International Journal for Rapid Research in Engineering Technology & Applied Science
ISSN (Online): 2455-4723
Volume II Issue IX October-2016

Survey on Generic Framework That Integrates Semantic Information

Upasana Choudhary[1],Maya Yadav[2]
M.Tech Scholar[1], Assistant Professor[2]
Department of computer science, Sanghvi Institute of Management & Science,Indore,PIN-453332
upasnachoudhary44@gmail.com[1], mayayadav55@gmail.com[2]

Abstract: Identification of the current interests of the user based on the short-term navigational patterns instead of explicit user information has proved to be one of the potential sources for prediction of pages which may be of interest to the user. This would help organizations in various analyses such as web site improvement. Various techniques are employed for achieving personalized recommendation. In this research employs web usage mining techniques for determining the interest of "similar" users, technique for classifying and matching an online user based on his browsing interests. A novel approach for prediction of unvisited pages has been employed. The complete process for next page prediction, represented in the architecture broadly consists of two components: offline component and online component. The offline component involves Data Preprocessing, Pattern Discovery and Pattern Analysis. The outcome of the offline component is the derivation of aggregate usage profiles using web usage mining techniques. The online component is responsible for matching the current user's profile to the aggregate usage profiles. The scope of this work is to match an online user's navigational activity with the aggregate usage profiles obtained through mining tasks and provides suitable page next page prediction, which may be of interest to the user. The recommendation process is an online phase and consists of two sub-phases: matching profile and recommendation.

Keywords:Web Usage Mining, Semantic Web, Domain Ontology, Sequential Pattern Mining, Recommender Systems

1. 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].

Three main concepts for data preprocessing of log files: filtering, normalization, and correlation. Filtering is the act of taking in raw log data, determining if you want to keep it. The output of filtering is a normalized log data. This data is an input to correlation. Correlation is the act of matching a single normalized piece of data, or a series of pieces data, for the purpose of taking an action.

Below is description for each step in the process:

1. *Raw Log Data:* This is what you start with. This is the first input into the process.
2. *Filter:* In the filter stage we look for log messages that we care about and don't care about. The ones we don't care about can be "dropped" in order to reduce load on the overall system. In Figure 9.1, this is shown with an arrow going to an exceptions store. This can be used to review the less-interesting log messages at a later time.
3. *Normalization:* In this step we take the raw log data and map its various elements (source and destination IP, etc.) to a common format. This is important for the correlation step. When a raw log message is normalized, the typical term for what results is an *event*. This term will be used throughout this chapter to denote a normalized log message. Another step in the normalization process is that of categorization. This means that a log message is transformed into a more meaningful piece of information.
4. *Correlation:* Correlation will often lead to groups of individually unimportant events to be flagged. A connection here, a failed login there and an application launch in some other place might mean a system compromise or insider abuse of system privileges. The two basic forms of correlation are rules-based and statistical.
5. *Action:* An action is generally what you do after a correlation has occurred. Figure 9.1 shows several kinds (this list is by no means exhaustive):
    a. *To Analysts:* If you have a log monitoring interface (generally some sort of GUI), this is where you send your high-priority events that require immediate attention.
    b. *Alerts:* This is generally a hybrid of sending an event to an analyst. In this scenario an alert might be a grouping of events which indicate something at a high level has occurred.
    c. *Email:* This can be used as a means to alert on-call staff after hours.
    d. *Long-term Storage:* Long-term storage is where you keep your log data and normalized events. This is a prerequisite for reporting, auditing, long-term analysis, etc.

2. LITERATURE REVIEW

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 ata [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.

Seidenberg[5] developed a methodology for extraction of related concepts from GALEN based on one or more classes given as input by the user. However, it is unclear how the algorithm can be generalized and applied to other ontologies apart from GALEN.

Jalalii in al al[6] put forth a web usage mining technique based on LCS algorithm for online predicting recommendation systems. The Internet is one of the quickly developing fields of intelligence gathering. Through their navigation Web users provide several records of their action. This vast quantity of information can be a helpful resource of knowledge. Advanced mining techniques are required for this information to be extracted, understood and used. Web Usage Mining (WUM) scheme is particularly proposed to perform this process by examining the data indicating usage data concerning a particular Web site.

3. PROBLEM DOMAIN

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 from a number of drawbacks, some of which include:

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

This research we will 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.

4. PROPOSED WORK:

We will 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 will use in the adopted sequential pattern-mining algorithm to prune the search space and partially relieve the algorithm from support counting. we will build A 1st-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 higher order Markov models.

- We will 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 will 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.
- We will propose a novel method for enriching the Markov transition probability matrix with semantic information and solve the problem of tradeoff between accuracy, complexity in

Markov models use for prediction, as well as the problem of ambiguous predictions.

We will perform following task in proposed work:

- Data Preprocessing and Usage Mining
- Pattern Discovery
- Pattern Analysis
- Recommendation Process
- Experimental Design
- Results
- Recommendations

5. Conclusion

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 1st-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.

Reference

[1]B. Mobasher, Robert Cooley and JaideepSrivastava, (2000) " Automatic personalization basedon Web usage mining", Communications of the ACM, 43(8), pp. 142-151.

[2]J Vellingiri, S.ChenthurPandian,"A Survey on Web Usage Mining"Global Journal of Computer Science and Technology Volume 11 Issue 4 Version 1.0 March 2011.

[3] Honghua Dai and BamshadMobasher, (2005) " Integrating Semantic Knowledge with WebUsage Mining for Personalization " , Web Mining: Applications and Techniques, Anthony scime(eds.), IRM Press, Idea Group Publishing, 2005.

[4] 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).

[5]J. Srivastava, R. Cooley, M. Deshpande and P. Tan, (2000) " Web usage mining: Discovery andapplications of usage patterns from Web data" ,SIGKDD Explorations, Vol. 1, No. 2, pp. 12-23, 2000.

[6]Ezeife, C. I. and Lu, Y. (2005). Mining web log sequential patterns with position codedpre-order linked wap-tree. Data Mining and Knowledge Discovery, 10(1):5{38. 5, 8, 10,14, 17, 36, 61, 64

[7]L. Wei and S. Lei, (2009) " Integrated Recommender Systems Based on Ontology and UsageMining" , Active Media Technologies, 5820, Springer-Verlag, Berlin Heidelberg, pp. 114-125,2009.

[8]Middleton, S. E., Roure, D. D., and Shadbolt, N. R. (2009). Ontology-based recommendersystems. In Staab, S. and Studer, R., editors, Handbook on Ontologies, InternationalHandbooks Information System, pages 779 796. Springer Berlin Heidelberg.

[9]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– No.9, October- 2012.

[10]Amit Bose, KalyanBeemanapalli, JaideepSrivastava and Sigalsahar, (2006) " IncorporatingConcept hierarchies into Usage Mining Based Recommendations " , Proceedings of WEBKDD'06, Pennsylvania.

[11]Li Xue Ming Chen Yun XiongYangyong Zhu," User Navigation Behavior Mining using Multiple Data Domain Description" IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-2010.

[12] H. Dai and B. Mobasher, (2002) "Using Ontologies to discover domain- level Web Usageprofiles", Proc. of the 2nd Semantic Web Mining Workshop at ECML/PKDD 2002, Helsinki,Finland, 2002.

[13] J. Seidenberg, "Web Ontology Segmentation: Extraction, Transformation, Evaluation," Modular Ontologies, LNCS 5445, Springer-Verlag, 2009, pp. 211-243.

[14]Facca, F. M. and Lanzi, P. L. (2005). Mining interesting knowledge from weblogs: a survey.Data and Knowledge Engineering, 53(3):225{241. 8, 14

[15]NizarMabroukeh and C.I. Ezeife, (2009) "Using domain ontology for Semantic Web usagemining and next page prediction", Proceedings of the 18th ACM Conference on Information andKnowledge Management (CIKM), Hong Kong, November 2-6, 2009, pp. 1677-1680.

[16]Miki Nakagawa and BamshadMobasher, (2003) " Impact of site characteristics onRecommendation Models Based on Association Rules and Sequential Patterns", Proceedings ofthe IJCAI'03 Workshop on Intelligent Techniques for Web Personalization, Acapulco, Mexico,August 2003.

[17]F. Khalil, J. Li, and H. Wang. A framework for combiningmarkov model with association rules for predicting web pageaccesses. In Proceedings of the Fifth Australasian DataMining Conference (AusDM2006), pages 177–184, 2006.