Semantic Approach for Improving Pattern Quality in Web Usage Mining

Nandkishor P.Jilhedar
M.E. Student, CSE Department,
D. Y. Patil College of Engineering & Technology,
Kolhapur, Maharashtra, India.

Dr. S. K. Shirgave
Associate Professor & Head,
IT Department, DKTES. T.E.I.,
Ichalkaranji, Maharashtra, India.

Abstract - Users' experience in accessing websites can be personalized by processing web logs, using different Web Usage Mining (WUM) techniques. These web logs, if treated with such techniques and matched with semantic information of a website, can enrich the websites' personal access for an individual user. Web logs could be used to identify frequent navigation patterns as well as re-structuring the website for rich navigation experience of a user. The systematic approach presented in this paper focuses on generating frequent patterns and combining with semantic information. The evaluation of generated web usage pattern quality is done with standard metrics. Experiment results confirm that more accurate results can be obtained using this approach.

Keywords: Web Usage Mining, Frequent Patterns, Semantic

I.INTRODUCTION

The growth in the size of World Wide Web (WWW) has made it the commonplace for interest in the works related to web sites such as ecommerce and recommendation. While tremendous research is taking place to enhance the benefits of using the sites for web applications, the researchers are keep focusing on more systematic approach to integrate different parameters of web services with web applications such as semantic information, ontology etc. The site's success could be measured in terms of the ability to keep user navigating on site providing ongoing support with user's personal interest or choice, without losing the main aim of navigation[1]. But the issue in achieving this is the size, architecture and complexity of web is unknown many times, which makes it difficult to reach towards relevant information on the site. This problem could be rectified by using Web Usage Mining techniques which extract the information of interest from analysis of usage data of a specific user. Server logs are the sources of these usage data. The analyzed data obtained by WUM processing which includes capturing, processing and analyzing, is said to be more to a specific user oriented, and can be used for many applications such as web recommender systems, personalization, target advertising etc.

Semantic Web focuses on imparting the contents of websites understandable by humans as well as computers. To achieve this, the software agents are used to look out interested contents. Semantic information using ontology (like concept hierarchies or product catalog) has obtained more attention in representing Web pages and Web objects. The domain related knowledge can be used to develop ontological instances of a web site.

Web Mining techniques are used to find and predict the users' interests and recommending them with proper navigation on a web site[4]. This can improve the personalization experience of a user. It focuses on analyzing standard navigation of a Web user or the skeleton of a website and/or the contents of that site. Huge amount of research has been carried out to improve the process of web customization using the web access patterns of a user. But many times the context of the web site is not considered while customizing the web pages. Hence in this paper, the combined approach of semantic information and frequent pattern generation using conventional WUM[2][3], is presented to obtain more mature and semantically rich navigation patterns.

II. LITERATURE STUDY

There are a variety of data mining techniques which could be used to model activities of Web user[4]. But the actual process of WUM can be further subdivided into following interdependent stages: data collection and preprocessing, pattern discovery and pattern analysis. In the process of preprocessing, the clickstream data is cleaned which is obtained from the server logs and is further categorized into group of user transactions which could be used to identify activity of a specific user.

In pattern discovery, different statistical and database operations are performed. These operations try to bring out the behavior of users which reflects in discovered patterns. Techniques such as association and correlation analysis, navigation pattern analysis, cluster analysis are comprised into pattern discovery stage.

III. PROPOSED APPROACH

In this approach, first step is to obtain web requests made by different users to a website. The server logs are used for this purpose. A web server collects this information and stores it into a web log, which can be used for further processing. The following Figure 1 shows the block diagram of the proposed work.

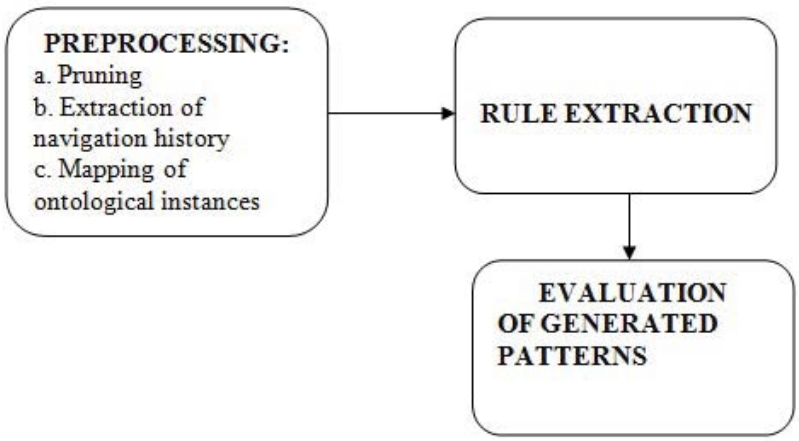

Figure 1. Block Diagram of Proposed System

A. Preprocessing

This step involves cleaning of server logs. It mostly includes steps such as removing irrelevant data/references, and removing noisy data. This step focuses on extraction of navigation history of a user and its mapping to corresponding ontology class.

A.1 Pruning

In this step non-responded requests and requests made by software agents such as crawlers are removed.

A.2 Navigation History Extraction

Navigation history is the collection of web objects requests made by users in the session. This step identifies individual users and their corresponding sessions and groups them accordingly with the link visits made by them. This information is useful in pattern generation and results evaluation.

A.3 Ontology Mapping

Ontology instances and web address requests captured in the log are mapped to fine tune the navigation recommendation. This ontology is built on the basis of the domain knowledge of the website such as site map.

B. Pattern Generation

This step incorporates sequential association rule mining, in which the frequent patterns are generated maintain the sequential association in the discovered items[6]. This pattern generation is carried out taking into consideration the minimum support threshold values. Frequent access links are shown to the user as an output by using these generated patterns, stored in a database.

Following Figure 2 shows the schematic of frequent pattern extraction process.

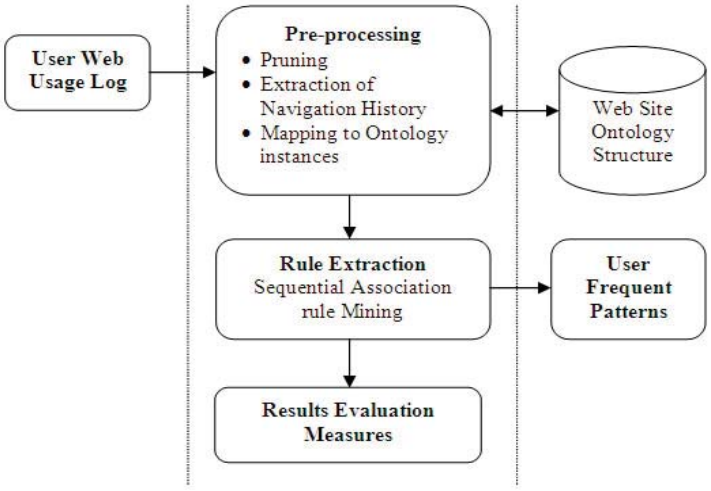

Figure 2. Frequent Pattern Extraction Process

III. EXPERIMENT EVALUATION

A. Web Logs

The large volume of data collected by webservers from a website is stored in the web logs. As well with the log files, other data such as user profiles , contents, information type and structure can be used in Web Usage Mining. For this setup of experiment dataset is collected fromthe web logs of the website reachouthyderabad.com. This dataset is further processed and its structure is modified as the IP addresses are replaced by user ids for evaluation purpose. Figure -3 shows a typical snap of a dataset- Weblog. Typical structure of Weblog is found to be similar for most of the webservers.

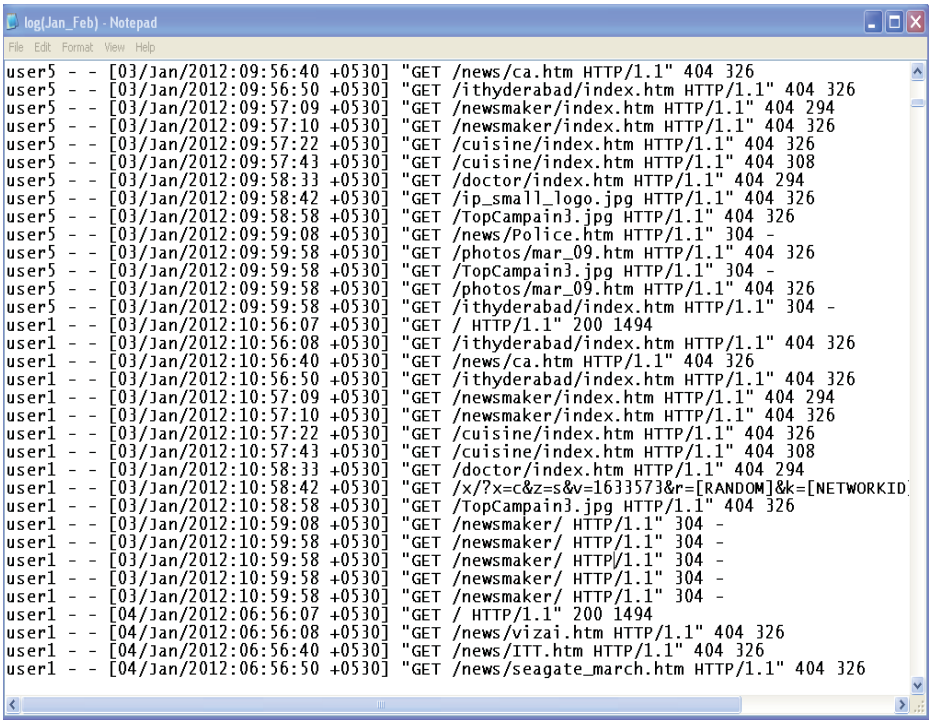

Figure 3. Dataset- Web Usage Log

B. Ontology instances

Ontology, once when described can be mapped with the links in the website and can be maintained in the database. This procedure is critical and requires maximum attention in the development. Because skipping of any link can result into incorrect output.

In the process of mapping ontology instances, the links are associated with numbers in order to use it in the recommendation phase easily[5][11]. The developed ontology instances are shown in the following figure.

head_id	head_name
1	City News
2	Leisure
3	IT Hyderabad
4	Students Den
5	Features
6	Biz Hyderabad
7	About Hyderab

Figure-4.a. Heads in the ontology

The following Figure-4 b shows the overall compilation of the links into the ontology, which can be referred to map the ontology instances to deeper details. More the depth of the ontology instances, more will be the accuracy of the mapping.

link_id	head_id	shead_id	link_value	links
1	1	1	Dr.Kapila Vatsy	/news/uoh9.htm
2	1	1	GHIAL obtains I	/news/ghail.htm
3	1	1	CM instructs of	/news/cmo38.htm
4	1	1	Reasoning Ann	/news/MartJack.htm
5	1	1	News Archives	/morenews.htm
6	1	2	Aircel launches	/newsmaker/hydwatch.l
7	1	2	More 240 articl	/newsmaker/hydwatch.l
8	1	3	Images: This w	/photos/mar_09.htm
9	1	3	Photo News Fe	/photos/Feb.htm
10	2	4	Gulal,13B,Delh	/moviereviews/filmrevi
11	2	4	Film personalit	/moviereviews/index.ht
12	2	4	Theatres/Imax,	/moviereviews/index.ht
13	2	5	Your Horoscop	/astro/index.htm
14	2	5	Know Your Birt	/astro/index.htm
15	2	6	Thought for Ha	/cheers/index.htm
17	2	7	Shopping areas	/shopping/index.htm
18	2	7	Places to relax	/recreation/index.htm
19	2	7	Plan your Trave	/travel/index.htm
20	2	7	Weekend trip I	/travel/beyondhyd.htm
21	3	8	IT Companies |	/ithyderabad/index.htm
24	4	10	List of City Coll	/students_den.htm
25	4	11	Free All India S	/news/gate.htm
27	5	12	P Abraham:'Me	/newsmaker/index.htm
28	5	12	View 175 More	/newsmaker/index.htm

Figure 4. Lookup for Links in the ontology

C. *Generated Frequent Patterns*

Pruned and preprocessed log data is used as input to generate frequent patterns, as explained. The sequential patterns generated with respect to different users for a given threshold value. Following are the obtained results for generated frequent patterns.

```
============================
USER2
============================
Threshold        Freq. Pattern
------------------------------------------------------------------------
2       [5, 1, 3, 8, 21, 27, 28, 29, 51, 31, 13, 14, 15, 20, 19, 60, 17, 63, 39]
4       [5, 1, 3, 8, 21, 27, 28, 29, 51, 31, 13, 14, 15, 20, 19, 60, 17, 63]
6       [5, 1, 3, 8, 21, 27, 28, 29, 51, 31, 13, 14, 15, 20, 19, 60, 17, 63]
8       [5, 1, 3, 8, 21, 27, 28, 29, 51, 31, 13, 14, 15, 20, 19, 60, 17, 63]
10      [5, 1, 3, 8, 21, 27, 28, 29, 51, 31, 13, 14, 15, 19, 60, 17, 63]
============================
USER3
============================
Threshold        Freq. Pattern

2       [13, 14, 15, 20, 17, 63, 5, 37, 38, 21, 27, 28, 29, 51, 31, 8, 19, 60, 1, 3, 45, 46, 47,
4       [13, 14, 15, 20, 17, 63, 5, 37, 38, 21, 27, 28, 29, 51, 31, 8, 19, 60, 1, 3, 46, 47, 59,
6       [13, 14, 15, 20, 17, 63, 5, 37, 38, 21, 27, 28, 29, 51, 31, 8, 19, 60, 1, 3, 47, 59, 52]
8       [13, 14, 15, 20, 17, 63, 5, 37, 38, 21, 27, 28, 29, 51, 31, 8, 19, 60, 47, 59]
10      [13, 14, 15, 20, 17, 63, 5, 37, 38, 21, 27, 28, 29, 51, 31, 8, 19, 60]
============================
USER4
============================
Threshold        Freq. Pattern
------------------------------------------------------------------------
2       [50, 52, 53, 18, 44, 47, 59, 42, 54, 49, 17, 63, 24, 5, 31, 37, 38, 25, 15]
4       [50, 52, 53, 18, 44, 47, 59, 42, 54, 49, 17, 63, 24, 5, 31, 37, 38, 25]
6       [50, 52, 53, 18, 44, 47, 49, 17, 63, 24]
8       [53, 18, 44, 47, 17, 63]
10      [18, 44, 47, 17, 63]
============================
```

Figure 5. Generated Frequent Patterns for different Support Threshold values

III. RESULTS EVALUATION

A. *Evaluation Measures*

Following evaluation metrics are used to evaluate the obtained results of pattern generation . Rules are mapped back to the Webpage locations once the recommendation link set is obtained[10][14]. Then it is recommended to users. Statistical techniques- Tenfold cross validation is used to define the effectiveness in terms of coverage, precision, F1 measure and R measure.

- *Precision & Coverage:* The ratio of relevant recommendations to the number of all recommendations is known as precision.

$$precision(R,t) = \frac{|R| \cap |t - w|}{|R|}$$

The effectiveness of recommendation system is measured by calculating coverage, to incur more accuracy in the recommendation.

- *F-1 Measure:* High precision and coverage is expectable and F1- measure is used to accomplish this. F-1 measure is calculated by,

$$F1(R,t) = \frac{2 \times precision(R,t) \times coverage(R,t)}{precision(R,t) + coverage(R,t)}$$

The F1 measure gains its maximum value when both precision and coverage are maximized.

- *R-Measure:* This measure is calculated by dividing the coverage by the size of recommendation set.

$$R(R,t) = \frac{coverage(R,t)}{|R|}$$

B. *Results Analysis*

The evaluation of the proposed system was carried out for different users with different support thresholds. The Table-1 shows the result for the explained metrics. This obtained result shows the improvement in precision and coverage, F-1 measure and R-1 Measure.

Log Users	Sup. Th.	Precision	Coverage	F1-Measure	R1-Measure
User1	2	0.5	0.5	0.5	0.011363637
	4	0.5176471	0.48235294	0.4993772	0.010962567
	6	0.5641026	0.43589744	0.49178174	0.00990676
	8	0.5714286	0.42857143	0.48979595	0.0097402595
	10	0.60273975	0.39726028	0.47888914	0.009028642
User2	2	0.5	0.5	0.5	0.02631579
	4	0.5135135	0.4864865	0.49963477	0.025604552
	6	0.5135135	0.4864865	0.49963477	0.025604552
	8	0.5135135	0.4864865	0.49963477	0.025604552
	10	0.5277778	0.4722222	0.49845678	0.024853801
User3	2	0.5	0.5	0.5	0.011904762
	4	0.54545456	0.45454547	0.4958678	0.010822511
	6	0.64615387	0.35384616	0.45727813	0.008424909
	8	0.67741936	0.32258064	0.43704474	0.0076804915
	10	0.7	0.3	0.42000002	0.0071428576
User4	2	0.5	0.5	0.5	0.02631579
	4	0.5135135	0.4864865	0.49963477	0.025604552
	6	0.6551724	0.3448276	0.45184305	0.01814882
	8	0.76	0.24	0.36479998	0.012631578
	10	0.7916667	0.20833333	0.3298611	0.010964912

Table 1. Result Evaluation Metrics for different users

IV. CONCLUSION

Web servers capture the website browsing actions and provide web logs. These logs could be used to generate frequent navigation pattern of a website. These frequent patterns provide valuable information which could be used for many applications such as website personalization, link recommendations etch. In conventional WUM process, the semantic information is not taken into consideration for carrying out data mining tasks. In the presented work, the impact of combining frequent patterns and semantic representation of the website is investigated, which shows the improvement in the quality of generated pattern, considerably as shown in the result analysis. From the work, it can be included that, deeper the semantic information more is the quality obtained in the pattern generation process.

REFERENCES

[1] Gerd Stumme, Andreas Hotho, Bettina Berendt "Semantic Web Mining -State of the Art and Future Directions" Journal of web semantics 2006.

[2] Ahu Sieg, Bamshad Mobasher, Robin Burke "Learning Ontology-Based User – Profiles: A Semantic Approach to Personalized Web Search" IEEE Int. Informatics Bulletin, Nov 2007

[3] Pinar Senkul, Suleyman Salin "Improving pattern quality in web usage mining by using semantic information" Springer –Verlag London Limited 2011

[4] Bamshad Mobasher, "Data Mining for Web Personalization" Springer-Verlag Berlin Heidelberg 2007

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

[6] Costantinos Dimopoulos, Christos Makris "A Web page prediction scheme using sequence indexing and clustering techniques" Springer-Elsevier 2009

[7] JIE TANG, LIMIN YAO, DUO ZHANG, JING ZHANG "A Combination Approach to Web User Profiling" ACM Transactions on Knowledge Discovery from Data, Vol. 5, No. 1, Article 2, Pub. date: December 2010.

[8] G.Sudhamathy, "Mining Web Logs – An Automated Approach" 2010 978-1-4503-0194-7/10/0009

[9] Tahira Hasan, Sudhir Mudur, Nematollaah Shiri "A Session Generalization Technique for Improved Web Usage Mining" 2009 ACM 978-1-60558-808-7/09/11

[10] Mehdi Adda, Petco Valtchev, Rokia Missaoui "A framework for mining meaningful usage patterns within a semantically enhanced Web portal", 2010 ACM 978-1-60558-901-5/10/05

[11] R G Tiwari, Mohd. Husain, V Srivastava, A Agrawal, Ambalika IMT, "Web Personalization by Assimilating Usage Data and Semantics Expressed In Ontology Terms" International Conference and Workshop on Emerging Trends in Technology (ICWET 2011) – TCET, Mumbai, India

[12] D Vora, S Bojevar "Design of a Tool Using Statistical Approach for Personalization and Usability Improvement" International Conference and Workshop on Emerging Trends in Technology (ICWET 2011) – TCET, Mumbai, India

[13] NIZAR R. MABROUKEH, C. I. EZEIFE, "A Taxonomy of Sequential Pattern Mining Algorithms" ACM Computing Surveys, Vol. 43, No. 1, Article 3, Publication date: November 2010.

[14] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns"