An Intelligent Web Page Prediction Based On WUM and Domain Ontology Using Key Information Extraction Algorithm

Jyoti B. Patil, Hridaynath P. Khandagale

***Abstract*— Web-page recommendation plays a vital role in intelligent Web systems. Useful knowledge discovery from Web usage data and satisfactory knowledge representation for effective Web-page recommendations are crucial and challenging work. Here we introduce a method to efficiently provide better Web-page recommendation generations through semantic-enhancement by integrating the domain and Web usage knowledge of a website. The models are proposed to represent the domain knowledge. This model uses ontology to represent the domain knowledge. This model uses one automatically generated semantic network to represent domain terms, Web-pages, and the relations between them. Another new model, the conceptual prediction model is proposed to automatically generate a semantic network of the semantic Web usage knowledge, which is the integration of domain knowledge and Web usage knowledge. A number of effective queries have been developed to query about these knowledge bases. Based on these queries, a set of recommendation strategies have been proposed to generate Web-page candidates. A key information extraction algorithm will be developed to compare with the term extraction method in this work, and perform intense comparisons with the existing semantic Web-page recommendation systems.**

***Index Terms*—Web session logs, domain knowledge, navigation pattern, ontology.**

I. INTRODUCTION

An intelligent web-page prediction system plays an important role in intelligent Web systems. Useful knowledge discovery from Web usage data and satisfactory knowledge representation for effective Web-page prediction are crucial and challenging. This work proposes a novel method to efficiently provide better Web-page recommendation through semantic-enhancement by integrating the domain and Web usage knowledge of a website.

The models are proposed to represent the domain knowledge. 1. This model uses ontology to represent the domain knowledge. 2. This model uses one automatically generated semantic network to represent domain terms, Web-pages, and the relations between them. 3. Another new model, the conceptual prediction model, is proposed to automatically generate a semantic network of the semantic Web usage knowledge, which is the integration of domain knowledge and Web usage knowledge.

A number of effective queries have been developed to query about these knowledge bases. Based on these queries, a set of predictions strategies have been proposed to generate Web-page candidates.

Webpage prediction has become increasingly popular, and is shown as links to related stories, related books, or most viewed pages at websites. When a user browses a website, a sequence of visited Web-pages during a session (the period from starting, to existing the browser by the user) can be generated. This sequence is organized into a Web session S = d1d2 . . . dk, where di (i = [1 . . . k]) is the page ID of the ith visited Web-page by the user. The objective of a Web-page recommender system is to effectively predict the Web-page or pages that will be visited from a given Web-page of a website. There are a number of issues in developing an effective Web-page recommender system, such as how to effectively learn from available historical data and discover useful knowledge of the domain and Web-page navigation patterns, how to model and use the discovered knowledge, and how to make effective Web-page recommendations based on the discovered knowledge.

A great deal of research has been devoted to resolve these issues over the past decade. It has been reported that the approaches based on tree structures and probabilistic models can efficiently represent Web access sequences (WAS) in the Web usage data. These approaches earn from the training datasets to build the transition links between Web-pages. By using these approaches, given the current visited Web-page (referred to as a state) and k previously visited pages (the previous k states), the Web-page(s) that will be visited in the next navigation step can be predicted. The performance of these approaches depends on the sizes of training datasets. The bigger the training dataset size is, the higher the prediction accuracy is. However, these approaches make Web-page recommendations solely based on the Web access sequences learnt from the Web usage data. Therefore, the predicted pages are limited within the discovered Web access sequences, i.e., if a user is visiting a Web-page that is not in the discovered Web access sequence, then these approaches cannot offer any recommendations to this user. We refer to

Manuscript received Sept, 2015.

***Jyoti Patil**, Computer Science And Technology, Department of Technology, Shivaji University, Kolhapur, India , 9404423933*

***Hridaynath Khandagale**, Computer Science And Technology, Department of Technology, Shivaji University, Kolhapur, India , 9881014185 .*

ISSN: 2278 – 1323

this problem as "new-page problem" in this study.

This work presents a method to provide better Web-page recommendation based on Web usage and domain knowledge, which is supported by three new knowledge representation models and a set of Web-page recommendation strategies. The first model is an ontology based model that represents the domain knowledge of a website. The construction of this model is semi-automated so that the development efforts from developers can be reduced. The second model is a semantic network that represents domain knowledge, whose construction can be fully automated. This model can be easily incorporated into a Web-page recommendation process because of this fully automated feature. The third model is a conceptual prediction model, which is a navigation network of domain terms based on the frequently viewed Web-pages and represents the integrated Web usage and domain knowledge for supporting Web-page prediction. The construction of this model can be fully automated. The recommendation strategies make use of the domain knowledge and the prediction model through two of the three models to predict the next pages with probabilities
for a given Web user based on his or her current Web-page navigation state. To a great extent, this new method has automated the knowledge base construction and alleviated the new-page problem as mentioned above. This method yields better performance compared with the existing Web usage based Web-page recommendation systems.

II. Related Work

Web-page recommendation plays an important role in intelligent Web systems. Useful knowledge discovery from Web usage data and satisfactory knowledge representation for effective Web-page recommendations are crucial and challenging. This paper proposes a novel method to efficiently provide better Web-page recommendation through semantic-enhancement by integrating the domain and Web usage knowledge of a website [1].

Nowadays more and more people are willing to do B2B transactions over the internet. Semantic Web Mining aims at combining the two fast-developing research areas. In this paper present architecture for integrating semantic information about the products with web log data and generates a list of recommended products by using LCS Algorithm. The implementation shows good performance in terms of precision, recall and F1 metrics [2].

Content adaptation on the Web reduces available information to a subset that matches a user's anticipated needs. Recommender systems rely on relevance scores for individual content items; in particular, pattern-based recommendation exploits co-occurrences of items in user sessions to ground any guesses about relevancy. To enhance the discovered patterns' quality, the authors propose using metadata about the content that they assume is stored in domain ontology. Their approach comprises a dedicated pattern space built on top of the ontology, navigation primitives, mining methods, and recommendation techniques [3].

The flow of information in a Web personalization system can be prone to significant amounts of error and uncertainty. This uncertainty pervades all stages from the user's Web navigation patterns to the final recommendations, including the intermediate stages of logging Web usage, preprocessing and segmenting Web log data into Web user sessions, clustering these sessions, and computing Web user profiles from these clusters. Fuzzy approximate reasoning can offer a general framework for the, recommendation process. It is this framework that is investigated in this paper. This paper presents a simple, intuitive, and fast approach to provide dynamic predictions in the Web navigation space. Real Web usage data is used as a simulation testbed for the fuzzy approximate reasoning based recommendation system [4].

In applying sequence learning models to Web-page recommendation, association rules and probabilistic models have been commonly used. Some models, such as sequential modeling, have shown their significant effectiveness in recommendation generation [7].

On the other hand, by mapping Web-pages to domain concepts in a particular semantic model, the recommender system can reason what Web-pages are about, and then make more accurate Web-page recommendations [12], [13].
Alternatively, since Web access sequences can be converted into sequences of ontology instances, Web-page recommendation can be made by ontology reasoning [11], [14]. In these studies, the Web usage mining algorithms find the frequent navigation paths in terms of ontology instances rather than normal Web-page sequences. Generally, ontology has helped to organize knowledge bases systematically and allows systems to operate effectively.

In order to model the transitions between different Web-pages in Web sessions, Markov models and tree-based structures are strong candidates [7], [16][17][18][19]. Some surveys [20], [21] have shown that tree-based algorithms, particularly Pre- Order Linked WAP-Tree Mining (PLWAP-Mine for short) [18], are outstanding in supporting Web-page recommendation, compared with other sequence mining algorithms.

Furthermore, the integration of PLWAP-Mine and the higher-order Markov model [17] can significantly enhance mining performance [23].The semantic-enhanced approaches integrate semantic information into Web-page recommendation models. By making use of the ontology of websites, Web-page recommendation can be enriched and improved significantly in the systems [23], [24].

In Web usage mining, (WUM) is a type of Web mining, which exploits data mining techniques to extract valuable information from navigation behavior of World Wide Web users. The data should be preprocessed to improve the efficiency and ease of the mining process. Field extraction algorithm performs the process of separating fields from the single line of the log file. Data cleaning algorithm eliminates inconsistent or unnecessary items in the analyzed data [26].

In the systems, domain ontology is often useful for clustering documents, classifying pages or searching

ISSN: 2278 – 1323

subjects. A domain ontology can be obtained by manual or automatic construction approaches, for example, ontologies have been developed for distance learning courses [25], course content [26], personalized e-learning [27], contracts [28], and software [29].

Depending on the domain of interest in the system, we can reuse some existing ontologies or build a new ontology, and then integrate it with Web mining. For example, ontology concepts are used to semantically enhance Web logs in a Web personalization system [29]. In this system, ontology is built with the concepts extracted from the documents, so that the documents can be clustered based on the similarity measure of the ontology concepts. Then, usage data is integrated with the ontology in order to produce semantically enhanced navigational patterns. Subsequently, the system can make recommendations, depending on the input patterns semantically matched with the produced navigational patterns.

Liang Wei and Song Lei [23] employ ontology to represent a website's domain knowledge using the concepts and significant terms extracted from documents. They generate online recommendations by semantically matching and searching for frequent pages discovered from the Web usage mining process. This approach achieves higher precision rates, coverage rates and matching rates.

Web Mining is one of the most propitious fields of Data Mining, which deals with the extraction of meaningful or relevant knowledge from the World Wide Web [30]. the source data is mainly composed of the (textual) logs that are gathered when the users access .the web servers and might be depicted in standard formats; and classic applications are those based on user modeling approaches, namely web personalization, adaptive web sites, and user modeling [31].

Web personalization [19] refers to any action that adapts the information or services provided by a Web site to the needs of a particular user or a set of users by using the knowledge procured from the navigational activities and individual interests of users recorded in the web usage logs, in conjunction with the content and the structure of the Web site. The role of Web personalization system is to provide the users with an information they desire or need, without expecting from them to inquire for it explicitly [31].

III. PROPOSED WORK

An outline of the proposed system architecture is shown in following figure.

Input Dataset The dataset is divided into two sub-sets, one for training and one for testing. The two subsets are pre-processed in the format of WAS.

A. Data Preprocessing

This step is used to extract the useful and relevant information from raw web logs. This raw web logs need to be processed analyzed and converted into proper format of sequential database to mine the weighted sequential patterns. There are no of data preprocessing method are available i.e. stop word removal, stemming etc.

B. Ontology Learning Module:

In the context of Web-page recommendation, the input data is Web logs that record user sessions on a daily basis. The user sessions include information about users' Webpage navigation activities. Each Web-page has a title, which contains the keywords that embrace the semantics of the Web-page. Based on these facts, aim to discover domain knowledge from the titles of visited Web-pages at a website and represent the discovered knowledge in a domain.

C. Semantic Network Analyzer:

This module is design to capture the domain knowledge of a website for supporting Web-page recommendation and to capture the semantics of Web-pages within a website. Also Semantic Network Analyzer is used to solve new page problem in web-page recommendation .Although they are efficient for capturing the domain knowledge and semantics of a given website, they are not sufficient on their own for making effective Web-page recommendations.

D. Domain Term Extraction:

It allows sorting of terms and Web-pages in the queries, so the most possible items are taken into account in the term prediction and Web-page recommendation processes.

Domain Knowledge Extraction fully models terms of Web-pages of user interest, and the relationships between terms and Web-pages. Hence, the interesting Web-pages can be interpreted by the machine.

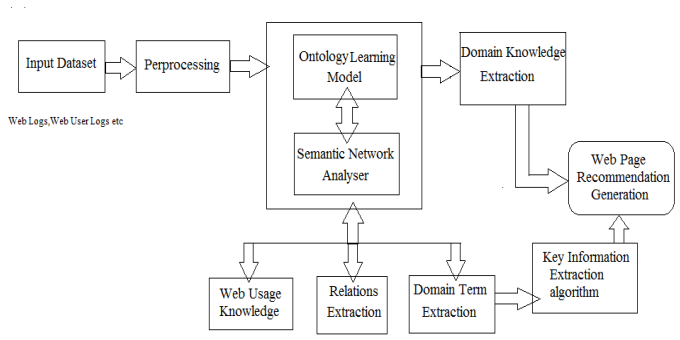


Figure 1: Proposed System Architecture

E. Key Information Extraction Algorithm

It was developed to compare with the term extraction method in this work, and will perform intense comparisons with the existing semantic Web-page recommendation systems.

Extracting key phrases - When extracting key phrases from new information, Key Extraction takes the domain term data file as an input. This file should contain manually assigned key phrases, one per line and feature values for each candidate phrase and computes its probability of being a key phrase. Phrases with the highest probabilities are selected into the final set of key phrases. The user can specify the number of key phrases that need to be selected. Features Extracted during key phrase matching will be TFxIDF, First occurrence, Length, Node degree.

ISSN: 2278 – 1323

F. Web Page Recommendation Generation

The output of the system is predicted web-page, or web pages. The domain ontology based model can improve significantly the performance of web-page predictions using ontology learning model and semantic network analyzer with key information extraction.

IV. IMPLEMENTATION

Following algorithm works based on link dataset and query link. First, it retrieves Log history to form array of links then we retrieve features according to page on link and will compare with that of query link.

Input : Log-History (Dataset), qry_Link)

Output: Recommended Pages

Step 1: link [ ]= load (log_history);

Step 2: for each ll in link

Follow step 3 to step 5

Step 3: Set d_features=load_domain (ll);

Step 4: Set s_feature=load_semantic (ll);

Step 5: Set p_feature[ll]=d_feature+ s_feature;

end for.

Step 6: Set q_d_features=load_domain(qry_link);

Step 7: Set q_s_feature= load_semantic(qry_link);

Step 8: return recommended (p_feature, q_d_feature + q_s_feature)

V. EXPERIMENTAL RESULTS

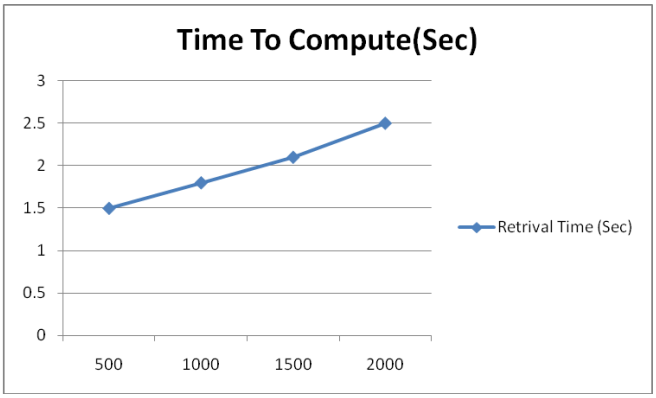
The experimental results perform on varying data set size from 500 links to 200 links to check computation time, precision, recall and accuracy of our algorithm.

TABLE I. NUMBER OF LINKS TO TIME, PRECISION, RECALL AND ACCURACY

No. of links	Computation time (sec.)	Precision	Recall	Accuracy
500	1.5	0.6	0.76	0.58
1000	1.8	0.75	0.65	0.69
1500	2.1	0.79	0.44	0.77
2000	2.5	0.88	0.34	0.91

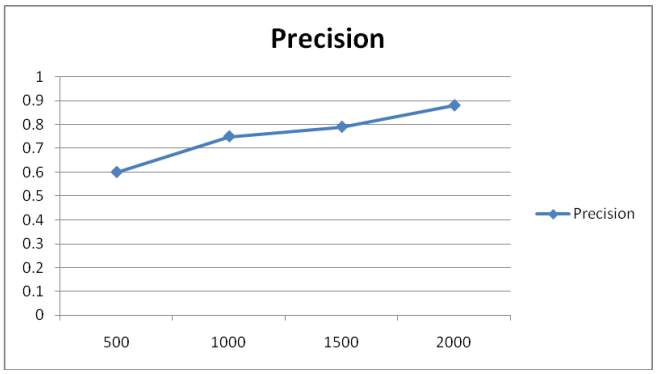
Line chart I shows that, retrieval time increases as increase in number of dataset links.

LINE CHART I. NUMBER OF LINKS TO TIME

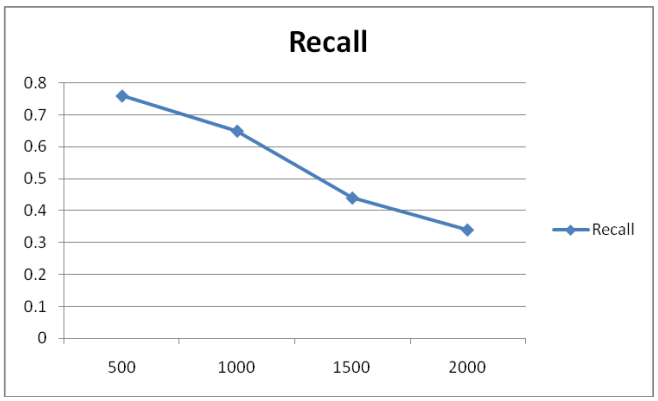


Line chart II and Line chart III shows that, precision and recall of the algorithm improves as increase in number of dataset links.

LINE CHART II. NUMBER OF LINKS TO PRECISION

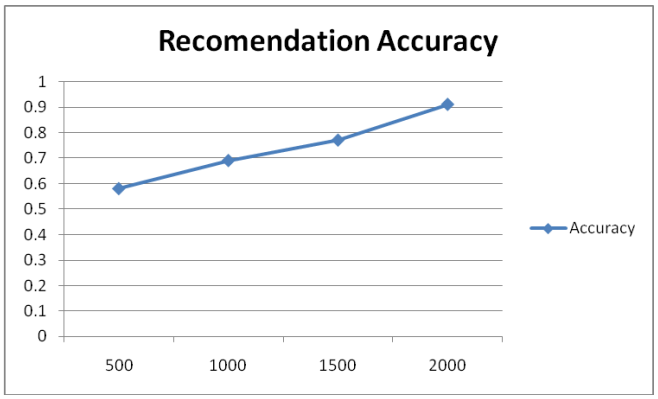


LINE CHART III. NUMBER OF LINKS TO RECALL



Line chart IV shows that, accuracy of the algorithm increases as increase in number of dataset links.

LINE CHART IV. NUMBER OF LINKS TO ACCURACY



VI. CONCLUSION

Finally, we conclude that the web page recommendation performed with the help of domain knowledge and text found in that web page. This domain knowledge will be used to find similarity between user's current page and dataset pages. The experiment shows that accuracy and precision-recall get directly affected by number of records, this also affects computation-cost.

ACKNOWLEDGMENT

Our thanks to the Department of Technology, Shivaji University for allowing us to go ahead with this system. My thanks to all persons who directly or indirectly helped me to

ISSN: 2278 – 1323

complete this work.

REFERENCES

[1] Thi Thanh Sang Nguyen, Hai Yan Lu, and Jie Lu"Web-Page Recommendation Based on Web Usage and Domain Knowledge" IEEE Transaction on Knowledge and data engg.vol 26.no 10 October.
[2] Valtchev, P. ; Missaoui, R. ; Djeraba, C. "Toward Recommendation Based on Ontology-Powered Web-Usage Mining" Internet Computing, IEEE (Volume:11 , Issue: 4) July-Aug. 2007.
[3] J. M. Gascuena, A. Fernandez-Caballero, and P. Gonzalez, "Domain ontology for personalized e-learning in educational systems," in Proc. 6th IEEE ICALT, Kerkrade, Netherlands, 2006,pp. 456–458.
[4] Petenes, C."An intelligent Web recommendation engine based on fuzzy approximate reasoning" Fuzzy Systems, 2003. FUZZ '03. The 12th IEEE International Conference on (Volume:2) 25-28 May 2003.
[5] Mahadevan, G. , Madhura Prakash, M. "An online recommendation system based on web usage mining and Semantic Web using LCS Algorithm" Electronics Computer Technology (ICECT), 2011 3rd International Conference on (Volume:2) 8-10 April 2011
[6] B. Mobasher, "Data mining for web personalization," in The Adaptive Web, vol. 4321, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Berlin, Germany: Springer-Verlag, 2007, pp. 90–135.
[7] G. Stumme, A. Hotho, and B. Berendt, "Usage mining for and on the Semantic Web," in Data Mining: Next Generation Challenges and Future Directions. Menlo Park, CA, USA: AAAI/MIT Press, 2004, pp. 461–480.
[8] H. Dai and B. Mobasher, "Integrating semantic knowledge with web usage mining for personalization," in Web Mining: Applications and Techniques, A. Scime, Ed. Hershey, PA, USA: IGI Global, 2005, pp. 205–232.
[9] S. A. Rios and J. D. Velasquez, "Semantic Web usage mining by a concept-based approach for off-line web site enhancements," in Proc. WI-IAT'08, Sydney, NSW, Australia, pp. 234–241.
[10] S. Salin and P. Senkul, "Using semantic information for web usage mining based recommendation," in Proc. 24th ISCIS, Guzelyurt, Turkey, 2009, pp. 236–241.
[11] A. Bose, K. Beemanapalli, J. Srivastava, and S. Sahar, "Incorporating concept hierarchies into usage mining based recommendations," in Proc. 8th WebKDD, Philadelphia, PA, USA, 2006, pp. 110–126.
[12] N. R. Mabroukeh and C. I. Ezeife, "Semantic-rich Markov models for Web prefetching," in Proc. ICDMW, Miami, FL, USA, 2009, pp. 465–470.
[13] M. O'Mahony, N. Hurley, N. Kushmerick, and G. Silvestre, "Collaborative recommendation: A robustness analysis," ACM Trans. Internet Technol., vol. 4, no. 4, pp. 344–377, Nov. 2004.
[14] G. Stumme, A. Hotho, and B. Berendt, "Semantic Web mining: State of the art and future directions," J. Web Semant., vol. 4, no. 2, pp. 124–143, Jun. 2006.
[15] B. Zhou, S. C. Hui, and A. C. M. Fong, "CS-Mine: An efficient WAP-tree mining for Web access patterns," in Proc. Advanced Web Technologies and Applications. vol. 3007. Berlin, Germany, 2004, pp. 523–532.
[16] J. Borges and M. Levene, "Generating dynamic higher-order Markov models in Web usage mining," in Proc. PKDD, Porto, Portugal, 2005, pp. 34–45.
[17] C. I. Ezeife and Y. Lu, "Mining Web log sequential patterns with position coded pre-order linked WAP-tree," Data Min. Knowl. Disc., vol. 10, no. 1, pp. 5–38, 2005.
[18] B. Zhou, S. C. Hui, and A. C. M. Fong, "Efficient sequential access pattern mining for web recommendations," Int. J. Knowl.-Based Intell. Eng. Syst., vol. 10, no. 2, pp. 155–168, Mar. 2006.
[19] C. Ezeife and Y. Liu, "Fast incremental mining of Web sequential patterns with PLWAP tree," Data Min. Knowl. Disc., vol. 19, no. 3, pp. 376–416, 2009.
[20] T. T. S. Nguyen, H. Lu, T. P. Tran, and J. Lu, "Investigation of sequential pattern mining techniques for Web recommendation," Int. J. Inform. Decis. Sci., vol. 4, no. 4, pp. 293–312, 2012.
[21] S. T. T. Nguyen, "Efficient Web usage mining process for sequential patterns," in Proc. IIWAS, Kuala Lumpur, Malaysia, 2009, pp. 465–469.
[22] L. Wei and S. Lei, "Integrated recommender systems based on ontology and usage mining," in Active Media Technology, vol. 5820, J. Liu, J. Wu, Y. Yao, and T. Nishida, Eds. Berlin, Germany: Springer-Verlag, 2009, pp. 114–125.
[23] A. Loizou and S. Dasmahapatra, "Recommender systems for the semantic Web," in Proc. ECAI, Trento, Italy, 2006.
[24] D. Dzemydiene and L. Tankeleviciene, "On the development of domain ontology for distance learning course," in Proc. 20th EURO Mini Conf. Continuous Optimization Knowledge-Based Technologies, Neringa, Lithuania, 2008, pp. 474–479.
[25] S. Boyce and C. Pahl, "Developing domain ontologies for course content," Educ. Technol. Soc., vol. 10, no. 3, pp. 275–288, 2007.
[26] Aye, T.T. ; Univ. of Compute. Studies, Mandalay, Mandalay, Myanmar Web log cleaning for mining of web usage patterns "Computer (Volume:2)"11-13 March 2011.
[27] Y. Yalan, Z. Jinlong, and Y. Mi, "Ontology modeling for contract: Using OWL to express semantic relations," in Proc. EDOC'06, Hong Kong, China, pp. 409–412.
[28] D. Oberle, S. Grimm, and S. Staab, "An ontology for software," in Handbook on Ontologies, vol. 2, S. Staab and R. Studer, Eds. Berlin, Germany: Springer, 2009, pp. 383–402
[29] M. Eirinaki, D. Mavroeidis, G. Tsatsaronis, and M. Vazirgiannis, "Introducing semantics in Web personalization: The role of ontologies," in Proc. EWMF, Porto, Portugal, 2006, pp. 147–162.
[30] Mulvenna. M. D, Anand. S. S, and Buchner. A. G, "Personalization on the Net using Web Mining", Communications of the ACM, vol. 43, no. 8, pp. 123– 125, August 2000.
[31] C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, "Rotation, scale, and translation resilient public watermarking for images," IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.

Jyoti Patil received B.E. degree in Computer Science and Engineering from Shivaji University, India in 2011. She persuing the M.Tech degree in Computer Science and Technology from Shivaji University, Kolhapur, India in 2015. Her current research interests include Web Mining.

.

Hridaynaynath Khandagale received B.E. and M.Tech degree in Computer Science and Engineering from Shivaji University, India. Now. He is Assistant Professor at Department Of Technology, Shivaji University, Kolhapur. His current research interests include Web page change detection, clustering algorithms.

.

ISSN: 2278 – 1323