An Efficient Hierarchical Frequent Pattern Analysis Approach for Web Usage Mining

G. Sudhamathy
Department Of Computer Applications,
Velammal College of Engineering & Technology,
Madurai 625 009, India

C. Jothi Venkateswaran
Department of Computer Science,
Presidency College (Autonomous),
Chennai 600 025, India

ABSTRACT
Web usage mining aims to discover interesting user access patterns from web logs. Web usage mining has become very critical for effective web site management, creating adaptive web sites, business and support services, personalization and so on. In this paper, an efficient approach for frequent pattern mining using web logs for web usage mining is proposed and this approach is called as HFPA. In this approach HFPA, the proposed technique is applied to mine association rules from web logs using normal Apriori algorithm, but with few adaptations for improving the interestingness of the rules produced and for applicability for web usage mining. This technique is applied and its performance is compared with that of classical Apriori-mined rules. The results indicate that the proposed approach HFPA not only generates far fewer rules than Apriori-based algorithms (FPA), the generated rules are also of comparable quality with respect to three objective performance measures, Confidence, Lift and Conviction. Association mining often produces large collections of association rules that are difficult to understand and put into action. In this paper effective pruning techniques are proposed that are characterized by the natural web link structures. Experiments showed that interestingness measures can successfully be used to sort the discovered association rules after the pruning method was applied. Most of the rules that ranked highly according to the interestingness measures proved to be truly valuable to a web site administrator.

General Terms
Pattern Recognition.

Keywords
Web Usage Mining, Web Logs, Web Personalization, Association Rules, Interestingness Measures.

1. INTRODUCTION
When web sites are visited by users, web log files are generated on web servers and these files contain a huge amount of information. Data mining techniques like Clustering and Association Rule mining can be applied on this data to discover interesting information which when analyzed by the web site maintenance engineer can reveal vital information required for web site improvement and there by attract more users to access the web site. Web usage mining has various application areas such as web pre-fetching, link prediction, site reorganization and web personalization.

Originally, association rule mining algorithms were applied for Market Basket Analysis which contained transaction data. The transaction data may include many records of which each record has a transaction id and a list of items purchased during that transaction. But when the same Apriori algorithm has to be applied for web log data, it has to be transformed to the same format as that of the transactions [11]. To make this happen, web log data has to be cleaned, split and preprocessed into sessions and the list of web pages navigated during each session. Once this data transformation is done, association rules can be mined as it is done for market basket analysis. However, the threshold selection, pruning method, interesting measures used and ranking of the rules needs some modifications to suit the needs of web usage mining [8].

The association rule mining algorithm can find all rules that satisfy defined constraints, they often result in a large set of rules that is difficult to exploit and find those rules that are truly interesting to the user [1]. Web log data differs from the market basket data in the sense that it contains a large number of tightly correlated web pages due to the link structure of a website. Web pages that are tightly linked together often occur in the same transaction, which is why the generated set of association rules are high and they have very high confidence, but are not truly interesting to the user. So, in this approach the item set that includes directly linked pages are pruned as the interest is only on the information that can prompt actions leading to enhancement of a website and improving the browsing experience for visitors [10].

As a case study to prove the efficiency of proposed approach the web log files of www.eretailstore.biz for a period of six months, 07/2010 to 12/2010 is used. These web log files were cleaned, preprocessed, transformed to match the format suitable for Apriori algorithm. The minimum threshold for Support and Confidence that will suit web logs analysis is set. After acquiring rules from frequent item sets produced by Apriori algorithm, support, confidence, lift and conviction are calculated [2]. Sort by descending order of lift value and then by descending order of conviction. The top ranked rules, say top 20% of the overall rules produced after pruning and threshold limit exceeding are found to be most useful and interesting rules that can recommend better web site reorganization or personalization [3].

The approach is compared with the traditional approach of using just support and confidence to find the interesting rules as in the case of market basket analysis. It was found that the proposed approach HFPA is performing better with respect to computation time, number of rules produced finally, percentage of accuracy of interestingness, memory usage and number of rules reduction by applying the pruning techniques [6].

2. PROPOSED APPROACH

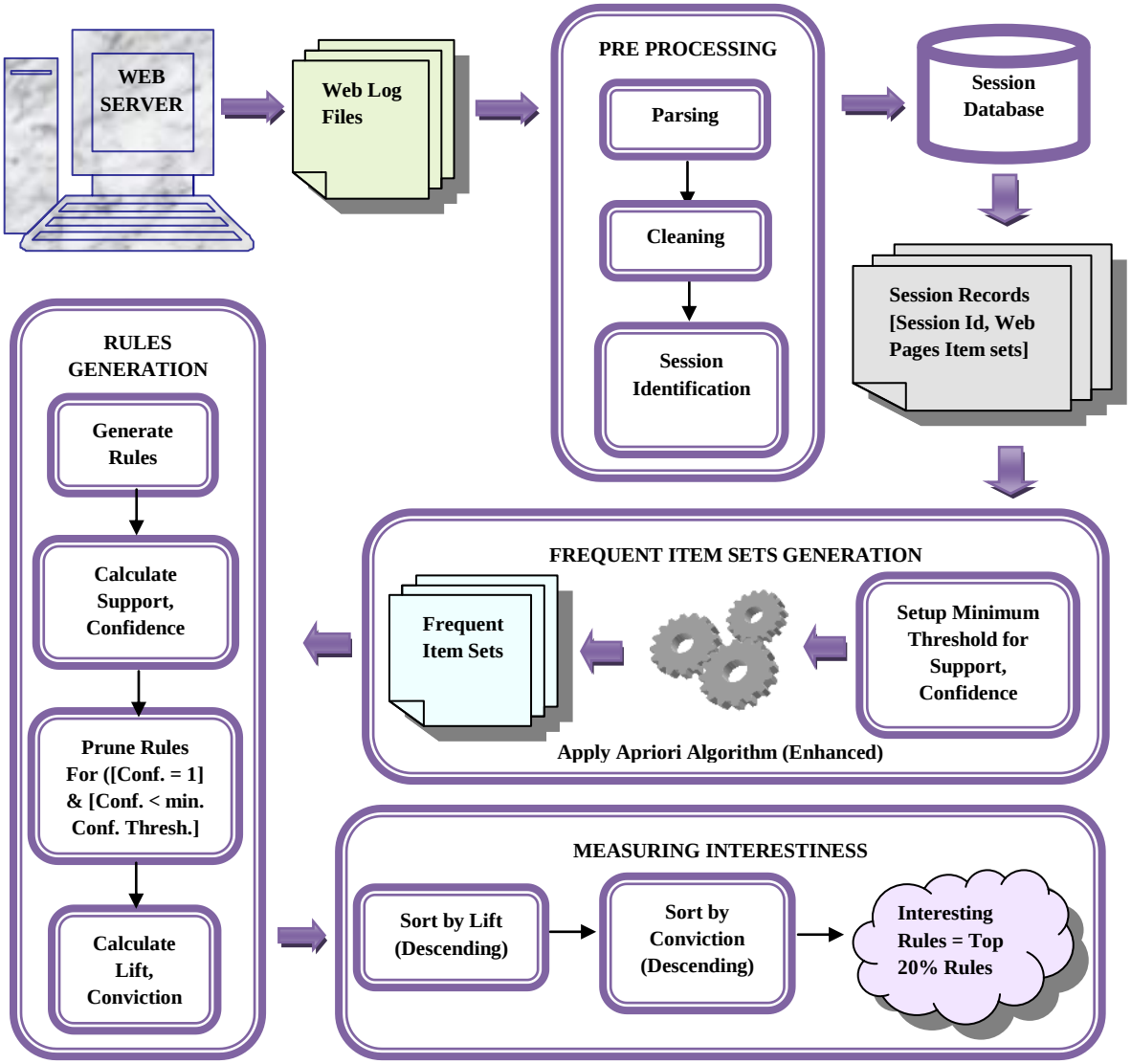

Fig 1: Proposed Approach – HFPA for Web Usage Mining.

Overview of the HFPA is represented pictorially by the figure 1. This approach includes four main steps, namely, Pre-Processing, Frequent Item Sets Generation, Rules Generation and Measuring Interestingness.

In the light of web usage mining, any page of a web site is referred as an item. Consider a web page http://eretailstore.org/logguest.php and say this is mapped to an alias name P1. A candidate item set is any set of web pages in a web site. Say a web site has the web pages P1, P2… P9, then one of the candidate item sets can be {P1, P2, P3, P4, P8}. A session is a set of web pages visited by a user from a client IP address continuously, without a break of 5 minutes. Data set is a set of sessions of a particular website.

Support count of a candidate item set is the measure of how frequently all items in the candidate item set occurs together in the set of all sessions of a data set. Suppose the candidate item set is {P1, P2}, there are 30 sessions totally and of these 12 sessions included the items P1 and P2 together, then the support count of the candidate item set {P1, P2} is 12. Minimum threshold for support is user specified. That is if the user sets the minimum support threshold as 0.1 (or 10%), it means that any candidate item set should occur in at least 10% of the total sessions to be a frequent item set. A candidate item set is said to be frequent item set, if its support exceeds the minimum support threshold. Consider a candidate item set {P1, P2} occurs in 12 sessions out of the total 30 sessions, which means its support is 12/30 = 40% and it exceeds the minimum support threshold and hence it is a frequent item set.

Association Rule is the correlation between the two candidate item sets of the form $X \rightarrow Y$, where $X \cap Y = \varnothing$ and X is called the rule antecedent and Y is called the rule consequent. Support for a rule $X \rightarrow Y$ is the support of the set $X \cup Y$. Also, Support $(X \cup Y)$ = Support$(X \cap Y)$. Confidence of a rule $X \rightarrow Y$ = Support$(X \cup Y)$ / Support(X). That is if a rule $X \rightarrow Y$ in a set of sessions D has confidence = c%, then c% of the sessions in D that contain X also contains Y. Say, there are 30 sessions totally, X is present in 10 sessions and Y is

present in 3 of the 10 sessions where X is present, then the confidence of the rule X → Y is 30%.

Lift of a rule X → Y = [Support(X ∪ Y)] / [Support(X) * Support(Y)]. Conviction of a rule X → Y = [1 – Support(Y)] / [1 – Confidence(X → Y)]. The interestingness of an association rule refers to the practical usefulness of the knowledge discovered by the association rule data mining to the web site administrator or designer so as to enhance the web site structure and there by enhance the business via the web site.

2.1 Pre Processing

During user visit to the web pages in a website, web log files are created in the Web Server in which the web site is hosted. These web log files are extracted from the web server (from the location where it is stored in the web server) and it is preprocessed. The Preprocessing includes the steps of Parsing, Cleaning and Session Identification. The preprocessing step is executed for each web log file at a time.

Actually the web log files are flat text files that contain many space / tab delimited fields. The important fields in any web log file are Data, time, Client IP address, Server IP address, Server Port, URL Visited and User Agent filed that gives details of the browser and operating system versions. The Parsing step does splitting of the text file into specific fields and extracts the required fields into a database table. In this case, we need the fields, date, time, Client IP address, URL Visited and User Agent.

Once these fields are split, extracted and stored in a database table, the extracted records are then cleaned to remove the images, icons and unwanted requests. So delete all records that have .JPEG, .GIF and .CSS files in the URL Visited field. As a result the cleaned database with relevant records is obtained.

The next step in preprocessing is user session identification [16]. The log entries in the web log files are chronologically ordered based on the different user's requests from their client machine to the web server.

The user sessions are identified based on the following assumptions:

1. Each user has a unique Client IP address while browsing the website. The same IP address can be assigned to other users after the user finishes browsing.
2. The couple of client IP address and user-agent are considered for single user identification as different users can come from the same proxy.
3. Any user accessing the website from a unique client IP address and user-agent details will be active in a session only if the user does not exceed the maximum idle time. This maximum idle time is optimally 5 minutes. [That is if the difference between the time of the web log entries from the same client ip address and user agent field is more than 5 minutes then both the log entries belong to different user sessions.]

Thus one user session is the set of records that have the same client IP address and User agent for which the consecutive date and time does not exceed the idle time of 5 minutes. See the below example table 1. for better understanding. In the below table first three log records belong to one user session and the remaining three records belong to the next user session. This is because even though the Client IP address and User agent fields are same the time difference between the third and fourth record is more than 5 minutes. During this time the same machine can be used by another user and hence will belong to a different user session.

Table 1. Example Web Log Records that belong to different User Sessions

Date (yyyy-mm-dd)	Time (hh:mm:ss)	URL Visited	User Agent
2009-10-24	03:46:22	http://eretailstore.org/plist.php	HTTP/1.1 Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+en-US;+rv:1.8.1.6)+Gecko/20070725+Firefox/2.0.0.6
2009-10-24	03:48:33	http://eretailstore.org/noplist.php	HTTP/1.1 Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+en-US;+rv:1.8.1.6)+Gecko/20070725+Firefox/2.0.0.6
2009-10-24	03:51:09	http://eretailstore.org/viewabs.php	HTTP/1.1 Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+en-US;+rv:1.8.1.6)+Gecko/20070725+Firefox/2.0.0.6
2009-10-24	03:58:14	http://eretailstore.org/logguest.php	HTTP/1.1 Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+en-US;+rv:1.8.1.6)+Gecko/20070725+Firefox/2.0.0.6
2009-10-24	04:01:34	http://eretailstore.org/noplist1.php	HTTP/1.1 Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+en-US;+rv:1.8.1.6)+Gecko/20070725+Firefox/2.0.0.6
2009-10-24	04:02:52	http://eretailstore.org/noplist2.php	HTTP/1.1 Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+en-US;+rv:1.8.1.6)+Gecko/20070725+Firefox/2.0.0.6

This is the end of the pre-processing step. As a result of the pre-processing of web log records, a session database which is ready to be used for association rule mining is obtained. This session database will resemble the traditional market basket analysis transaction database that has item sets of the items purchased by the user. In this case the item set will be the set of web pages visited by a user in a session. Thus the session database consists of records that have the session Id and Web Pages Item set as listed in the example table 3.

In the example table 3, the web pages are renamed as P1, P2, … for simplicity and easy usage. The actual web page to its alias names mapping is still maintained in a data dictionary inside the session database. Apart from this the session database also stores the details of the direct web page linkages

as designed for the web site. A sample web page links are shown as in the below figure 2.

Table 2. Example Market Basket Transaction Database Records

Transaction ID	List Of Items Purchased
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3
T901	I3, I5
T902	I2, I5
T903	I1, I5

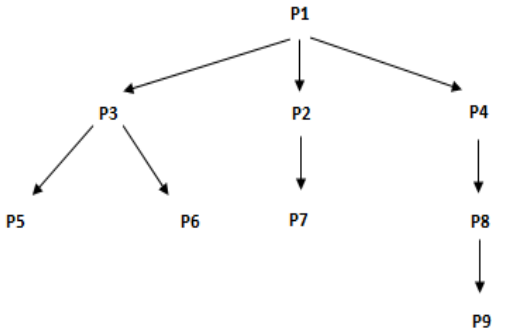

Fig 2: Sample Web Site Link Structure

2.2 Frequent Item Sets Generation

Frequent Item sets can be generated by applying the enhanced version of Apriori algorithm, but before applying this algorithm the minimum threshold for support and confidence is set up. The ideal minimum threshold for support is set to be 10% (or 0.1) and for confidence is set to be 45% (or 0.45) for web usage mining [7].

Recall the definition for frequent item sets. A candidate item set is said to be frequent item set, if its support exceeds the minimum support threshold.

E.g. Candidate item set {P1, P2} occurs in 12 sessions out of the total 30 sessions, which means its support is 12/30 = 40% and it exceeds the minimum support threshold and hence it is a frequent item set.

Before going into the Apriori algorithm, it is required to understand two properties of the algorithm. Firstly the Apriori Property, which states that "Any subset of frequent item set must be frequent". Example - If {P1, P2} is a frequent item set, then its sub sets {P1} and {P2} are also said to be frequent. Secondly, the Join Operation which states that "To find a set of candidate k item sets, it is generated by joining the frequent k-1 item sets with itself". Example - If there are frequent 2 item sets {P1, P2} and {P1, P3}, then the candidate 3 item sets are obtained by joining {P1, P2} and {P1, P3}. The candidate 3 item set formed will be {P1, P2, P3}.

Table 3. Example Web Usage Session Database Records

Session ID	List Of Web Pages Visited
S1	P1, P3, P5, P4, P8, P9
S2	P1, P3, P5, P6
S3	P1, P3, P6, P2, P7
S4	P1, P2, P7
S5	P1, P4, P8
S6	P1, P4, P8, P9
S7	P1, P4, P3, P6
S8	P1, P2, P7, P4, P8
S9	P1, P3, P6, P5
S10	P1, P2, P4, P8, P9
S11	P1, P3, P5
S12	P1, P3, P6
S13	P1, P4, P8, P9
S14	P1, P2, P3, P5
S15	P1, P3, P4, P8
S16	P1, P4
S17	P1, P2
S18	P1, P3, P4
S19	P1, P4, P2, P3
S20	P1, P2, P3, P4, P8
S21	P1, P3, P2, P7
S22	P1, P4, P3, P6
S23	P1, P2, P3, P5
S24	P1, P3, P5, P1, P3
S25	P1, P3, P6, P1, P3
S26	P1, P2
S27	P1, P3, P4, P8
S28	P1, P4, P3, P6
S29	P1, P4, P8, P9
S30	P1, P2, P7

In the enhanced Apriori algorithm of HFPA, consider C_k as the Candidate item set of size k and L_k as the frequent item set of size k.

Pseudocode for Enhanced Apriori Algorithm:

Let L_1 = {Frequent 1 item sets};
For (k=1; L_k != ∅; k++) do begin
  Generate C_{k+1} from L_k using the Join Operation;
  Initialize the count of all candidates in C_{k+1} to 0;
  For each record r in the session database do
    If (number of items in the record r >= k+1) then
      Increment the count of all candidates in C_{k+1} that are contained in r by 1;
    End If;
  End For;
  L_{k+1} = Candidates in C_{k+1} with support greater than the minimum support threshold – Candidates in C_{k+1} that reflect the direct link structure of the web site.
End For;
Return L1 ∪ L2 ∪ … L_k

There are two enhancements done to the normal Apriori algorithm in the above algorithm. Firstly, while Scanning the session database for support count for a candidate item set, skip the records that have less number of items (or pages) than

that is there in the Candidate Item set [5]. This is because; the pages in the candidate item set will not be present in such records that has less number of pages than in the candidate item set. Example - If the support count is calculated for the Candidate item set {P1, P2, P3} and the session database has the records S16, S17 and S26 (refer table 3) that has just two pages, these records can be skipped while calculating the support count for the candidate item set said above. Secondly, prune the item sets that reflect the direct link structure of the web site from the frequent item set list [9]. This is because the rules produced by these item sets will represent the website link correlation and hence are not truly interesting to the user. Example - Consider the frequent item set list includes the item set {P1, P2, P7}, then this has to be removed from the frequent item set list as it reflects the direct web site link and hence will not generate interesting rules (refer figure 2).

Now as a result of applying the above said algorithm the optimal frequent item sets are obtained from which next the rules are generated [14].

2.3 Rules Generation

Consider all the Frequent Item sets except the one frequent item sets for rules generation.

Rules generation is done in two steps. As a first step, for each frequent item set I, generate all non empty subsets of I. As a second step, for every nonempty subset S of I, output the rule S → I – S. Example - Consider the frequent item set {P1, P5}. The sub sets are – {P1} and {P5}. Hence the rules are – {P1} → {P5} & {P5} → {P1}. In the above, note that the item on the left hand side of the rule is called the rule antecedent and the item on the right hand side of the rule is called the rule consequent.

Next calculate the support and confidence for each of the rules as per the definitions said above. The support for all the rules will definitely be greater than the minimum support threshold as item sets are already pruned and its support is less than the minimum support threshold.

But, there can exist many rules whose confidence is less than the minimum confidence threshold. Hence it is required to prune the rules whose confidence is less than the minimum specified confidence threshold. Also the rule whose confidence is equal to one is also pruned as they mean the strong correlation of pages due to link structure of the web site [17]. These rules with confidence = 1 will not be interesting for the user.

After pruning the unwanted rules based on confidence it is required to calculate lift and conviction for the remaining rules. This is done as per the formulae mentioned in the definition above. Now there is a list of rules and their corresponding support, confidence, lift and conviction.

2.4 Measuring Interestingness

From the list of rules with their support, confidence, lift and conviction, it is necessary to arrive at the most interesting rules that can help a web site administrator to improve the web site [4]. For this sort the rules by descending order of lift and then by descending order of conviction and rank them accordingly. The top 20% of the overall pruned rules are found to be the most interesting and expected to be taken up for further action by the web master [12].

3. PERFORMANCE EVALUATION

The proposed approach HFPA has to be evaluated and compared with the traditional market basket analysis way of association rule mining which is called as FPA. To make this possible, a software is developed using java that incorporates all the steps and algorithm as proposed in the approach. As said before, the web log files of www.eretailstore.biz web site are taken for a period of six months from 07/2010 to 12/2010 as a test data to evaluate the proposed approach against the traditional approach. The results proved that the proposed algorithm is seen to perform better in all aspects such as computation time, number of rules produced, percentage of accuracy of interestingness, memory usage and the percentage of rules pruned from the original set of rules.

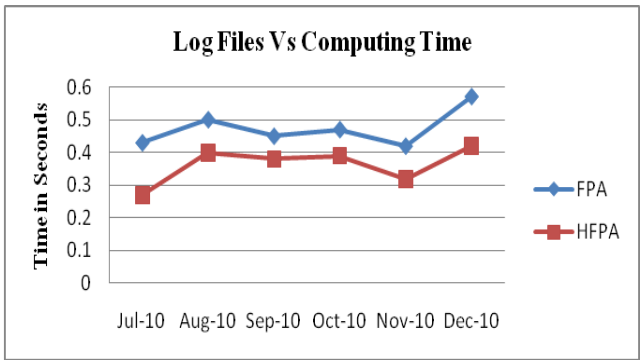

Fig 3: Chart Comparing the Run Times of the HFPA and FPA Approaches

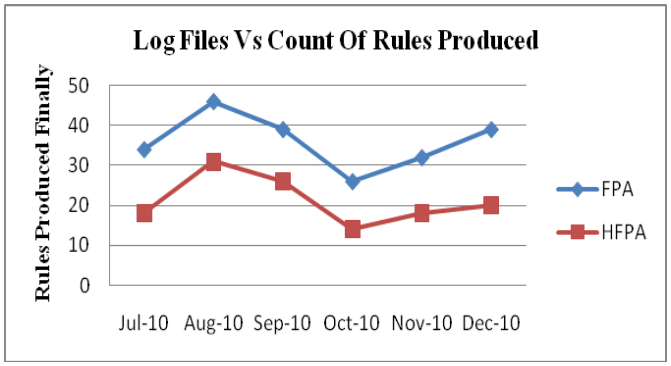

Fig 4: Chart Comparing the Rules Produced Finally by the HFPA and FPA Approaches

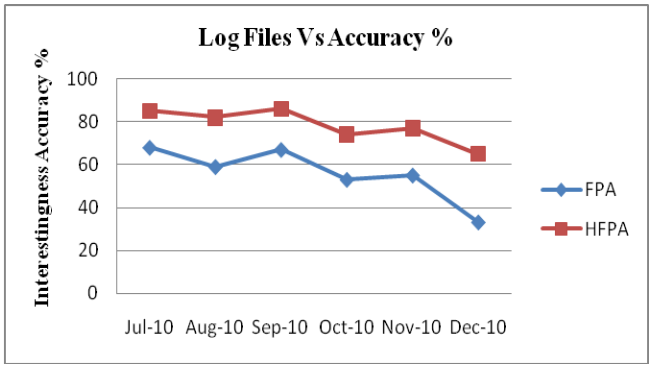

Fig 5: Chart Comparing Interestingness Accuracy % by the HFPA and FPA Approaches

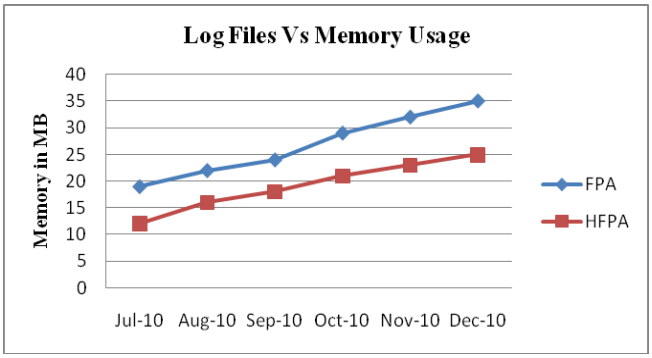

Fig 6: Chart Comparing Memory Usage by the HFPA and FPA Approaches

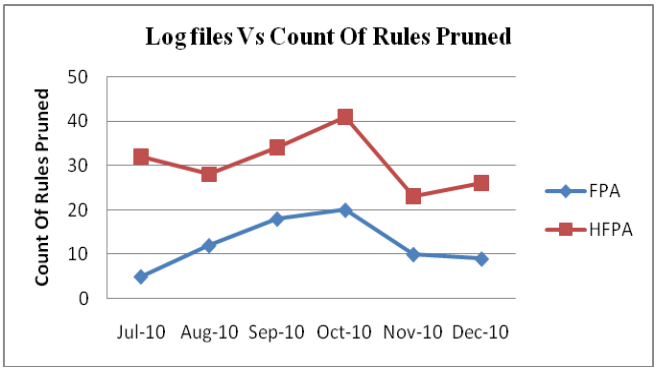

Fig 7: Chart Comparing Count of Rules Pruned by the HFPA and FPA Approaches

Thus HFPA approach takes lesser run time, produces less number of rules from which interestingness have to be evaluated, gives high percentage of interestingness accuracy, uses less memory for processing and prunes more rules than the traditional frequent pattern analysis, FPA approach. Thus this is a hierarchical frequent pattern mining approach that is found suitable for analyzing web log data and to predict useful information from the analyzed data.

The java software tool developed to do this performance evaluation also provides the opportunity for the users to set up different support and confidence threshold apart from what is suggested in this paper. By this way the approach can be tested for different web sites log files. Also the tool allows the user to select the top ranking percentage by which the bottom line is set for selecting interesting rules. By this way there is flexibility to get more interesting rules which might be found useful for the web master [13]. The rules can be categorized as "Expected to cause an action", "Might cause an action" and "Not expected to cause an action". The domain expert can then take the necessary action accordingly.

4. CONCLUSION

One of the major problems in the domain of web usage mining is that, the size of association rules produced increases dramatically due to the existence of rules that have very high confidence because of the interconnectedness of web pages through the link structure. In order to deal with this issue of rule over-generation, pruning of item sets is proposed at the initial stage itself that causes such uninteresting rules. The HFPA approach is also beneficial in that it reduces the number of records scanned in the session database while counting for the frequent item sets. This may be particularly useful for sparse data, where candidates do not occur in too many sessions. Even though there are many interesting measures available for association rule mining, the optimal minimum threshold for support and confidence is found and the ordering of lift and conviction values suit ranking of rules formed out of web log files. The count of interesting rules considered by the web site administrator to improve the browsing experience of the users is left to the individual performing the analysis, even though an optimal percentage of top ranked rules is proposed. The HFPA approach is compared with the traditional FPA approach and found that the new HFPA approach outperformed the existing FPA approach in many aspects like run time, memory usage, rules pruned, rules produced and accuracy percentage. There are scopes for future work in this proposal, like applying the same for different web sites to confirm the results and other interestingness measures can be explored to see if they give better results.

5. REFERENCES

[1] Kannan, S., & Bhaskaran, R. (2009) Association rule pruning based on interestingness measures with clustering. International Journal of Computer Science Issues, IJCSI, 6(1), 35-43.

[2] Liqiang Geng and Howard J. Hamilton, "Interestingness Measures for Data Mining: A Survey", ACM Computing Surveys, Vol. 38, No. 3, Article 9, September 2006.

[3] P. Tan, V. Kumar, and J. Srivastava. "Selecting the Right Interestingness Measure for Association Patterns". Technical Report 2002-112, Army High Performance Computing Research Center, 2002.

[4] Liaquat Majeed Sheikh, Basit Tanveer, Syed Mustafa Ali Hamdani. "Interesting Measures for Mining Association Rules", In Proceedings of INMIC 2004. 8th international Multitopic Conference, 2004, pp 641-644.

[5] Tianyi Wu, Yuguo Chen, and Jiawei Han, "Association Mining in Large Databases: A Re-Examination of Its Measures", In Proceedings of PKDD-2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Warsaw, Poland, September 17-21, 2007, pp 621-628.

[6] R. Iváncsy and I. Vajk, "Time- and Memory-Efficient Frequent Itemset Discovering Algorithm for Association Rule Mining." International Journal of Computer Applications in Techology, Special Issue on Data Mining Applications

[7] Huang, X. (2007). Comparison of interestingness measures for web usage mining: An empirical study. International Journal of Information Technology & Decision Making (IJITDM), 6(1), 15-41.

[8] Iváncsy, R., & Vajk, I. (2008). Frequent pattern mining in web log data. Journal of Applied Sciences at Budapest Tech, 3(1), Special Issue on Computational intelligence.

[9] Jaroszewicz , S., & Simovici, D. A. (2002). Pruning redundant association rules using maximum entropy principle. Advances in Knowledge Discovery and Data Mining, 6th Pacific-Asia Conference,PAKDD'02.

[10] H. Han and R. Elmasri, "Learning rules for conceptual structure on the web," J. Intell. Inf. Syst., Vol. 22, No. 3, pp. 237-256, 2004

[11] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization," ACM Trans. Inter. Tech., Vol. 3, No. 1, pp. 1-27, 2003

[12] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, "Mining access patterns efficiently from web logs," in PADKK '00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications. London, UK: Springer-Verlag, 2000, pp.396-407

[13] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," SIGKDD Explorations, Vol. 1, No. 2, pp. 12-23, 2000

[14] J. Punin, M. Krishnamoorthy, and M. Zaki, "Web usage mining: Languages and algorithms," in Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, 2001

[15] P. Batista, M. ario, and J. Silva, "Mining web access logs of an on-line newspaper," 2002

[16] Sudhamathy, G. 2010. Mining web logs: an automated approach. In Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in india (Coimbatore, India, September 16 - 17, 2010). A2CWiC '10. ACM, New York, NY, 1-4. DOI= http://doi.acm.org/10.1145/1858378.1858435

[17] J. Hou and Y. Zhang, "Effectively finding relevant web pages from linkage information." IEEE Trans. Knowl. Data Eng., Vol. 15, No. 4, pp. 940-951,2003