International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-2, Issue-4 E-ISSN: 2347-2693

Advance Approach of Integrate Semantic Information Usage Mining for next Page Prediction

Ankit Mahajan[1*] and Megha Singh[2]

[1*,2]Central India Institute of Technology, Indore, India

www.ijcseonline.org

Received: 17/03/2014 Revised: 26/03/2014 Accepted: 24/04/201x4 Published: 30/04/2014

***Abstract*— An** online navigation behavior grows each passing day, and thus extracting information intelligently from it is a difficult issue. Web Usage Mining (WUM) is the process of extracting knowledge from Web user's access data by exploiting Data Mining technologies. It can be used for different purposes such as personalization, system improvement and site modification. In our propose work, user navigation patterns describe as the common browsing behaviors among a group of users. Since many users may have common interests up to a point during their navigation, navigation patterns should capture the overlapping interests or the information needs of these users. In addition, navigation patterns should also be capable to distinguish among web pages based on their different significance to each pattern. We would perfect the algorithm and apply some classification methods for classifying user request. This can be used in WUM based prediction systems. We proposed a complete generic framework that utilizes underlying domain ontology available at web applications. On which any sequential pattern mining algorithm can fit.

***Index Term*—**Ontology, Semantic Distance , Next Page Request Prediction ,Web Prefetching

I. INTRODUCTION

Semantic Web is to address the current web problems by structuring the content of the web, add semantics and extract maximum benefit from the processing power of machines and web. As defined by Sir Tim Berner's LEE, "The semantic web is an extension of the current web in which information is given well distinct meaning, better enabling computers and people to work in co-operation [1]. It is a vision: the thought of having data on the Web definite and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications [2]. Web Mining plays a pivot role in achieving this as it enables to quickly and easily find the information we need. It is mostly for obtain functional information and knowledge from a large number of web pages of websites, and it can be regarded as the data mining continuing to use on the web, which can draw automatically, standardization and analyzing, explaining the data [3]. There are various types of Web mining in which web usage mining is of concern which is used to discover automatic knowledge mining of user access patterns from different web servers. Web usage mining is defined as the extraction of meaningful user patterns from web server access logs using data mining techniques [4]. Web scraping is the process of automatically collecting useful information from web[5].It is also referred as web data extraction and extracts useful information from HTML pages in various ways. It may be performed as text grapping which was performed for Unix originally and may use a scripting language known as Prolog Server Pages(PSP) based on Prolog language where PSP is embedded in HTML language for scrapping HTML pages. An important aspect of semantic web is to add a formal structure and semantics to the textual unstructured or semi-structured content of present web and semantic annotation is a technique which can make it possible for a more useful or efficient information extraction. The semantic annotation provides a more precise description of the knowledge contained in the document and it's semantics in the domain[6][7]. Modeling the user web navigation behavior is becoming the challenging task as the growth of the World Wide Web is increasing rapidly. Web Usage Mining is the field of web mining which deals with finding the interesting usage pattern from the logging information. The logging information is stored in a file known as web log file. Web log file contains lot of information like IP address, date, time, web page requested etc. Web log file can be retrieved from web server, proxy server or client side. This web log contains lot of information so it is preprocessed before modeling. The web log file is preprocessed and converted into the sequence of user web navigation sessions. The web navigation session is the sequence of web page navigated by a user during time window. The user navigation session is finally modeled through a model. Once the user navigation model is ready, the mining task can be performed for finding the interesting pattern. Modeling of web log is the essential task in web usage mining. The prediction accuracy can be achieved through a modeling the web log with an accurate model. We Appling hybrid Markov model is widely used for modeling the user web navigation sessions. The traditional Markov model is having its own limitation. First-order Markov model is less complex but the accuracy is low because of lack of looking into the depth. As we move to the second-order Markov model it is accurate as compared to the first-order Markov model but the coverage of prediction state is less and the time complexity get increased. There are wide application areas of the analysis of user web navigation behavior in web usage mining. The analysis of user web navigation behavior can help for improving the organization of the web site and improvement of web performance by pre-fetching and caching the most

Corresponding Author: *Ankit Mahajan*
Central India Institute of Technology, Indore, India

probable next web page in advance. We proposed hybrid Markov model for Approach of Integrate Semantic Information Usage Mining for next Page Prediction.

II. RELATED WORKS

Sneha Y.S at al[1] in this paper has used OWL technology to add semantics to the existing navigational paths. consequences explain that their approach fetched better accuracy than the existing web based approach. This research they present a framework for integrating semantic information along with the navigational patterns. This research evaluated the framework and it illustrates promising results in terms of quality recommendation of products.

J Vellingiri in at al[2] Examination of user actions in communication with web site can offer insights causing to customization and personalization of a user's web practice. As a consequence of this, web usage mining is of extreme attention for e-marketing and ecommerce professionals. Web usage mining involves of three phases, namely, preprocessing, pattern discovery and pattern analysis. There are different techniques available for web usage mining with its own advantages and disadvantages. This paper provides some discussion about some of the techniques available for web usage mining.

Li Xue at a[3] User Navigation Behavior Mining (UNBM) mainly studies the problems of extracting the interesting user access patterns from user access sequences (UAS), which are usually used for user access prediction and web page recommendation. Through analyzing the real world web data, they find most of user access sequences carrying hybrid features of different patterns, rather than a single one. Therefore, the methods that categorize one access sequence into a single pattern, can hardly obtain good quality consequences. multi-task learning approach based on multiple data domain description model (MDDD), which simultaneously captures correlations among patterns and allowing categorizing one UAS into more than one patterns.

Grau et al.[4] propose the notion of conservative extensions to support partial reuse of ontologies where the objective is to extract from a foreign ontology a small fragment that captures the meaning of terms used in a local ontology. However, determining whether a particular extension is a conservative extension or not is computationally unsolvable and various approximation techniques have to be employed in practice.

Nizar R. Mabroukeh in at al[5]The integration of semantic information, drawn from an essential domain ontology, into probabilistic low-order Markov models they have proposed. Semantic information is infused into the Markov transition probability matrix to change it to a matrix of weights for better-informed prediction, and to overcome the problem of contradict prediction. In this work they have take this idea a step more by proposing to use maximum semantic distance as a compute for pruning higher-order Markov models.

Jawahir Che Mustapha Yusuf in at al[6]in This research they have discuss the architecture of an ontology-based system that consent to practiced users to contact multiple media documents inside heterogeneous storage environments via rich metadata. It is planned to make easy knowledge level interoperability by permit the automated system to give a set of semantically rich ontology-based metadata relating multiple media documents to be exposed, and to be used in disaster management activities.

III. PROPOSED METHODOLOGY

Web usage mining is concerned with finding user navigational patterns on the world wide web by extracting knowledge from web logs. Finding frequent user's web access sequences is done by applying sequential pattern mining techniques on the web log . Its best characteristic is that it fits the problem of mining the web log directly. On the other hand, current sequential pattern mining techniques suffer from a number of drawbacks, some of which include:

(1) Support counting has to be maintained at all times during mining, which adds to the memory size required.

(2) the sequence data base is scanned on nearly every pass of the algorithm or a large data structure has to be maintained in memory all the time

(3) most importantly they do not incorporate semantic information into the mining process and do not provide a way for predicting future user access patterns or, at least, user's next page request, as a direct result of mining. Predicting user's next page request usually takes place as an additional phase after mining the web log. Towards the semantic annotation recognition, the Web pages may embrace metadata or semantic markups/annotations which can be made use of to locate specific data snippets. The annotations may be organized into a semantic layer where they are stored and managed separately from Web pages, so that the Web scrapers can retrieve data schema and instructions from this layer before scraping the pages. Semantic Annotation may be seen as enriching a document by creating a connection between the text uses information extraction technique to build a large knowledge base of annotation in the form of metadata in the form of named entities. referring annotating a web-page depicts the information extraction of the entities that are being given the semantic description. Here, we can simply perform the various searches provided by the Knowledge Infrastructure Management Platform. Without this semantic description or information about the entity, the search cannot be performing.

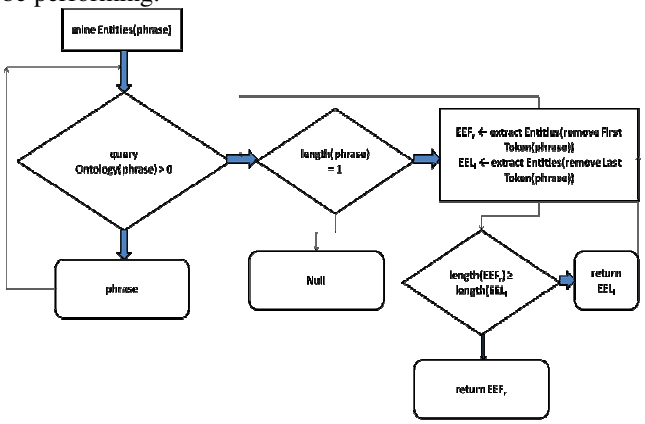

Figure 1: Ontology constriction process

IV. NEXT PAGE PREDICTION

After all knowledge is extracted, the system represents this knowledge into the Semantic Web form and stores it in a semantic knowledge base. For each survey record, we create objects of a customer, survey and location. For fields related to customer profile, the system maps them to the properties in the survey ontology whose domains are class customer. For fields related to geographical locations, the system maps them to the properties in the survey ontology whose domains are class location. The class location and some properties about it are imported from the popular used geographical ontology, GeoNames. For all other fields that are about survey itself are mapped to the properties whose domains are class survey. Finally, the system connects these three objects: customer, survey and location by appropriate properties. In addition to the original fields in a survey, all other knowledge, i.e. concepts and sentiment related terms extracted, is connected to the survey object by using property has Keyword. Thus all knowledge in a survey is fully and semantically represented in the Semantic Web form.There are a number of choices to store these extracted Semantic Web data. Compared to traditional databases or other solutions, RDF store has the biggest advantage that it naturally support heterogeneous survey data. Because when information resources commit to the same ontology then the same meaning is anticipated for any term from that ontology. Even data that commit to different ontologies can be integrated in a information repository, as long as the ontologies have certain relationships, e.g. their concepts are defined in terms of a common ontology or an alignment is provided. Another obvious advantage of a RDF store is that it can give more flexible query support through SPARQL query interface. SPARQL is standardized by the RDF Data Access Working Group of the World Wide Web Consortium, and is considered a key semantic web technology

This research we propose a integrate semantic information, in the form of domain ontology from an e-Commerce application into the pattern discovery and prediction phases of web usage mining, for intelligent and better performing web usage mining As to the classification problem, there are many successful methods, such as Support Vector Machine (SVM), Artificial Neuron Network, Nearest Neighbor Method, etc. However, the preliminary SVM method is only used to solve binary classification problems. Recently, methods are proposed to solve the multi-category problems, such as one-vs.-one and one-vs.-all. which is an extension of data domain description model [6], current methods are unable to capture the inherent co-relation among classes. We apply semantic-aware techniques in this research to pattern-growth sequential pattern mining algorithms. Use the semantic distance matrix as a measure for pruning states in SMM [3]. Investigate concept generalization, and the effect of semantics inclusion on answering more complex pattern queries with improved accuracy. propose a complete generic framework that utilizes an underlying domain ontology available at web applications . on which any sequential pattern mining algorithm can fit. The feasibility of this integration is characterized by the fact that the domain ontology is separated from the mining process. We propose an approach for incorporate semantic information in the heart of the mining algorithm. Such integration allows more pruning of the search space in sequential pattern mining of the web log. Our propose a novel method for enriching the hybrid Markov model with semantic information and solve the problem of trade off between accuracy, complexity in Markov models use for prediction, as well as the problem of ambiguous predictions. Predicting user's next page request on the World Wide Web is a problem that affects web server's cache performance and latency. Different methods exist that can look at the user's sequence of page views, and predict what next page the user is likely to view so it can be perfected. One way is to use association rules as a result of sequential pattern mining [6]. Another way is to model the user's accessed web pages as a Markov process with states representing

The accessed web pages and edges representing transition probabilities between states computed from the given user sequence in the web log. In this case, a trained Markov model can be used to predict the single next state, given a set of k previous states. recently, more businesses on the internet are starting to include domain ontologies in their online applications (e.g. Amazon.com1, eBay2). Domain ontology provides a useful source of semantic information that can be used in next page prediction systems. The availability of this information and the tradeoff problem between state space complexity and accuracy in Markov models [8], trigger a need to integrate semantic information in the mining process.

The integration of semantic information directly in the transition probability matrix of lower order Markov models, was presented as a solution to this tradeoff problem [5]. This integration also solves the problem of contradicting prediction. , we propose to use semantic information as a criteria for pruning states in higher order (where $k > 2$) Selective Markov models [4], and compare the accuracy and model size of this idea with semantic-rich markov models and with traditional Markov models used in the literature

V. CONCLUSION

Semantic annotation in information extraction on web in a better and efficient way. Semantic annotation may be extended to an ontology as a significant semantic annotation Platform We propose generic framework that integrates semantic information into all phases of web usage mining. Semantic information can be integrated into the pattern discovery phase, such that a semantic distance matrix is used in the adopted sequential pattern mining algorithm to prune the search space and partially relieve the algorithm from support counting. We build A 1*st*-order Markov model during the mining process and enrich with semantic information, to be used for subsequently page request prediction, as a solution to ambiguous predictions problem and providing an informed lower order Markov model without the need for complex hybrid order Markov models.

ACKNOWLEDGMENT

We would like to express our gratitude to all those who gave us the possibility to complete this paper. We want to thank the of the Department of Computer Science & Engineering, Central India Institute of Technology, for giving me permission to commence this paper in the first instance, to do the necessary research work and to use departmental data.

We are deeply indebted to our Master of Engineering supervisor Megha Singh from the CSE Department CIIT whose help, stimulating suggestions and encouragement

REFERENCES

[1] Sneha Y.S, G. Mahadevan," Semantic Information and Web based Product Recommendation System – A Novel Approach" International Journal of Computer Applications (0975 – 8887) Volume- 55 Issue-9,Page no.(10-14) October 2012.

[2] J Vellingiri, S.Chenthur Pandian," A Survey on Web Usage Mining" Global Journal of Computer Science and Technology Volume 11 Issue 4 Version 1.0, Page no.(67-72) March 2011.

[3] Li Xue Ming Chen Yun Xiong Yangyong Zhu," User Navigation Behavior Mining using Multiple Data Domain Description" IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Volume 3 Page no.(132-135) September 2010.

[4] B. C. Grau, I. Horrocks, Y. Kazakov, and U. Sattler, "Extracting Modules from Ontologies: A Logic-Based Approach," Modular Ontologies, LNCS 5445, Springer-Verlag, Page no.(159-186) , 2009

[5] N. R. Mabroukeh and C. I. Ezeife. A taxonomy of sequential and web pattern mining algorithms. ACM Computing Surveys, Volume 43 Issue 1 Article 3 March 2011.

[6] Jawahir Che Mustapha Yusuf, Mazliham Mohd Su'ud, Patrice Boursier, Muhammad Alam, " Extensive Overview of an Ontology-based Architecture for Accessing Multi-format Information for Disaster Management" IEEE Page no (294-299), 2012

[7] J. Bao, G. Slutzki, and V. Honavar, "A Semantic Importing Approach to Knowledge Reuse from MUltiple Ontologies," Proc. The 2nd AAAI Conference on Artificial Intelligence, AAAI Press, Page no(1304-1309) ,2007

[8] M. Deshpande and G. Karypis. Selective markov models for predicting web page accesses. Transactions on Internet Technology,Volume 4 Issue 2 Page no(163– 184), 2004.

AUTHORS PROFILE

Ankit Mahajan pursuing M. Tech Computer Science and Engineering in Central India Institute Of Technology in Indore under Rajiv Gandhi Proudyogiki Vishwavidyala University Bhopal.I have completed My B.E Computer Science & Engineering in Jawaharlal Institute of Technology,Khargone under Rajiv Gandhi Proudyogiki Vishwavidyala University Bhopal.Specialized area is Data Mining and Web Usage Mining, Data warehousing.

Megha Singh working as Assistant Professor in Central India InstituteOf Technology in Indore under Rajiv Gandhi Proudyogiki Vishwavidyala University Bhopal.and had 4 year of experience. She completed here M.TECH Computer Science and Engineering in S.O.I.T. in Bhopal RGPV University and completed B.E Computer Science in R.I.T.S under R.G.P.V University published seven papers in International journal. Area of interest is Data Warehousing and Data Mining ,Network Security

