Web Log based Analysis of User's Browsing Behavior

Ashwini ladekar
B.E Student,
Department of Computer,
JSPM's BSIOTR
,Wagholi,Pune, India

Pooja Pawar
B.E Student,
Department of Computer
JSPM's BSIOTR
,Wagholi,Pune, India

Dhanashree Raikar
B.E Student,
JSPM's BSIOTR
,Wagholi,Pune, India

Jayashree Chaudhari
Department of Computer,
JSPM's BSIOTR
,Wagholi,Pune, India

ABSTRACT

In The increasing craze of internet has geared a number of modern firms for using web technology in their day to day lives. A remarkable ability to analyze web log data is provided to them by the web mining technology, that are putatively full of information, but frequently lacking with meaningful information. This signifies the need to the development of an inference mechanism which is advanced and that draws out richer implication from mining's output. This paper presents a web mining algorithm that aims at amending the interpretations of the draft's output of association rule mining. This algorithm is being tremendously used in web mining. The results obtained prove robustness of the algorithm proposed in this paper.

Keywords

Page Interest Estimation, Web log, data mining, Apriori algorithm.

1. INTRODUCTION

We With the ample amount of information residing on the World Wide Web (WWW), the problem of retrieving meaningful information from the Web has obtained attention between researchers in mining of the data. Web services are becoming a gaining importance for oraganizations due to today's agonistic business environment, not only for collecting data but also for discovering patterns from the collected data[3-5]. Business strategies can then be developed from the collected data on the basis of the knowledge gained. To create loyal customers and achieve competitive advantage most of the oraganizations are implementing value-added services on the Web. So to retain long-term relations with the customers, companies have realized the importance of providing personalized products and services. A variety of methods focused on discovering individuals needs are achieved through personalization[6-7]. Web mining helps to retrieve such knowledge that is useful for personalization and improved services if the Web[8].

The process of discovering knowledge from web is known as the Web Mining. It is the procedure to apply the data mining techniques to retrieve meaningful knowledge from ample amount if information present on the web[9]. The objective of web mining being same as that of data mining, both search for meaningful knowledge from the web log and the database[10-11]. Data mining deals with more structural database and extraction of knowledge[13].

Web mining is divided into three areas majorly[14]:web structure, web content and web usage mining. Inferring data from organization and links on web is known as Web structure mining [15], the process of extracting data from multiple number of web document contents is known as web content mining[15] whereas automatic discovery of user patterns from web servers is known as the web usage mining. The web content mining [17] uses the web search engines, its main role is to discover the contents from the web which are as per the user's requirements and constraints.

Recently, web content mining has migrated to agent based[18] and database driven mining by the usage of traditional search engine. More intelligently the consideration of domain characteristics and user profiles are taken by agents for search towards more relevant web contents. They also support the users to figure out the discovered contents of the web. The external, internal structures and the URL are most of the agents for web structure. The investigation of the hyperlinked relationships between web pages is known as external structure mining[19], while the analysis of relationships of information within web pages is known as internal structure mining. The extraction of URL that is important to decision maker's purpose is known as URL mining[20].

Web usage mining is to apply the concept of data mining to the web log file data, and automatically discover user access patterns towards a specific web page. Another source for the web usage mining is web logs, which contain information about the referring pages for each page reference. Results gained from web usage mining deliver decision makers with important data about the lifetime value of the customers, cross-marketing strategies across products. Among other things, the web usage mining [21] helps organization analyze user access patterns to targeting ads or web pages and provide with the restructuring of a web site to generate more effective management of the workgroup communication and infrastructure of organization.

This paper contracts with web usage mining, for which multiple data mining techniques such as statistical analysis, clustering, classification, association rules, sequential pattern discovery, and dependency modeling have been smeared to web server logs. Amongst them, association rules forms the focus of this study.

2. ASSOCIATION RULES MINING

All Among the various data mining techniques, association rules mining [22] is popular in marketing intelligence fields. However, association rule mining is applied to the task of web mining. The derivation of association rule mining is done from the the approach of data mining, known to well discover meaningful rules from large collections of information. The association rule mining emphases on finding association rules such as X→Y with support A% and confidence B%, where X and Y are the item sets, from the database transactions. The rule specifies that the records of transactions in database that contain X tend to contain Y. The support A % of the rule X→Y is the transactions percent that contain X. The rule X→Y grips in the database transactions with confidence B%, if B% of transactions that support X also support Y. Thus, the support signifies the usefulness of the discovered association rules, while the confidence specifies certainty of the perceived association rules. Therefore, the goal of the association rule mining is to find out all rules that have support and confidence larger than the user-specified minimum support and minimum confidence. This definition of the association rule mining specifies that numerous redundant rules may be found, which also substantiates the algorithm presented in this paper to cultivate a new web mining approach to enrich the interpretation from the results of web mining like association rule mining [23].

Apriori algorithm [24], which is espoused in this paper as on the association rule mining techniques, is one of the widespread association rule mining techniques. Apriori algorithm functions in two phases. In the first phase, all transaction sets are produced with the downward closure property of support. In the second phase, the association rules are produced from all the transaction sets.

Our proposed algorithm is poised of the support number one, the support number two and Rel-confidence. The first stage is to abstract association rules from the web log database in which all the web surfing log files users generated on a target web site are stored. The purpose why we use association rules is that a web site shows a set of interrelated hyperlinks each of which may signify either a menu, a data set , a figure, or another web document, etc., and that the objective of the web mining is usually to find out the hidden informative relationship between those inter-related hyperlinks users visited while web surfing.

3. DESCRIBING METHODOLOGY

To discover the unknown informative relationship between those interrelated hyperlinks users visited while surfing the web, a new algorithm of determining association rules that contain the support number one, the support number two and Rel-confidence is offered in this paper. K-candidate item sets, where k is the number of data items in its each itemset, is denoted as Ck. Ik denotes the candidate itemset from Ck. Ik.support denotes the support for candidate itemset Ik. The algorithms for producing frequent itemsets that include semi-frequent itemsets are provided as follows:

```
Algorithm1 (function for generating candidate itemsets,
candi-gen(Lk-1))
    if (k=2) then insert into Ck;
        select p.iteml; q.item1 from L1
    else
        insert into Ck
        select p.iteml; p.item2;…; p.itemk-1; q.itemk-1
        where p.item1=q.item1, …, p.itemk-2=q.itemk-2,p.itemk-1 < q.itemk-1
Algorithm2:
  D: Database (a set of transactions)
  I = i1, i2, . . . , in: a set of data items accessed by all transactions in D
  C1 ={ i1, i2, . . . , ik};
  for ( ik∈I)
      if (ik.support≥1st support) then ik∈L1
      else
          if (ik.support≥2nd support) then ik∈L1
          else remove ik from C1
end
for (k=2; Lk-1 != Ø; k++) do
          Ck = Candi-gen(Lk-1);
          T = { };
          for all transactions t in D do
              Ct ={ }; /* Ct ={Ik | Ik∈Ck and all data items in Ik are included by transaction t}*/
          for all Ik∈Ck do
            if (all data items in Ik are included by transaction t )
              then Ct = Ct∪Ik
      end
              for all candidates Ik∈Ck do
      Ik.count++
              End
                T = T∪Ct
      end
for all candidates Ik in T do
   if (Ik.count≥1st support) then Lk ={ Ik| Ik∈T and Ik.count≥1st support}
   else
Rel-confidence (i1, i2, … , ik) = max(sup(i1, i2, … ,ik)/sup(i1), sup(i1, i2, … , ik)/sup(i2), …,sup(i1,i2, … , ik)/sup(ik))
   if (Ik.count≥2nd support and Ik. max-conf≥min-conf)
      then Lk ={ Ik| Ik∈T and Ik. max-conf≥min-conf }
      else remove Ik from Ck
  end
  Answer =∪k Lk
End
```

The original web log data is adulterated by various types of irrelevant, redundant and void data, for instance, filename suffixes like map, jpeg, gif and count. cgi, etc. For mining a significant set of association rules from the web log database, the first step is to cleanse the target web log data so that the preprocessed web log data becomes extra traceable to particular web documents, and free from meaningless noisy data like the local caches, proxy servers, corporate firewalls, etc. After completion of the preprocessing task, the target web log data then becomes clean and ready for further web mining procedures. It is ordinary to sort out the preprocessed web log data file by the same IP address because an unidentified user leaving some trails on the web log database can be identified uniquely by his/her own IP address. Table 1 depicts raw web log database and the corresponding preprocessed web log database.

The web mining algorithm we espoused here is the one projected above, which is built based on the known Apriori Algorithm. Based on the preprocessed web log database in Table 1. The equivalent association rules were mined with a threshold of 60% confidence. Table 1 shows an extract of the derived association rules.

4. EXPERIMENTAL EVALUATIONS

The methods and techniques which are used for estimating the interest of the user are web usage mining, DFS and apriori algorithm using association rule are being explained under in this section. Various modules such as server module, client module.

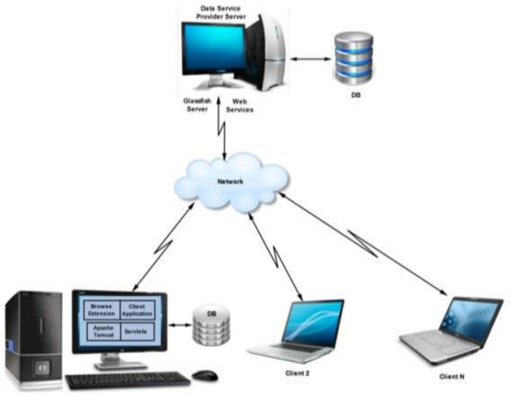


Fig 1: Proposed System

Steps on client and server side :

Client:

- The client machine's browser tracks user's behavior of browsing and the time details that are relates.
- The extension makes use of servlets for logging data on database(local)by making use of apache tomcat.
- The client application supports the feature of storing the browsing history(local), the filtering data and it applies user statistics to the cloud.
- The user can thus estimate its time that was needed to open a particular page or site.

Server:

- Server obtains log which consists of user's action from the client machines.
- Server applies mining algorithms so that better business intelligence solutions can be .
- Server application serves with a graphical analysis of user's usage patterns.

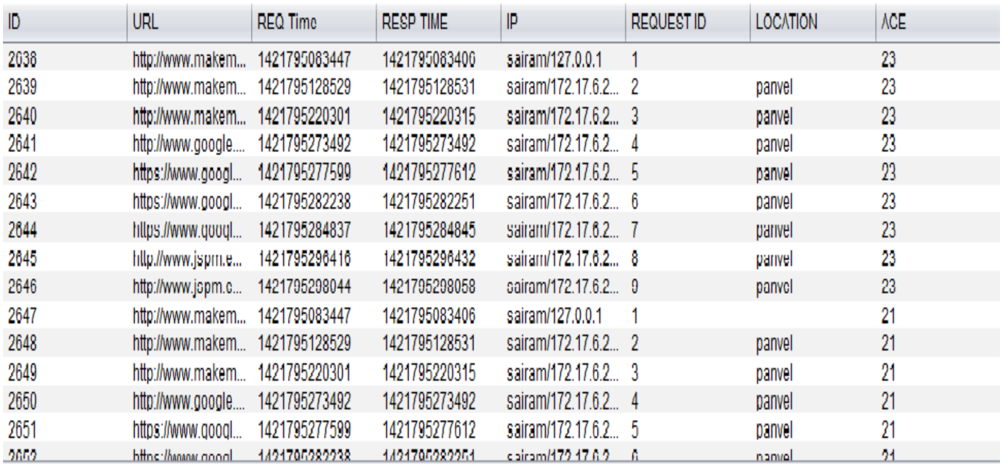
We have also created an extension for the browser like chrome, local log manager, etc. Their functions are as follows

Chrome Extension:
It helps to track the user actions on the web and stores the data in the web log database. Whenever any user enters an URL in the address bar, its details are stored immediately in the local database.

Local Log Manager:
It contains the browsing details of a particular user such as ip address, url, request and response time.

4.1 Datasets

Table 1. Summary description of datasets



ID	URL	REQ Time	RESP TIME	IP	REQUEST ID	LOCATION	ACE
2038	http://www.makem...	1421795083447	1421795083406	sairam/127.0.0.1	1		23
2639	http://www.makem...	1421795128529	1421795128531	sairam/172.17.6.2...	2	panvel	23
2640	http://www.makem...	1421795220301	1421795220315	sairam/172.17.6.2...	3	panvel	23
2641	http://www.google....	1421795273492	1421795273492	sairam/172.17.6.2...	4	panvel	23
2642	https://www.googl...	1421795277599	1421795277612	sairam/172.17.6.2...	5	panvel	23
2643	https://www.googl...	1421795282238	1421795282251	sairam/172.17.6.2...	6	panvel	23
2644	https://www.googl...	1421795284837	1421795284845	sairam/172.17.6.2...	7	panvel	23
2645	http://www.jspm.e...	1421795296416	1421795296432	sairam/172.17.6.2...	8	panvel	23
2646	http://www.jspm.c...	1421795298044	1421795298058	sairam/172.17.6.2...	9	panvel	23
2647	http://www.makem...	1421795083447	1421795083406	sairam/127.0.0.1	1		21
2648	http://www.makem...	1421795128529	1421795128531	sairam/172.17.6.2...	2	panvel	21
2649	http://www.makem...	1421795220301	1421795220315	sairam/172.17.6.2...	3	panvel	21
2650	http://www.google....	1421795273492	1421795273492	sairam/172.17.6.2...	4	panvel	21
2651	https://www.googl...	1421795277599	1421795277612	sairam/172.17.6.2...	5	panvel	21
2652	https://www.googl...	1421795282238	1421795282251	sairam/172.17.6.2...	6	panvel	21

5. ACKNOWLEDGMENTS

We would like to express gratitude to our staff, family and friends for guiding us throughout this paper pub-lication process and providing us with excellent sup-port.

6. REFERENCES

[1] B. Hay, G. Wets, K. Vanhoof. Segmentation of visiting patterns on Web sites using a sequence alignment method. Journal of Retailing and Consumer Services, 2003, 10 (3) :145–153.

[2] K.A. Smith, A. Ng. Web page clustering using a self-organizing map of user navigation patterns. Decision Support Systems, 2003, 35 (2): 245–256.

[3] J.D. Martin-Guerrero, A. Palomares, E. Balaguer-Ballster, E. Soria-Olivas, J. Gomez-Sanchis, A. Soriano-Asensi. Studying the feasibility of a recommender in a citizen Web portal based on user modeling and clustering algorithm. Expert Systems with Applications, 2006, 30 (2) :299–312.

[4] S.K. Rangarajan, V.V. Phoha, K.S. Balagani, R.R. Selmic, S.S. Iyengar. Adaptive neural network clustering of Web users. IEEE Computer, 2004, 37 (4) :34–40.

[5] R.J. Kuo, J.L. Liao, C. Tu. Integration of ART2 neural network and genetic K-means algorithm for analyzing Web browsing paths in electronic commerce. Decision Support Systems, 2005, 40 (2) :355–374.

[6] C. Shahabi, F. Banaei-Kashani. Efficient and anonymous web-usage mining for Web personalization. Journal on Computing, 200315 (2):123–147.

[7] Q. Yang, J.Z. Huang, M. Ng. A data cube model for prediction-based Web prefetching. Journal of Intelligent Information Systems, 2003, 20(1) :11–30.

[8] F.M. Facca, P.L. Lanzi. Mining interesting knowledge from weblogs: a survey. Data and Knowledge Engineering, 2005, 53 (3):225–241.

[9] G. Paliouras, C. Papatheodorou, V. Karkaletsis, C.D. Spyropoulos, V. Malaveta. Learning user communities for improving the services of information providers. 1998, Comput. Sci, 1513: 367–384.

[10] J.S. Park, M.S. Chen, P.S. Yu. Using a hash-based method with transaction trimming for mining association rules. IEEE Trans. Knowledge Data Engng, 1997, 9 (5): 813–825.

[11] M.S. Chen, J.S. Park, P.S. Yu. Efficient data mining for path traversal patterns. IEEE Trans. Knowledge Data Engng, 1998, 10 (2) :209–221.

[12] J. Borges, M. Levene. Data mining of user navigation patterns, in: Web Usage Analysis and User Profiling, Lecture Notes in Computer Science, Springer, Berlin, 2000, 1836: 92–111.

[13] Kuo, R. J., Chen, J. H., Hwang, Y. C. An intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural network and artificial neural network[J]. Fuzzy Sets and Systems, 2001, 118(1):21–45.

[14] Weigen, A. S., Rumelhart, D. E.Generalization by weight-elimination with application to forecasting. Advances in Neural Information Processing Systems[J]. 1999, 3:875–882.

[15] Chen, M, S, Han, J. Data mining: an overview from a database perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 8(6): 866–883.

[16] Schafer, J. B., Konstan. E-commerce recommendation application[J]. Journal of Data Mining and Knowledge Discovery, 2001, 16:125–153.

[17] Giudici, P, Passerone, G. Data mining of association structures to model e-shopper behavior. Computational Statistics and Data Analysis[J]. 2002, 38:533–541.

[18] P. Kumar, R.S. Bapi, P.R. Krishna. A sequence clustering algorithm for Web personalization, International Journal of Data Warehousing and Mining, 2007, 3 (1):29–53.

[19] P. Kumar, P.R. Krishna, R.S. Bapi, S.K. De. Rough clustering of sequential data. Data and Knowledge Engineering, 2007, 63 (2) 183–199.

[20] R. Sen, M.H. Hansen. Predicting Web users next access based on log data. Journal of Computational and Graphical Statistics, 2003, 12:143–155.

[21] S. Park, N.C. Suresh. Performance of fuzzy ART neural network and hierarchical clustering for part-machine grouping based on operation sequences. International Journal of Production Research, 2003, 41 (14) :3185–3216.

[22] Changchien, S. Mining association rules procedures to support on-line recommendation by e-shoppers and products fragmentation [J]. Expert Systems with Applications, 2001, 20(4):325–335.

[23] Song, H, Kim, J. Mining the change of e-shopper behavior in an Internet shopping mall[J]. Expert System with Applications, 2001, 21(3):157–168.

[24] Anand, S, Patrick, A. A data mining methodology for cross-sales[J]. Knowledge-Based Systems, 2006, 10:449-461.