20] for Semantic formalization of browsing events. And further frequent Sequential rules are generated by applying Sequential rule mining.

In the proposed system, we preferred Sequential rule mining over Frequent Pattern Mining, since the sequence information in the navigation is retained in the generated rules. In the Proposed System, ERMiner (Equivalence class based Sequential Rule Miner) [25] algorithm is employed to generate Frequent Sequential rules. It relies on a vertical representation of the database to avoid performing database projection and adopts the novel idea of exploring the search space of rules using equivalence classes of rules having the same antecedent or consequent.

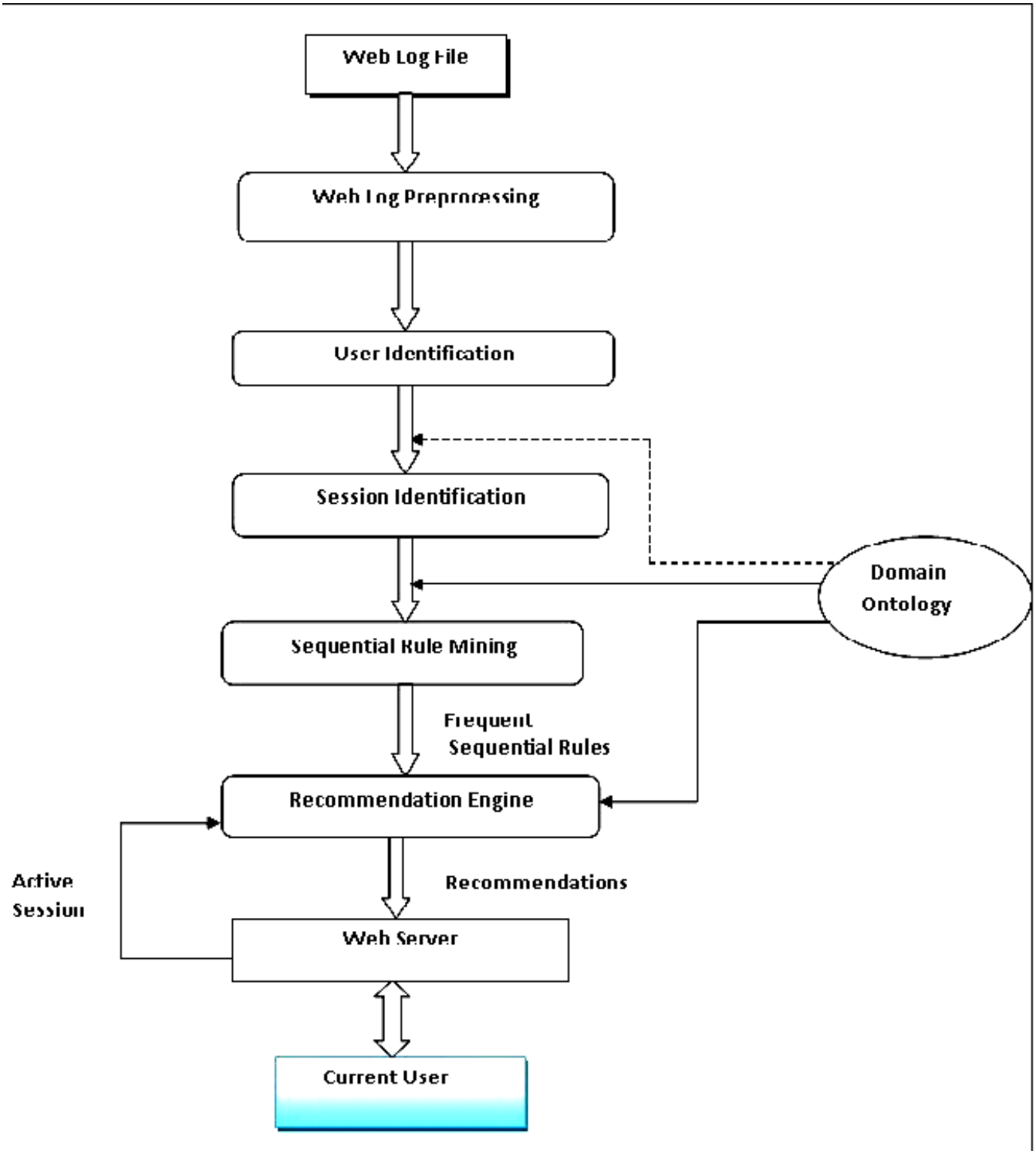

Figure 1 : An Architecture for Online Recommendations using Web Usage Mining and Domain Ontology

DOI:10.32377/cvrjst0702

SPMF [26] is an open source data mining framework implemented in Java. For extracting frequent Sequential rules we have implemented ERMiner algorithm using this framework.

D. *Generating Recommendation*

In recommendation phase, Sequential rules extracted in the above process and active user's navigation session are compared in order to recommend a new page or pages to the user in real time.

Generally, not all the Web pages in the active session path are taken into account while generating a recommendation set. Window count, a parameter which defines the maximum number of previous page visits to be used while generating a recommendation set to the current user is defined.

The recommendation set constitutes the set of Sequential rules which will be used for generating recommendations. After constructing the recommendation set, the Web Page recommendation begins. The Sequential rules in the recommendation set are ordered by their confidence value and the highest one is taken first for the recommendation. For each Sequential Association rule in the recommendation set, its consequent part is extracted and used for recommending Web resources. The ontological instances of consequent part are reverse mapped to page views before applying recommendation

IV. EXPERIMENTS AND PERFORMANCE EVALUATION

A. *DataSet Description :*

Experiments were conducted on two publicly available real datasets SWDF (Semantic Web Dog Food) and DBpedia. SWDF is a very active Web site of publications, people and organizations in the Semantic Web fields, covering several of the major conferences and workshops. DBpedia is shallow, cross domain ontology representing the Wikipedia information in structured format ie.. in the form of classes and properties.

TABLE I
THE SUMMARY STATISTICS OF THE EXPERIMENTAL DATA SETS.

	SWDF	DBpedia
#sessions	890	1020
Avg #sessions/day	150	187
#triples	27790	34870
Period of Usage data	17-07-09 to 22-07-09	21-04-11 to 26-04-11

B. *Evaluation:*

To evaluate the performance of our system we have employed the evaluation metrics discussed in [27]. We have used the measures such as precision and recall. As precision and coverage are inversely related, a combination measure called the F1-measure giving equal weight to both precision and coverage can also be used. Precision measures the degree to which the recommendation engine produces accurate recommendations. Coverage measures the ability of the recommendation engine to produce all of the pages that are likely to be visited by the user.

10 –fold cross – validation is performed on each of the datasets. Each session t in the test session set *ts* is divided into two parts. The first *n* web pages of test session are used for generating recommendations, and the second part is simulated as the future requests (page visits) which are compared with the output of the recommendation system. w is called the window count , which represents the last n pages in the first part of session called active session window (*asw*).

The recommendation engine takes *asw* and the recommendation threshold μ as the input and generates a recommend list which is denoted by *Rec(asw,μ)*. Note that *Rec(asw, μ)* contains all pages whose recommendation score is at least μ. The set of pages *Rec(asw, μ)* can now be compared with the remaining $|t|-n$, pages in *t*. We denote this portion of t by Eval.

$$Precision\ (Rec(asw,\ \mu)) = \frac{|Rec(asw,\mu) \cap Eval|}{|Rec(asw,\mu|}$$

$$Coverage\ (Rec(asw,\ \mu)) = \frac{|Rec(asw,\mu) \cap Eval|}{|Eval|}$$

We performed experiments with recommendation threshold ranging from 0.1 to 1.0.

The results of these experiments are given below.

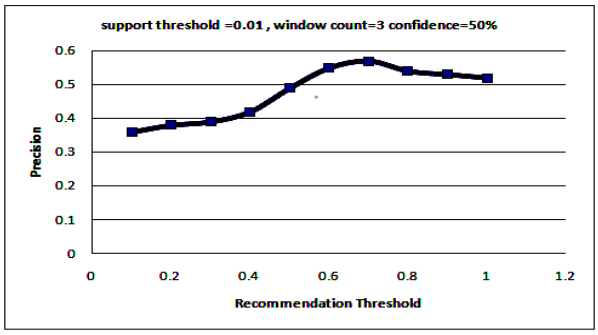

Figure. 2: Recommendation Precision

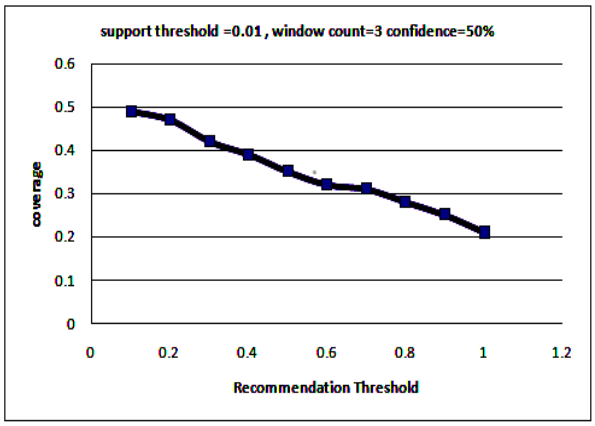

Figure.3 : Recommendation Coverage

V. CONCLUSIONS AND FUTURE WORK

The proposed work extracts interesting Sequential rules using Semantic Web Usage Mining. We have applied an extended version of the state-of-the art Sequential rule mining algorithm ERMiner over Ontological space to extract frequent and interesting Sequential rules. The generated Sequential rules are in terms of ontological instances instead of Web page views. The discovered Semantic Sequential rules form the basis of recommendation engine of the proposed model. Compared with the recommendation system based on classical Web Usage Mining, our proposed model shows promising results.

Experimental results are promising and we believe that the successful integration of Semantic knowledge with Web Usage Mining is likely to lead to the next generation of Personalization tools which will be more intelligent and more useful for Web Users.

Future work includes the development of techniques related to the acquisition of domain ontology, when this is not provided, since it is a crucial component of the Semantic enrichment of usage data with concepts from the application domain. Proposed work can also be extended by combining the clustering and sequential rule mining techniques incorporating domain ontology, making a hybrid recommendation system.

REFERENCES:

1. M. Eirinaki, M. Vazirgiannis, " Web Mining for Web Personalization", ACM Transaction on Internet Technology.Vol. 3(1), pp.1-27, 2003.
2. M. Eirinaki, M. Vazirgiannis and I. Varlamis, (2003) "SEWeP: Using Site Semantics and aTaxonomy to Enhance the Web Personalization Process", *Proc. of the 9th SIGKDD Conf*, 2003.
3. C.Ramesh, K.Chalapati Rao, A.Govardhan, "A Semantically Enriched Web Usage Based Recommendation Model", International Journal of Computer Science and Information Technology (IJCSIT), vol.3,no.5, pp.193-202.
4. B. Zhou, S. C. Hui, and K. Chang, "An intelligent recommender system using sequential Web access patterns," in IEEE conference on cybernetics and intelligent systems, pp. 393–398,2004.
5. Sarabjot Singh Anand and Bamshad Mobasher, "Intelligent techniques for Web Personalization ", LNCS,vol.3169, Springer, 2005.
6. B.Mobasher, R.Cooley, J.Srivastava, "Creating Adaptive Web Sites through Usage-Based Clustering of URLs", In proc. of the IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99),November 1999.
7. Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava, "Automatic personalization based on Web usage mining,"Communications of the ACM, vol. 43, no. 8, pp. 142–151, 2000.
8. M. Perkowitz, O.Etzioni, "Adaptive Web Sites : Conceptual Framework and Case Study", In proc. of WWW,1999.
9. B.Berendt, A. Hotho and G. Stumme, (2002) "Towards Semantic Web Mining", *Horrocks, I.,Hendler, J. (eds.) ISWC 2002, LNCS*, Vol. 2342, pp. 267-278, Springer, Heidelberg (2002).
10. D.oberle, B.Berendt , A. Hotho, and J. Gonzalez, " Conceptual User Tracking". In E.M. Ruiz , J.Segovia , and P.S. Szczepaniak, editors , AWIC ,volume 2663 of Lecture Notes in Computer Science, pages 155 – 164,springer , 2003.
11. Xin Jin, Yan Zan Zhou and Bamshad Mobasher, " A unified Approach to Personalization Based on Probabilistic Latent Semantic Models of Web Usage and Content." In AAAI workshop on Semantic Web Personalization(SWP'04), July 2004.
12. G.Stumme, B. Berendt and A Hotho, (2004) " Usage Mining for and on the Semantic Web",*Data Mining: Next Generation Challenges and Future Directions*", pp. 461-480, AAAI/MIT Press.
13. O.Nasraoui, Maha Soliman, Esin Saka, Antonio Badia and Richard Germain, (2008) " A Web Usage Mining Framework for mining evolving user profiles in dynamic Web sites", *IEEE Trans.Knowl. Data Eng*. Vol. 20, No. 2, pp. 202-215.
14. R.Agrawal and R. Srikant, " Mining Sequential Patterns", In proceedings of the 11th International Conference on Data Engineering.pp 3-14, Taipei, Taiwan,1995.
15. Amit Bose, Kalyan Beemanapalli, Jaideep Srivastava and Sigal sahar, (2006) "Incorporating Concept hierarchies into Usage Mining Based
16. M. Vanzin, K. Becker, and D. D. A. Ruiz. "Ontology-based filtering mechanisms for web usage patterns retrieval". In EC-Web'05, pp. 267-277, 2005.
17. Mehdi Adda, Petko Valtchev, and Rokia Missaoui, "A framework for mining meaningful usage patterns within a semantically enhanced web portal," in Proceedings of the Third C* Conference on Computer Science and Software Engineering C3S2E '10, New York, USA, 2010, pp. 138-147.
18. Nizar Mabroukeh and C.I. Ezeife, (2009) "Using domain ontology for Semantic Web usage mining and next page prediction", Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM), Hong Kong, November 2-6, 2009, pp. 1677-1680.
19. H.Yilmaz and P.Senkul, "Using ontology and sequence information for extracting behavior patterns from web navigation logs". In Data Mining Workshops (ICDMW), 2010 IEEE International conference on – pages 549-556, dec 2010.
20. Julia Hoxha, Martin Junghans, and Sudhir Agarwal, "Enabling Semantic Analysis of User Browsing Patterns in the Web of Datain Julia Hoxha, Martin Junghans, Sudhir Agarwal, Lyon, France, 2012.
21. T.Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing." International Journal of Human Studies.Vol.43,pp.907-928.(1995).
22. http://protege.stanford.edu/
23. http://data.semanticweb.org/ns/swc/swc_2009-05-09.html
24. http://wiki.dbpedia.org/Ontology
25. Fournier-Viger, P., Gueniche, T., Zida, S., Tseng, V. S., "*ERMiner: Sequential Rule Mining using Equivalence Classes*." Proc. 13th Intern. Symposium on Intelligent Data Analysis (IDA 2014), Springer, LNCS 8819, pp. 108-119.
26. http://philippe-fournier- viger.com/spmf/
27. H. Dai and B. Mobasher, (2002) "Using Ontologies to discover domain- level Web Usage profiles", *Proc. of the 2nd Semantic Web Mining Workshop at ECML/PKDD 2002*, Helsinki,Finland, 2002.

DOI:10.32377/cvrjst0702