International Journal for Rapid Research in Engineering Technology & Applied Science
Vol VIII Issue IV April 2022
ISSN Online:- (2455-4723)

Implementation to Find Navigational pattern of Log Files Using Hadoop Technology

Khushbu Ankil*, Prof. Mohit Jain**
BM College , RGTU Bhopal, Indore, 452001,India*
BM College , RGTU Bhopal, Indore, 452001,India**
ankilkhushbu@gmail.com*,bmctmohitcs@gmail.com**

Abstract
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. 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 to improve the performance of the servers, caching is used where the frequently accessed pages are stored in proxy server caches. Pre-fetching of web pages is the new research area which when used with caching greatly increases the performance. In this paper, a better algorithm for predicting the web pages is proposed. Clustering of web users according to their location using clustering is done and then each cluster is mined using FP-Growth algorithm to find the association rules and predict the pages to be pre- fetched for storing in cache.

Keywords: Web Usage Mining, Semantic Web, Domain, Sequential Pattern Mining, Recommender Systems, and Markov Model, Prediction, web log.

1. INTRODUCTION

1.1 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

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 pagerequested etc

1.2 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.

- Server Log: the server stores data regarding requests performed by the client, thus data regard generally just one source.ServerLog details are given in Fig 1.
- Client Log : it is the client itself which sends to a repository information regarding the user's behaviour (can be implemented by using a remote agent (such as Javascripts or Java applets) or by modifying the source code of an existing browser (such as Mosaic or Mozilla) to enhance its data collection capabilities.);
- 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.

1.3 Domain Knowledge:

The Pattern association or relationship among all this data can provided information.Inforamation can be converted into knowledge about historical patterns and future.Domain Knowledge consist of information about the data that is already available either through some other discovery or form a domain expert.Domain knowledge classify into three classes are Hierarchical Generalization Tree(HG Tree),Attribute Relationship Rule(AR-Rule) and Environment Based Constraints (EBC).

1.4 Ontology:

Popular definition of "ontology is the specification of Conceptualization". Ontology compartmentalizes the variable needed for some set of computations and

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...	MLXILAPP28	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...	MLXILAPP28	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...	MLXILAPP28	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...	MLXILAPP28	172.16.2.167	POST	/Ipartner.Qu...	trueClientIp:...	443	125.252.22...	HTTP/1.1
C:\Users\A...	185	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXILAPP28	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...	MLXILAPP28	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...	MLXILAPP28	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...	MLXILAPP28	172.16.2.167	GET	/Content/ilo...	eventID=tes...	443	23.57.75.56	HTTP/1.1
C:\Users\A...	189	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXILAPP28	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...	MLXILAPP28	172.16.2.167	GET	/Content/ilo...		443	125.56.222....	HTTP/1.1
C:\Users\A...	190	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXILAPP28	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...	MLXILAPP28	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...	MLXILAPP28	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...	MLXILAPP28	172.16.2.167	GET	/Content/ilo...	eventID=tes...	443	23.57.75.56	HTTP/1.1
C:\Users\A...	194	05/11/2012 ...	01/01/2000 ...	W3SVC171...	MLXILAPP28	172.16.2.167	GET	/WebPages...	TSM Hidd...	443	125.252.22...	HTTP/1.1

Domain Ontology is a representation by a set of concepts C within the domain and the relations R among them .For example the set C contains Product, Purchase, Supplier, and Warehouse as some of the concepts of the domain ontologyconsidered in our Model.

2. RELATED WORK

Sneha Y.S at al in this paper has used OWL technology to add semantics to the existing navigational paths. 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.

Amit Bose et al, proposed a framework for personalization combining usage information and domain knowledge based on ideas from bioinformatics and information retrieval. Unlike our model, these works do not integrate the domain knowledge of the Web application in all phases of Web usage Mining.

J Vellingiri in at alExamination 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.

The studies presented in , extracted domain level objects from user sessions and created a user profile for each user by aggregating these objects according to their weights and a merge function. It is assumed that there already exists a domain level ontology for the Website and merge functions have already been defined on every attribute of objects.

Seidenbergdeveloped a methodology for extraction of related concepts from GALEN based on one or more classes given as input by the user Pirolli and Pitkow's research in in addition to Sarukkai in lead to the use of higher order Markov models for link preiction. The order of a Markov model corresponds to the number of prior events used in predicting a future event.So, a k^{th}-order Markov model predicts the probability of the next event by looking at the past k events.

3. PROPOSED SYSTEM

We propose the use of domain ontology, and its integration in a complete web usage mining system, targeting web recommendation and web usage mining. Proposed architecture is

Proposed architecture divides into three phases:

In the first phase is preprocessor, a clean server-side web log is provided for preprocessing. This log contains all information about server. Web log maintain all the incoming request by users page views and items pi in the web log are mapped to their ontology concepts turning the web log into a sequence database D of semantic thus enriching the web log with semantic information.

The second phase of proposed architecture is Pattern Discovery. The rich sequence database D, is fed into this process. We proposed two Algorithms S_P_M and Join_Apriori for sequence data base in which the semantic information from the preprocessor phase is used to prune candidate sequence and sequence less than supporting count.These proposed algorithms show , with experimental results, that laptop base

domain ontology used in the web uses mining core of any algorithm S_P_M, and it improved its effectiveness and performance , without compromising the quality of the frequent patterns under few conditions.

Third phase of proposed architecture that uses them to generate semantics association rules and n top recommendations. This phase includes techniques that allow further reporting and filtering of results. these results of this phase represents them as a set of recommended items, to be used in web site restructuring or marketing campaigns or to the active user, or as decision making recommendationsto the administrator.

4. SEQUENTIAL PATTERN MINING

A proposed algorithm S_P_M , a variation of Apriori algorithm, improved by semantic information for generating frequent sequences and Join_Apriori() for generate candidate is described. S_P_M generates. A proposed Algorithm S_P_M taken inputs are sequence database (SQ_Database),Sequence matrix Mat[ ][ ] and minimum support and out in form of frequent sequences for rich semantic.Algorithm: S_P_M (SQ_Database, Mat[ ][ ], _, support) is shown in figure 3.

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

M1=Ø

for all P,Q _ Sk-1

with P = { i1,i2,….ik-2, ik-1}

and Q = {i1,i2,…..ik-2, i'k-1}

and D (ik-1,i'k-1) _ _ /* D(ik-1 , i'k-1) :defines the semantic distance between ik-,i'k-1 */

/* _: Maximum Semantic Distance -Maximum allowed semantic distance between any

two semantic objects. _ is a user defined value and can be determined as specified in [18]. D(ik-1 , i'k-1) is derived from M */

M = {i1, i2,…..ik-1,i'k-1}

Mk _ Mk U { M }

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

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

C1=Ø

for all P,Q _ Lk-1

with P = { i1,i2,….ik-2, ik-1}

and Q = {i1,i2,…..ik-2, i'k-1}

and D (ik-1,i'k-1) _ _ /* D(ik-1 , i'k-1)

:defines the semantic distance between ik-,i'k-1 */

D(ik-1 , i'k-1) is derived from M */

c = {i1, i2,…..ik-1,i'k-1}

Ck _ Ck U { c }

return Ck.

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

Another Proposed algorithm is Join_Apriori which take two input candidate set and calculate sequence matrix and distance of two objects. Algorithm: Apriori (Lk-1, Mat[ ][ ], _) is shown in figure 4.

5. NEXT PAGE REQUEST PREDICTION

The integration of semantic information directly in the transition probability matrix of lower order Markov models, was presented as a solution to this tradeoff problem . This integration also solves the problem of contradicting prediction. , we propose to use semantic information as a criteria for pruning states in higher order (where k > 2) Selective Markov models , and compare the accuracy and model size of this idea with semantic-rich markov models and with traditional Markov models.

Markov Model as a proposed solution to proved semantically meaningful and accurate predictions without using complicated all K^{th}order.The semantic distance matrix Weight Matrix W and Transition Matrix P is directly used in markov model

6. EXPERIMENTAL RESULTS:

All the experiments performed on system Intel B 960 processor,4GB RAM Windows XP Professional OS. Programs coded on .Net Framework with sql server as database

In Experiment we used web server's log file. In this file contain information is describe in figure 1. Log file Size 7MB approximately. The ontology model of the domain includes the concepts Laptop (prize,model,company,screensize).

The Algorithm S_P_M was run with Support count(Minimum) = 0.01 and Semantic Distance(Maximum) =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

Above table 1.1 shows that how many time particular host is access by users. For example www.icicilmbard hits 8685 which is maximum hits and minimum hits 6 on host www.icicilmbard.download.akamai.course

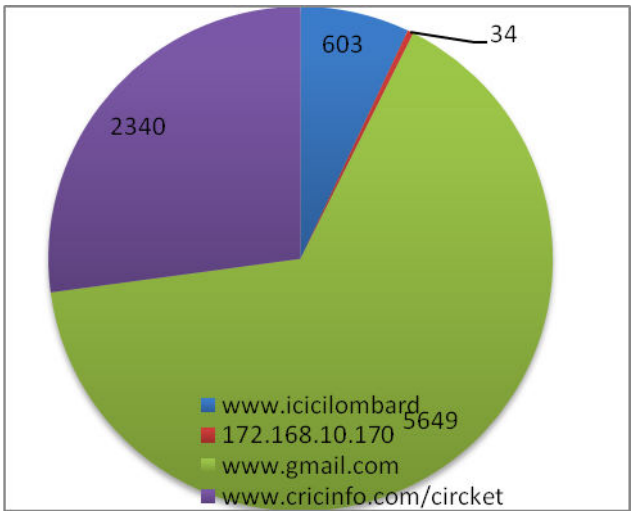

Fig 5: Top Client on the bases of Request

We used two servers 172.16.2.167 and 172.16.10.167.Load of each server is show in Fig5.

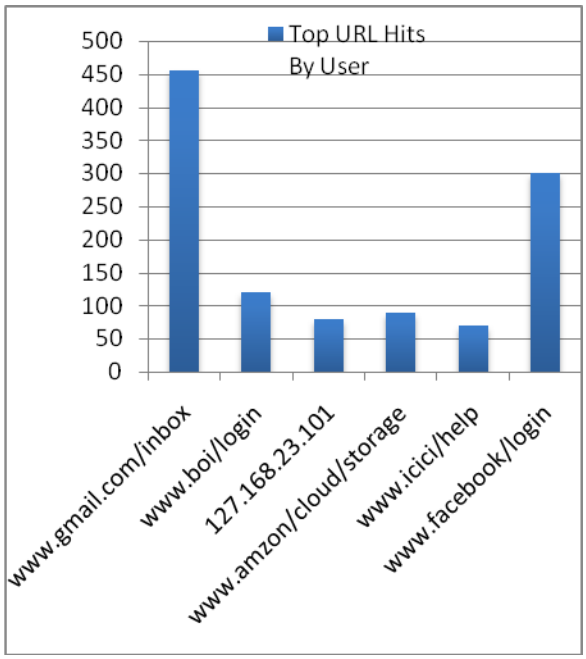

Fig 6.Top URL of Website

7. 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. In this paper we suggest offline recommender system using markov mode for next page prediction. We proposed new framework that integrates semantic information into all the phases of web usage mining. second phase pattern discovery phase that calculate semantic distance matrix and pattern mining algorithm to prune and support counting. Semantic annotation in information extraction on web in a better and efficient way.. 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.

In Future work can be

- Enhanced to live log analysis as currently this analysis is of off line analysis.
- Also it can be further enhanced to greater performance if we use parallel tasking or multi threading concept in programming.

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.