Web Page Annotation Using Web Usage Mining and Domain Knowledge Ontology

PoonamP.Chavan
ME II (Computer),
LokmanyaTilak College of Engineering,
New Mumbai, India
Email:poonampchavan@gmail.com

Prof.SonalBankar
Computer Dept.
LokmanyaTilak College of Engineering,
New Mumbai, India
Email:sonal.bhople@gmail.com

Abstract- Today's world the growth of the WWW has increased tremendously, the user is totally relying on web for information. Search engine provides the result pages to the user but all are not relevant so the challenging task is extracting the pages from web and provide to the user. WUM is an approach to extract knowledge and use it to the different purposes. In this paper new semantic approach is proposed based on WUM and Domain Knowledge Ontology. Ontology database preparation, it is also challenging task in this project.

Keywords - *WUM, Ontology, Domain Knowledge, Semantic.*

*****

I. INTRODUCTION

In today's world the growth of the e-commerce site increasing day by day so the web-page recommendation plays an important role in the site's final success. Accessing relevant information efficiently is the need of user for his satisfaction. Web Usage Mining (WUM) is the approach to extract the knowledge from analysis of web usage data about a particular website. This usage data can be obtained from server logs. WUM captures models and analyzes the behavioral patterns and profiles those interact with the web sites. This analyzed data is beneficial and can be used for different purposes such as web personalization, recommender systems, presentation of promotional contents etc.

Semantic Web focuses on making the contents of the web site understandable not only by humans but also by computers. To accomplish this, it helps the software agents to look for expected contents. Hence increase in the efforts in annotating Web pages and objects in form of semantic information using ontologies (such as product catalogs or concept hierarchies) are observed. Ontological instances can be built by using Web site specific domain knowledge. Hence, in this work,

Data cleaning is another important subtask of data pre-processing which consists of removing extraneous references to embedded objects which may not be important for analysis. For e.g. references to style files, semantic information of a web site can be combined with the patterns generated by conventional Web usage mining to generate frequent navigation patterns enriched with semantic information of the Web pages.

II. LITERATURE REVIEW

In literature various data mining techniques are being used to model and understand the Web user activity based on web usage [3], [X]. Web Usage Mining processes are divided into three inter dependent stages: pre-processing, pattern generation and pattern analysis.The pre-processing stage consists of cleaning the clickstream data obtained from server logs and partitioning into set of user transactions to represent individual users' activity and frequent usage patterns.

Data preparation process involves pre-processing the original data and transforming into a form to cope up with a specific data mining operation. The task of pre-processing is very important because the success of data mining depends on it. Basically the data sources used for Web usage mining are the server log files. This log data collected automatically by Web servers signifies the navigational behavior of visitors [11].

techniques such as association rules and sequential patterns.

Sequential pattern mining is used to find inter-session patterns. These patterns have the property that presence of set of items is followed by another item in sessions which are ordered with respect to time. Broadly, this approach is graphics or sound files. Data cleaning also involves removal of references due to crawler navigations.

1. *Web Usage Mining*

Web usage mining is the application of data mining methods for analyzing recordings of the use of the website, especially in the form of web server logs. A central problem is to be found in a large number of models in the standard model identified the following as the most interesting. The explosive growth of information sources available on the World Wide Web, it has become necessary for organizations to discover ways to gain an advantage over competitors to examine the patterns found.

Jespersen et al. [12] proposed a hybrid approach for analyzing the visitor click-stream sequences. A combination of hypertext probabilistic grammar and click fact table approach is used for Weblog mining that could also be used for general sequence mining tasks. Mobasher et al. [13] proposed the web personalization system, the offline tasks made mining related, if usage data and online process of automatic Web page customization based on the discovered knowledge. LumberJack by Chi et al [14] are user profiles built by combining both clustering of user sessions and traditional statistical traffic analysis with k-means algorithm.

2. *Pattern Generation*

Web usage mining is the application of data mining techniques to discover ways to use the website data, to understand and to better serve the needs of Web-based applications[11][15]. Web usage mining consists of three phases, namely pre-processing, pattern recognition and pattern analysis.R. Cooley et al. [16] propose a web mining process can be divided into two main sections. The first part of the transformation of the domain-dependent Web data into a suitable form of transaction processes. This pre-processing transaction data is for the identification and integration of the components. In the second part, data mining and pattern matching

III. PROPOSED SYSTEM

The propose system presents a framework for web usage processing using semantic classification approach for the information incorporated with domain web site ontology to generate navigation patterns which should tailor the effective recommendations of web pages and services which the user really wishes to visit.

It performs semantic classification using the pattern generation, rather than in post processing. The qualities of the generated patterns are evaluated within a recommendation environment. By making use of the used by marketers to predict future visit patterns or target advertisements. In its Web context, mining of sequential patterns is used to capture frequent navigational paths in user trails. Markov models are the best machines used for link predictions using sequential patterns. The "trie" structure is also used to find navigation patterns. Web Utilization Miner uses this type of structure, using MINT as mining query language.

Clustering is a technique which groups together a set of items having similar characteristics. One of the most commonly used tasks of analysis in WUM is clustering of user records [7]. This tries to form group of users tending to have similar browsing patterns. Standard clustering algorithms such as k- means is used to partition the domain space into number of clusters depending on a distance or similarity among the users.

IV. EXISTING SYSTEM

The previous work on Web usage mining with semantic information is very limited. There are studies that aim to generate patterns in terms of semantic information as in 'towards semantic web mining', 'integrating semantic knowledge with web usage mining for personalization', 'Usage mining for and on the semantic web: next generation data mining'.

However, generally, these works map out the results of classical Web usage mining with ontological terms and concepts. Clustering appears as the major pattern generation technique, and the resulting patterns are used for generating Web page recommendations Integrating web usage and content mining for more effective personalization, using ontology and sequence information for extracting behavior patterns from web navigation logs. Unlike most of the previous approaches, we preferred to use sequential association rule mining, since the sequence information in the navigation is retained in the generated patterns.

Pattern Generation phase incorporates sequential association rule mining. In this case of association rule mining, the frequent patterns generated tend to maintain the sequence relation between the set of items discovered.

Semantic Classification

Semantic classification phase classifies the generated log pattern of a user in relevant to a query. It takes the domain ontology knowledge for accurate classification of the pattern generated. Semantic classification performs relevance association computation to relate the semantic relevancy. When the product is a service of information, the semantic-based classification of metadata and identity of each service will compute the similarity between the terms information. To perform the association of the document semantically enhanced patterns, recommendations are generated for a set of experiment analysis. The results are evaluated by comparing the generated recommendation with the visitor's actual next page in terms of precision and coverage. A shorter description of the proposed work with limited experimental results is presented in Using semantic information for web usage mining based recommendation.

A. System Modules

The figure 1 will provides the system framework for Web-page Annotation and Recommendation, which can be defined in following system modules as,

1. Pre-processing
2. Pattern Generation
3. Semantic Classification
4. Webpage categorization
5. Web Page Recommendation

Pre-processing

Pre-processing consists of removal of noisy and irrelevant data, and in addition to this it also integrates semantic information of Web pages with their log data. In this step log server files are pruned, transactions are extracted and ontology class individuals are mapped to the Web page address.

Pattern Generation we present a Concept Similarity mechanism. The concept of such words and phrases or metadata considers it finds its relevance. It is also associated with other elements of the domain it belongs to.

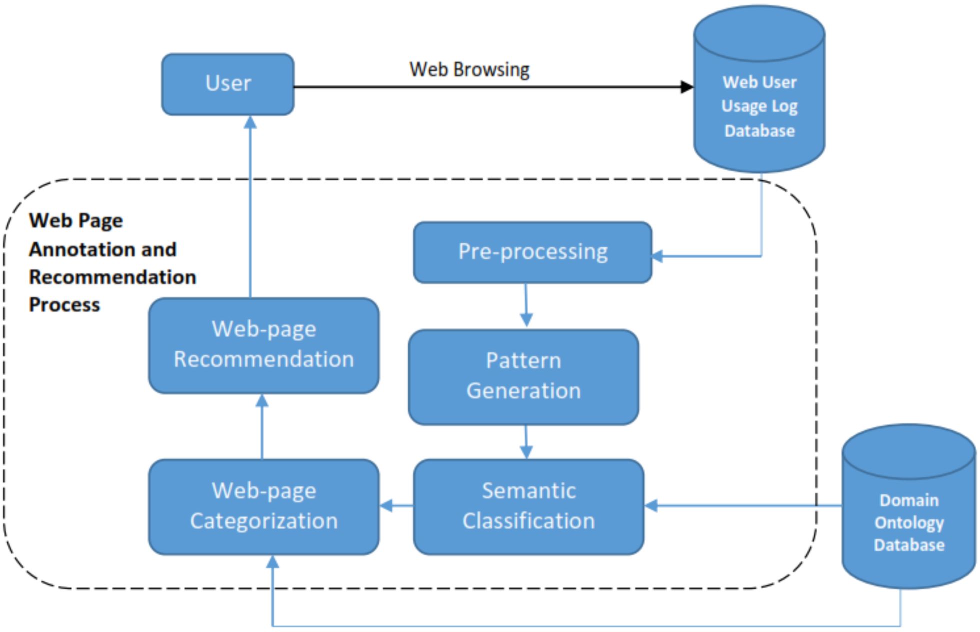
Let's consider a domain page, *Đ* which consists of set of terms belong a domain page as, *Đ* = *{t*$_1$*, t*$_2$*, ... ,t*$_n$*}*. To find the relation association we calculate the relation between the terms and the metadata of the domain page using a frequent of term calculation as, *f(t)* for the dependency of this term to other terms to this using the equation-1.

$$f(t) = \frac{T_n}{Z}, T \in S_k. \quad (1)$$

Where, *T* is the terms of domain page related to the terms of domain ontology S_k,and *Z* is the total number of extracted terms from web service document. On completion of *freq(t)* identification we compute the similarity relevancy probability as *P(t|s)*, of each term present in the domain of each document as *t*, against each domain service ontology as *s*,using the equation-2.

$$sim(d:s) = \frac{Prob\ (t \cap s)}{Prob\ (s)} \quad (2)$$

FRAMEWORK FOR WEB-PAGE ANNOTATION AND RECOMMENDATION



II. Fig.1Framework for Web-page Annotation and Recommendation

Web-page Categorization

Web-page categorization phase process the semantic classified pattern class label generated based on domain ontology and implements a *k-mean* method to relate the pattern categorization in respect to domain ontology knowledge.

Web-page Recommendation

Web-page recommendation process will implements a decision algorithm to identify the best page should recommend against the user query to meet the highest level of satisfactory.

To evaluate generated patterns, preference is given to build recommendation engines. In this step the approach to follow is that the page the user visits early in the visit, generally do not affect the next page, because users have the habit of clicking what is referred by recent pages. For this purpose the concept of window count can be used, which is the maximum number of previous pages to be taken into errors. Once the recommendation set is ready then the rules are mapped back to the Web page addresses and recommended to the users. The actual evaluation of this method is performed by using the statistical technique to define the effectiveness in terms of coverage and precision measure.

The obtained results data pattern will be evaluated based on account for recommendation, which user has visited already. Recommendation is generated by comparing active user's navigation history and sequential association rules. The following set of steps will be performed for the recommendation:

Algorithm:

Convert the navigation history to Patterns instances.

```
Initialize counter i=0.
do
{       Take the most recently navigated items from the Log;
        Scan using association rules;
        Build the pattern sets by adding rules which later part used for the recommendation process;
        i=i+1;
}
While (end of records);
```

[2]. Khalid Belhajjame, Suzanne M. Embury, and Norman W. Paton "Verification of Semantic Web Service Annotations Using Ontology-Based Partitioning", IEEE Transactions On Services the following measures.

Precision:Precision is defined as the proportion of the number of relevant recommendations to the *t* number of all recommendations. In other words, precision measures the accuracy of the recommendations.

$$precision = \frac{|R| \cap |eval|}{|R|}$$

Coverage:Coverage measures the ability of the recommendation system to produce all the page views that are likely to be visited by the user. In other words, it shows how well the recommendation covers all the pages that the user is likely to visit.

$$coverage = \frac{|R| \cap |eval|}{|eval|}$$

Where, *R* – is the recommendation phase produces for set of page views.

eval – is the set is compared with evaluated page views.

V. CONCLUSION

An optimized recommendation system has proposed in this paper. Algorithms for recommendation process have been proposed for implementing effective web search.

Calculating similarity measures is the major effective task in classification process, for finding the similarity measure, similarity equation has proposed in this paper and it will produce the effective result. The result will improve the relevancy of web pages and reduce the time to get accurate information. Time efficiency of recommender system will also improve.

REFERENCES

[1] ThiThanh Sang Nguyen, Hai Yan Lu, and Jie Lu, "Web-Page Recommendation Based on Web Usage and Domain Knowledge", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 10, October 2014

[2] Computing, Vol. 7, No. 3, July-september 2014

[3] B. Liu, B. Mobasher, and O. Nasraoui, "Web usage mining in Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", B. Liu, Ed. Berlin, Germany: Springer-Verlag, 2011, pp. 527–603.

[4] A Abello, O Romero, Torben B Pedersen, R Berlanga, Victoria N, M J Aramburu, and A Simitsis, "Using Semantic Web Technologies for Exploratory OLAP: A Survey", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 2, February 2015

[5] M Shirakawa, K Nakayama, Takahiro Hara And ShojiroNishio, "Wikipedia-Based Semantic Similarity Measurements for Noisy Short Texts Using Extended Naive Bayes", IEEE Emerging Topics in Computing, Volume 3, No. 2, June 2015

[6] Yiyao Lu, Hai He, Hongkun Zhao, WeiyiMeng, and Clement Yu, "Annotating Search Results from Web Databases", IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 3, March 2013

[7] Bin Jiang, Jian Pei, Yufei Tao, Member, and Xuemin Lin, "Clustering Uncertain Data Based on Probability Distribution Similarity", IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 4, April 2013

[8] Mclean, Y. Li, and Z.A. Bandar, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 4, pp. 871-882, July/Aug. 2003.

[9] Jason Deane, Praveen Pathak, "Ontological analysis of web surf history to maximize the click through probability of web advertisements" Springer - Elsevier 2009.

[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] R. Cooley, B. Mobasher, and J. Srivastava, “Web mining: Information and pattern discovery on the world wide web”, In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997, 1997.

[12] Jespersean S.E., Throhauge J., and Bach T., “A hybrid approach to Web Usage Mining, Data Warehousing and Knowledge Discovery”,SpringerVerlag Germany, pp73-82, 2002.

[13] Mobasher, B., Dai, H., Luo, T., & Nakagawa, M., “Effective Personalization Based on Association Rule Discovery from Web Usage Data”, Proceedings of the 3rd International Workshop on Web Information and Data Management, 2001.

[14] Chi E.H., Rosien A. and Heer J., LumberJack:IntelligentDiscovey and Analysis of Web User Traffic Composition. In Proceedings of ACMSIGKDD Workshop on Web Mining for Usage Patterns and User Profiles, Canada, ACM press, 2002.

[15] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. D.Spyropoulos, “Web Usage Mining as a Tool for Personalization: A Survey”, User Modeling and UserAdapted Interaction, 13(4):311-372, 2003.

[16] Robert Cooley, BamshadMobasher, and JaideepSrivastava, “Web Mining: Information and Pattern Discovery on the World Wide Web”, in Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, 1997.

[17] P. Senkul, S.Salin, "Improving pattern quality in web usage mining by using semantic information" Springer Verlag London Limited 2011.