Discovering Semantic Association Rules using Apriori & kth Markov Model on Social Mining

Ritesh Dubey[1] Dr. Kavita[2]
[1]Ph.D Scholar [2]Professor
[1,2]Department of Computer Science & Engineering
[1,2]JVWU, Jaipur, India

Abstract— The Semantic Web opens up new opportunities for the data mining research. 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.

Key words: Web Usage Mining, Semantic Web, Domain, Sequential Pattern Mining, Sentiment Analysis, Opinion Mining, Support Vector Machine, Term Frequency, TF-IDF

I. INTRODUCTION

Modeling the user web navigation behavior is becoming the challenging task as the growth of the World Wide Web is increasing rapidly. 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. Web scraping is the process of automatically collecting useful information from we]. 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. 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. Below figure1 shows that general architecture of recommendation system.

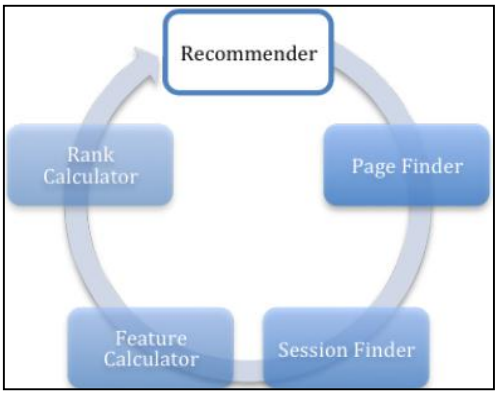


Fig. 1: General Architecture of Recommendation System

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. 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 probable next web page in advance.

Web Personalization, Adaptive web sites are some of the applications of web usage mining. Web usage mining can provide guidelines for improving ecommerce to handle business specific issues like customer attraction, customer retention, crosses sales, and customer departure.

A. Web Log

The web log is a registry of web pages accessed by different users at different times, which can be maintained at the server-side, client-side or at a proxy server.

– Server Log:
– Client Log:
– Proxy Log:

1) Web Usage Mining

In recent times, Web Usage Mining has emerged as a popular approach in providing Web personalization. Web usage mining is concerned with finding user navigational patterns on the World Wide Web by extracting knowledge from web usage logs (we will refer to them as web logs). The

assumption is that a web user can physically access only one web page at any given point in time that represents one item. The process of Web Usage Mining goes through the following three phases are.

- Preprocessing phase: The main task here is to clean up the web log by removing noisy and irrelevant data. In this phase also, users are identified and their accessed web pages are organized sequentially into sessions according to their access time, and stored in a sequence database.
- Pattern Discovery phase: The core of the mining process is in this phase. Usually, Sequential Pattern Mining (SPM) is used against the cleaned web log to mine all the frequent sequential patterns.
- Recommendation/Prediction phase: Mined Patterns.

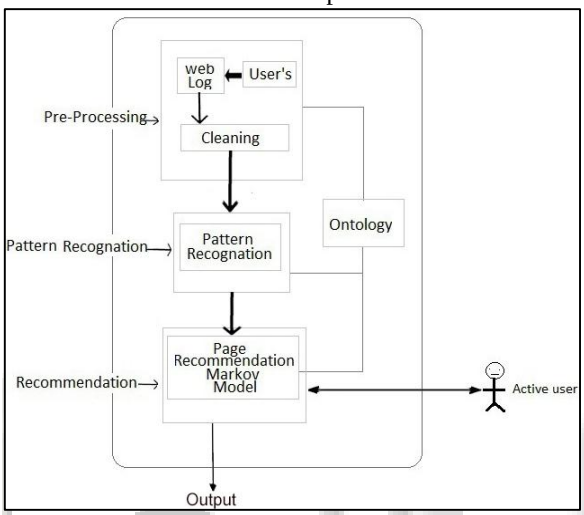


Fig. 1: Phases of Web Usage Mining

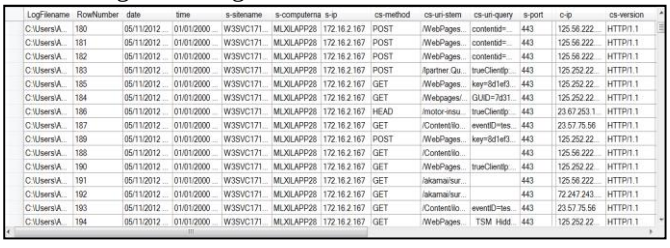
Web Usage Mining is the field of web mining which deals with finding the interesting usage pattern from the logging information. The logging information is stored in a file known as web log file. Web log file contains lot of information like IP address, date, time, web page requested etc.

B. Web Log

The web log is a registry of web pages accessed by different users at different times, which can be maintained at the server-side, client-side or at a proxy server, each having its own benefits and drawbacks on finding the users' relevant patterns and navigational sessions.

1) Server Log

The server stores data regarding requests performed by the client, thus data regard generally just one source. Server Log details are given in Figure 2.



LogFilename	RowNumber	date	time	s-sitename	s-computerna	s-ip	cs-method	cs-uri-stem	cs-uri-query	s-port	c-ip	cs-version
C:\Users\A...	180	05/11/2012...	01/01/2000...	W3SVC171...	MLXLAPP28	172.16.2.167	POST	/WebPages...	contentId=...	443	125.56.222...	HTTP/1.1
C:\Users\A...	181	05/11/2012...	01/01/2000...	W3SVC171...	MLXLAPP28	172.16.2.167	POST	/WebPages...	contentId=...	443	125.56.222...	HTTP/1.1
C:\Users\A...	182	05/11/2012...	01/01/2000...	W3SVC171...	MLXLAPP28	172.16.2.167	POST	/WebPages...	contentId=...	443	125.56.222...	HTTP/1.1
C:\Users\A...	183	05/11/2012...	01/01/2000...	W3SVC171...	MLXLAPP28	172.16.2.167	POST	/partner Qu...	trueClientIp...	443	125.252.22...	HTTP/1.1
C:\Users\A...	185	05/11/2012...	01/01/2000...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/WebPages...	key=8d1ef3...	443	125.252.22...	HTTP/1.1
C:\Users\A...	184	05/11/2012...	01/01/2000...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/Webpages/...	GUID=7d31...	443	125.252.22...	HTTP/1.1
C:\Users\A...	186	05/11/2012...	01/01/2000...	W3SVC171...	MLXLAPP28	172.16.2.167	HEAD	/motor-insu...	trueClientIp...	443	23.67.253.1...	HTTP/1.1
C:\Users\A...	187	05/11/2012...	01/01/2000...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/Content/lo...	eventID=tes...	443	23.57.75.56	HTTP/1.1
C:\Users\A...	189	05/11/2012...	01/01/2000...	W3SVC171...	MLXLAPP28	172.16.2.167	POST	/WebPages...	key=8d1ef3...	443	125.252.22...	HTTP/1.1
C:\Users\A...	188	05/11/2012...	01/01/2000...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/Content/lo...		443	125.56.222...	HTTP/1.1
C:\Users\A...	190	05/11/2012...	01/01/2000...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/WebPages...	trueClientIp...	443	125.252.22...	HTTP/1.1
C:\Users\A...	191	05/11/2012...	01/01/2000...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/akamai/sur...		443	125.56.222...	HTTP/1.1
C:\Users\A...	192	05/11/2012...	01/01/2000...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/akamai/sur...		443	72.247.243...	HTTP/1.1
C:\Users\A...	193	05/11/2012...	01/01/2000...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/Content/lo...	eventID=tes...	443	23.57.75.56	HTTP/1.1
C:\Users\A...	194	05/11/2012...	01/01/2000...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/WebPages...	TSM Hidd...	443	125.252.22...	HTTP/1.1

Fig. 2: A Sample of Serer Side Web Log

2) Client Log

It is the client itself which sends to a repository information regarding the user's behavior (can be implemented by using a remote agent (such as Java scripts or Java applets) or by modifying the source code of an existing browser (such as Mosaic or Mozilla) to enhance its data collection capabilities.);

3) Proxy Log

Information is stored at the proxy side, thus Web data regards several Websites, but only users whose Web clients pass through the proxy.

II. LITERATURE REVIEW

Due to rapid growth in the number of internet users, the user perceived latency has become a serious issue for the web service providers. Researches have been done which combines different techniques from multiple domains to overcome this issue.

To reduce perceivable network latency, researchers focused on pre-fetching popular documents. The integration of pre-fetching and caching techniques greatly improves the performance and also reduces the running time of the applications by 50%.

Harish Kumar et.al [1] The proposed research work uses hierarchical clustering technique with modified Levenshtein distance, Page Rank using access time length, frequency and higher order Markov model for prediction. Experimental results prove that the proposed approach for prediction gives better accuracy over the existing techniques. The proposed work can be used to prefetch the web pages before they are actually being requested by the user, this reduces the access latency.

Omair Shafiq et.al. [2] They have built Semantic FP-Trees based technique to perform association rule learning on functional and non-functional characteristics of Web Services. The process of automated execution of Web Services is improved in two steps, i.e., (1) we provide semantically formalized logs that maintain well- structured and formalized information about past interactions of Services Consumers and Web Services, (2) we perform an extended association rule mining on semantically formalized logs to find out any possible correlations that can used to pre-filter Web Services and reduce search space during the process of automated ranking and adaptation of Web Services. We have conducted comprehensive evaluation to demonstrate the efficiency, effectiveness and usability of our proposed approach.

Mayank Kalbhor [3] This paper has provided a more current evaluation and updating of web usage mining "future prediction of web user access" research available. Developed and tested our hybrid approach on different available datasets, and results are compared with markov model.

Lakshmana Phaneendra Maguluri et. al.[4] This paper proposes a method to reuse the previously computed values using a hash data structure thus reduce the execution time. To demonstrate the effectiveness of the proposed method, experiments were conducted on SWETO ontology. Results show that the proposed method is more efficient than the other existing methods. This paper proposed a new technique for defming the context more efficiently at different levels viz. at schema level, at instance level and at relationship level. Using this user can access more relevant associations. Additionally this paper proposed another

technique to reuse the previously computed values which reduced the total execution time.

Priyanka Bhart [10] In this paper users browsing behavior is firstly preprocessed using hierarchical clustering then prediction is done in two phases. In first phase category prediction is done using Markov model then in second phase page prediction is done.

III. IMPORTANCE OF STUDY

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 Pre-processing, 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 provide 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
- Recommendation

IV. RESEARCH OBJECTIVES & PROBLEMS

A. *Objectives are*

- To examine semantic association rules using Apriori algorithm and k^{th} Markov model.
- To design Apriori algorithm for association rules.
- To implement k^{th} Markov model using JDK and Netbean tools.
- To analyze pattern discovery on social mining.
- To predict memory problem in FP tree by using Apriori.

B. *Problems in Existing Systems are*

- Do Not provide a way for predicting future user access pattern
- Current sequential pattern mining techniques suffer from a number of drawbacks, some of which include:
- The sequence data base is scanned on nearly every pass of the algorithm
- A large data structure has to be maintained in memory all the time.
- Support counting has to be maintained at all times during mining, which adds to the memory size required.

1) *Disadvantages of FP-Growth*

- FP-Tree may not fit in memory
- FP-Tree is expensive to build

V. METHODOLOGY

In the proposed work, the research emphasis on the following points:

- Literature Survey: In depth analysis of existing work and their results.
- For Experimental and modelling some tool/library will be use like NetBeans, JfreeChart Library, Weblog file.

The Data will be analyzed and the results will be presented.

VI. 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 use for subsequently page request prediction, as a solution to ambiguous predictions problem and providing an informe lower order Markov model without the need for complex higher order Markov models.

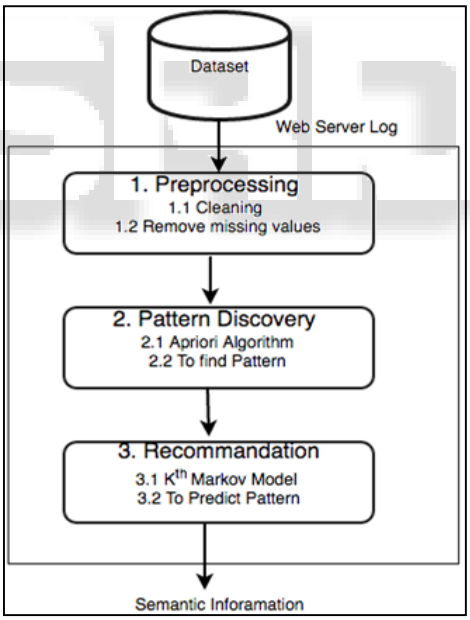


Fig. 3: Processing Step of the Proposed System

The processing steps of the system have three main phases. Preprocessing is performed in the first phase. The second phase is clustering web sessions using K-means clustering. In the final phase, Markov model is used to predict next page access based on resulting web sessions. The popularity and similarity-based page rank algorithm is used to decide the most relevant answer if the ambiguous result is found in Markov model prediction. The input of the proposed system is a web log file. A web log is a file to which the web server writes information each time a user requests a resource from that particular site

1. *Begin*
   a. *Data Preprocessing is carried out on the input web log file*
   b. *Find Pattern using FP Tree Algorithm*
   c. *Build a k-Markov model*
   d. *For Markov model*
      i. *states where the result is not clear*
   e. *End For*
2. *End*

3. *Prediction:*
4. *Begin*
5. *For each coming session*
   a. *Use Markov model to make prediction*
      i. *If the predictions are made with the ambiguous result*
      ii. *Use page rank algorithm to make a prediction*
      iii. *End If*
6. *End For*
7. *End*

The proposed system focuses on the improvements of predicting web page access. The process is as:

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

VII. Conclusion

Web usage mining model is kind of mining to server logs. Web usage mining used for the improvement of improving the requirement of the system performance, the customers relation and realizing enhancing the usability of the website design. The main goal of the proposed system is to identify semantic association rules using Apriori algorithm and kth Markov model & analyze pattern discovery on social mining.

References

[1] Harish Kumar B T, Dr. Vibha L and Dr. Venugopal K R,"Web Page Access Prediction Using Hierarchical Clustering Based on Modified Levenshtein Distance and Higher Order Markov Model " Region 10 Symposium (TENSYMP), Bali, Indonesia IEEE 2016.

[2] Omair Shafiq, Reda Alhajj, Jon G. Rokne "Reducing Search Space for Web Service Ranking using Semantic Logs and Semantic FP-Tree based Association Rule Mining " 9th International Conference on Semantic Computing (ICSC) IEEE 2015.

[3] Mayank Kalbhor, Kunl Jain "Fuzzy Based Hybrid Approach for User Request Prediction Using Markov Model" International Conference on Computer, Communication and Control (IC4) IEEE 2015.

[4] Lakshmana Phaneendra Maguluri, M Vamsi Krishna, P S S Sridhar "ANovel Approach for Discovering Relevant Semantic Associations on Social Web Mining "Conference on IT in Business, Industry and Government (CSIBIG), IEE 2014.

[5] Priyanka S. Panchal , Prof. Urmi D. Agravat "Hybrid Technique for User's Web Page Access Prediction based on Markov Model" 4th ICCCNT, Tiruchengode, India July 4-6, 2013

[6] Poornalatha G, Prakash S Raghavendra "Web Page Prediction by Clustering and Integrated Distance Measure" ACM International Conference on Advances in Social Networks Analysis and Mining IEEE 2012.

[7] Mamoun A. Awad and Issa Khalil "Prediction of User's Web-Browsing behavior: Application of Markov Model" IEEE Transactions on Systems, Man, and Bernetics—Part B: Cybernetics, Vol. 42, No. 4, August 2012.

[8] M. M. Group. Internet world stats - usage and population statistic http://www.internetworldstats.com/stats.htm/, 2011. Last Visit: October 2011.

[9] M. D. Kunder. World wide web size - daily estimated size of the world wide web. http://www.worldwidewebsize.com/, 2011. Last Visit: November 2011.

[10] Priyanka Bhart "Prediction Model Using Web Usage Mining Techniques "[IJCA-2014]

[11] N. Duhan, A. Sharma, and K. Bhatia. "Page ranking algorithms: A survey", International Advance Computing Conference (IACC) IEEE March 2009.

[12] Y. Z. Guo, K. Ramamohanarao, and L. Park. "Personalized pagerank for web page prediction based on access time-length and frequency", International Conference on Web Intelligence, IEEE/WIC/ACM, pages 687-690, Nov. 2007.

[13] Ruma Dutta, Anirban Kundu, and Debajyoti Mukhopadhyay "Offering Memory Efficiency utilizing Cellular Automata for Markov Tree based Web-page Prediction Model" 10th International Conference on Information Technology IEEE 2007.

[14] H. Liu and V. Ke_selj, "Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests", Data Knowl. Eng., vol 61, pp 304-330, May 2007.

[15] M. Eirinaki and M. Vazirgiannis, "Usage-based page rank for web personalization" 5th International Conference on Data Mining IEEE Nov. 2005.

[16] M. Eirinaki, M. Vazirgiannis, and D. Kapogiannis, "Web path recommendations based on page ranking and markov models", 7th annual ACM international workshop on Web information and data management, WIDM '05, pages 2-9, New York, USA, 2005.

[17] Diamanto Oikonomopoulou , Maria Rigou, Spiros Sirmakessis, thanasios Tsakalidis "Full-Coverage Web Prediction based on Web Usage Mining and Site Topology", International Conference on Web Intelligence IEEE/WIC/ACM 2004.

[18] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization", ACM Trans. Internet Technol, Vol.3, pp 1-27, February 2003.

[19] S. Gunduz and M. T. Ozsu. "A web page prediction model based on click-stream tree representation of user behaviour", 9th international conference on Knowledge

discovery and data mining ACM SIGKDD, KDD ' pages 535-540, New York, USA, 2003.

[20] T. Haveliwala. "Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search", IEEE Transactions on Knowledge and Data Engineering, Vol 15, pp 784-796, July-Aug. 2003.

[21] Devanshu Dhyani ,Sourav S Bhowmick, Wee-Keong ng "Modelling and Predicting Web Page Accesses Using Markov Processes" 14th International Workshop on Database and Expert Systems Applications IEEE 2003.

