Implementation to Find Navigational pattern of Log Files Using Hadoop Technology

Himanshu Kaushal [#1], Dr. Ravi Kumar Singh Pippal[*2]
[#]*M.Tech Scholar & Professor*
[1,2]*Department of Computer Science and Engineering*
[1,2]*Vedica Institute of Technology, Bhopal, India*
[1]himanshukaushal16@gmail.com
[2] ravesingh@gmail.com

Abstract

It is necessary to preprocess the site log before modeling when dealing with a substantial amount of data. The web log file that has been preprocessed is used to chronologically order the user's web navigation sessions. A web navigation session is the order in which a person visits several websites on the internet within a given period. The user navigation session is ultimately delineated by a paradigm. As soon as the user navigation model has been developed, it is possible to carry out mining to find the pertinent pattern. Specifically with regard to the field of web usage mining, modeling weblogs is a crucial task to do. It is possible to improve the accuracy of forecasts by modeling the weblog with an accurate model, accomplished through caching, which stores frequently requested websites in the caches of proxy servers. When paired with caching, the pre-fetching of web pages is a new research topic that has the potential to significantly increase performance. An enhanced method for predicting web pages is proposed in this research study. Clustering is used to categorize users of the web according to their locations. Finally, the FP-Growth algorithm mines each cluster to determine which pages should be pre-fetched and cached.

Keywords: Mining, semantic web, domain, web log, prediction, Markov model, sequential pattern mining, recommender systems

1. INTRODUCTION

1.1 Web Usage Mining:

Web Usage Mining has become a prominent method for delivering Web customisation in recent years. Finding patterns in how users navigate the internet by extracting information from web usage logs—which we will call web logs—is the focus of web usage mining. One physical web page, representing one thing, is assumed to be accessible to a web user at any given moment.

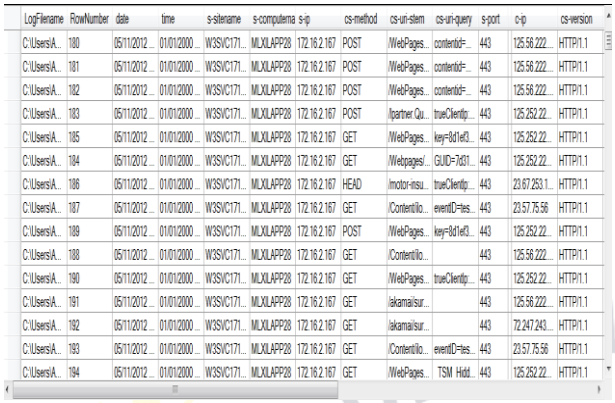
The following three stages comprise the Technique for Mining Web Usage.

- Preprocessing phase: The primary goal at hand is to purge the site log of any superfluous or distracting information. Users' identities are also verified in this phase, and the web sites they have viewed are sorted into sessions based on when they were accessed and saved in a database.
- Pattern Discovery phase : The Pattern Discovery phase is where the mining process is most centred. Most of the time, To extract all the frequently occurring sequential patterns from the cleansed web log, Sequential Pattern Mining (SPM) is applied. Recommendation/Prediction phase: Mined patterns

The area of web mining known as "web usage mining" searches log data for intriguing usage patterns. The logging information is kept in a file known as a "web log file." The web log file contains a wealth of information, such as the IP address, time, date, and desired web page.

1.2 Web Log: A web log records the URLs visited by various users over time. When determining functional user patterns and navigation sessions, there are pros and cons to keeping information on the server, at the client, or in a proxy server.

- Server Log: The server records client request data, which usually only records data from one source. In Figure 1, you can see specifics taken from the server log.
- Client Log: The Client Log feature uses a remote agent like JavaScript or Java applets to allow clients to send messages to a repository containing user behavior information. The Proxy Log feature maintains information on the proxy side, enabling tracking and data harvesting for users whose web clients go through the proxy. This is achieved by modifying the source code of existing browsers like Mozilla or Mosaic.
- Proxy Log: The proxy server stores data, making websites viewable to only those whose web clients pass via it.



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	/bpartner 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=0d1ef3...	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	/ContentInfo...	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=0d1ef3...	443	125.252.22...	HTTP/1.1
C:\Users\A...	188	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/ContentInfo...		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	/akamaisur...		443	125.56.222...	HTTP/1.1
C:\Users\A...	192	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/akamaisur...		443	72.247.243...	HTTP/1.1
C:\Users\A...	193	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXLAPP28	172.16.2.167	GET	/ContentInfo...	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 Hdd...	443	125.252.22...	HTTP/1.1

Fig:1 A Sample of Serer Side Web Log

1.3 Domain Knowledge:

The pattern relationship or link between all of these data can provide information. Knowledge about the past, present, and future can be derived from data. Knowledge regarding data that has already been discovered or obtained from domain experts is referred to as domain knowledge. Three classes of domain knowledge are identified: Environment Based Constraints (EBC), Attribute Relationship Rule (AR-Rule), and Hierarchical Generalisation Tree (HG Tree). The pattern relationship or link between all of these data can provide information. Data can be transformed into knowledge about the past, present, and future. Knowledge regarding data that has already been discovered or obtained from domain experts is referred to as domain knowledge. Three classes of domain knowledge are identified: Environment Based Constraints (EBC), Attribute Relationship Rule (AR-Rule), and Hierarchical Generalisation Tree (HG Tree).

1.4 Ontology:

"The specification of Conceptualization" is a common definition of ontology. Ontology establishes the relationship between the variables and compartmentalises the variables required for a certain set of calculations. Types of ontologies include higher, hybrid, and domain. We employed the electronic item domain ontology in the experiment. A domain ontology displays concepts that are specific to a region of the globe.

A model known as domain ontology describes a particular domain by means of a set of concepts (C) and their relations (R). Among the domain ontology concepts taken into account by our Model, set C includes Product, Purchase, Supplier, and Warehouse, to name a few.

2. RELATED WORK

Using OWL technology, Sneha Y.S. et al. have added semantics to the existing navigational paths in this work. A framework for merging semantic information with navigational patterns is presented in this paper. This study assessed the framework and presented encouraging findings about suggestions for product quality.

Amit Bose presented a personalization framework that integrates domain knowledge and usage data, drawing inspiration from concepts in information retrieval and bioinformatics. These studies, in contrast to our paradigm, do not incorporate the Web application's domain knowledge throughout the whole Web usage mining process.

J. Vellingiri among others Analyzing how users interact with websites can reveal information that helps users customize and personalize their online experiences. For this reason, online usage mining is receiving a lot of attention from e-marketing and e-commerce experts. Three stages comprise Pre-processing, pattern recognition, and pattern analysis in web usage mining. There are various online usage mining approaches, each having pros and cons of its own. This paper discusses a few web usage mining approaches that are currently accessible.

The research described in collected domain-level objects from user sessions and aggregated these data using a merge function and weights to produce a user profile for each user. It is presumed that the website already has a domain-level ontology and that merge functions have been developed for each attribute of the objects.

An elevated chain of command Because of the work of Sarukkai, Pitkow, Pirolli, and Seidenberg's method for extracting related concepts from GALEN based on one or more classes the user chose, Markov models were used for link prediction. The order of a Markov model depends on how many past events were used to predict a future event. Thus, a kth-order Markov model predicts the chance of the next event based on the past k events.

3. PROPOSED SYSTEM

In order to target web recommendation and web usage mining, we suggest using domain ontology and integrating it into a comprehensive system. The suggested architecture is

Algorithm: Combine_Apriori

(Set_Sk-1, Matrix[][], Delta)

Set_M1 = Empty_Set

for each pair (X, Y) in Set_Sk-1

where X = { item1, item2, …, item(k-2), item(k-1) }

and Y = { item1, item2, …, item(k-2), item'(k-1) }

and Relation(item(k-1), item'(k-1)) in Delta /* Relation(item(k-1), item'(k-1)): Defines the semantic distance between item(k-1) and item'(k-1) */

/* Delta: Maximum Semantic Distance - The maximum permissible semantic distance between any two semantic entities. Delta is a user-defined value, determined as specified in reference [18]. The distance between item(k-1) and item'(k-1) is derived from Matrix */

The proposed architecture is divided into three phases:

There are three stages to the suggested architecture. A clean server-side weblog is preprocessed in the first stage to extract pertinent data such as user requests, page visits, and concept maps. After that, this data is transformed into a semantic sequence database called D that has been enhanced with semantic data.

Pattern Discovery is the second stage in the suggested architecture that receives input from the sequence database D. We presented two algorithms, S_P_M and Join Apriori, to process the sequence database by pruning candidate sequences and those with less than the required supporting count.

Experimental findings have demonstrated that the laptop-based domain ontology used in the web can improve the efficacy and performance of any S_P_M algorithm, under certain conditions. This improvement is achieved without compromising the quality of the often-occurring patterns.

The third stage of the suggested architecture makes use of them to produce top recommendations and semantic association rules. This stage contains methods for additional reporting and result filtering. The outcomes of this phase are presented to the administrator as a set of suggestions for decision-making, which can be applied to marketing campaigns, website redesigns, active user guidance, or website management.

4. SEQUENTIAL PATTERN MINING

The proposed algorithm An Apriori algorithm version called S_P_M, which uses semantic information to generate common sequences and Join_Apriori() to generate candidates, is described. S P M creates. An algorithm proposed The inputs utilized by S_P_M include a sequence database (SQ_Database), a sequence matrix Mat[ ][ ], minimal support, and common sequences for rich semantic output. Here is the S_P_M algorithm in action, as seen in Figure 3.

Algorithm:Join_ Apriori (S_{k-1}, Mat[ ][ ], _)

Initialize an empty set M1 as Ø.

Consider two sets P and Q from Sk-1, where:

P = { i1, i2, …, ik-2, ik-1 }

Q = { i1, i2, …, ik-2, i'k-1 }

D(ik-1, i'k-1) is a function representing the semantic distance between ik-1 and i'k-1. This value is derived from a matrix M and must be less than or equal to a maximum semantic distance, denoted as "max", which is a user-defined parameter. The method to determine this value is outlined in reference [18].

Define a new set M as { i1, i2, …, ik-1, i'k-1 }.

Update Mk by adding the new set M to it, denoted as Mk ← Mk ∪ { M }.return Mk.

Fig 3: Algorithm: S_P_M (SQ_Database, Mat[ ][ ], _, support)

Algorithm:Join_ Apriori (L_{k-1}, Mat[ ][ ], _)

Initialize C_1 as an empty set.

For each pair of sequences P and Q in the set L_{k-1} :

Let P = { i_1 , i_2 , …, i_{k-2} , i_{k-1} }

Let Q = { i_1 , i_2 , …, i_{k-2} , i'_{k-1} }

Assuming D(i_{k-1} , i'_{k-1}) represents the semantic distance between i_{k-1} and i'_{k-1} , and it's derived from a matrix M.

Define a new sequence c = { i_1 , i_2 , …, i_{k-1} , i'_{k-1} }.

Update C_k by adding the new sequence c to it, i.e., C_k = C_k ∪ { c }.

Fig 4: Algorithm: Apriori (L_{k-1}, Mat[ ][ ], _)

Join_Apriori is a proposed algorithm that improves the efficiency of the Apriori algorithm. It takes two input candidate sets and generates a sequence matrix and

distance of two objects. The Apriori algorithm is a frequent itemset mining algorithm that generates candidate itemsets and prunes infrequent ones. Figure 4 shows the pseudocode of the Apriori algorithm.

5. NEXT PAGE REQUEST PREDICTION

One way to solve the tradeoff between accuracy and complexity is by incorporating semantic information directly into the transition probability matrix of lower-order Markov models. Our proposal is to use this semantic information as a criterion for pruning states in higher-order (k > 2) Selective Markov models. We compare the precision and model size of this approach with both semantic-rich and traditional Markov models. By integrating semantic information, this method also resolves any conflicting predictions.

The Markov Model is a recommended method for generating accurate and semantically meaningful predictions without requiring complex calculations for all-Kth-order models. This model relies on the semantic distance matrices, Weight Matrix W and Transition Matrix P.

6. EXPERIMENTAL RESULTS:

All trials were conducted on a computer with an Intel B960 processor, 4GB of RAM, and Windows XP Professional. SQL Server and Visual Studio.NET were used to code the programs.

We utilized the log file of the web server in the experiment. Figure 1 describes the information that this file contains. The log file is about 7 megabytes in size. The terms "laptop" (prize, model, firm, screen size) are part of the domain's ontology model.

The S_P_M algorithm was executed with a minimum support count of 0.01 and a maximum semantic distance of 10.

Table 1.1 No. hits by particular host

Host URL	Hits
www.icicilombard	603
172.168.10.170	34
www.gmail.com	5649
www.cricinfo.com/circket	2340

Users access a given host several times, as shown in Table 1.1. As an example, in Figure 5, the most significant number of hits for the website www.icicilmbard is 8685, and the minimum number of hits is 6. Most Valuable Customer Based on Request The two servers utilized were 172.16.2.167 and 172.16.10.167. Figure 5 displays the loads of all the servers.

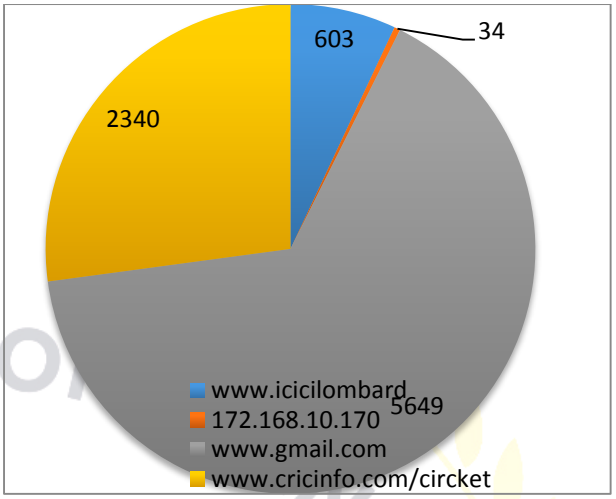


Fig 5: Top Client on the bases of Request

We utilized two servers: 172.16.2.167 and 172.16.10.167. The server load is depicted in Fig 5.

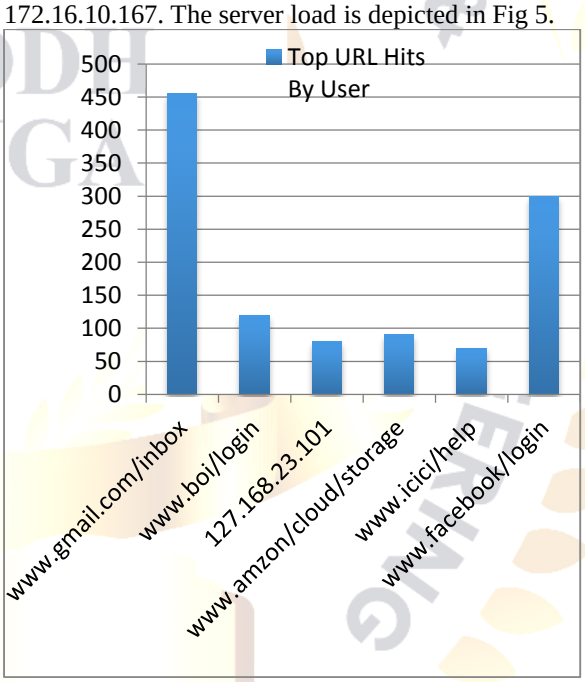


Fig 6. Top URL of Website

7. CONCLUSION

Mining server logs is comparable to a web usage mining model. Improved client connections, more efficient system performance, and user-friendly website design are all outcomes of web use mining. Our offline recommender system here uses Markov mode to forecast the following page. We proposed a novel approach to internet usage mining that uses semantic data throughout. Phase two of finding patterns involves using a pattern mining technique to trim and make counting easier and computing the semantic distance matrix. Superior semantic annotation for data extraction from the web. Rather than relying on complex hybrid-

order Markov models, which can lead to misleading predictions, we build a first-order Markov model while mining and adding semantic information to it. This way, we can deliver an educated lower-order Markov model. This is utilized to predict future page requests. It is possible to incorporate.

- This study is currently being done offline but has been improved to a live log analysis.

- Programming concepts such as parallel tasking and multi-threading might further improve it to even higher levels of performance.

REFERENCE

[1] B. Mobasher, Robert Cooley and JaideepSrivastava, (2000) "Automatic personalization based on 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 and applications 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 coded pre-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 Usage Mining", 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 recommender systems. In Staab, S. and Studer, R., editors, Handbook on Ontologies, International Handbooks 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) " Incorporating Concept 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 Usage profiles ", 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 usage mining and next page prediction", Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM), Hong Kong, November 2-6, 2009, pp. 1677-1680.

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

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