Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey on Different Techniques for Data Classification and Information Extraction from the Websites

Sonali Sonksusare[*]
M.E. Student
A.I., RGPV, India

Mr. Jayesh Surana
Assistant Professor
I.T., RGPV, India

Abstract— This paper is an attempt to make a survey on different techniques used for web information extraction, in order to remove the flaws of the old approaches in our new approach of information extraction and data classification. Due to rapid development of World Wide Web in its volume of traffic and the size the complexity of web sites increases day by day. So every web service provider wants to decrease this traffic load. For this, the classification of web data has been done. The data classification on the web sites and information extraction from the websites can be done by using different methods. The web mining is one of them. Aim of web mining is to retrieve useful and interesting patterns and then discover it from a large dataset. In Web mining knowledge extracted from web data, and at least one of structure or usage data is used in the mining process. In this paper we are making a survey on the different techniques used by different authors in their papers for web data classification and web information extraction. In this paper we are also presenting a conclusion of the entire study of these papers and on the basis of that we presenting an overview of our proposed approach in which the information extraction from websites is done by an advanced and easy way.

Keywords: Candidate sets, E-Web Miner Algorithm, Association rule, Apriori and AprioriAll Algorithm, Baum-Welch algorithm, Hidden markov model, Pre-Order Linked WAP-Tree (PLWAP-tree).

I. INTRODUCTION

Today every web service provider wants to discover the way to decrease the traffic load. This can be done by predicting the users' behaviors and personalize information from web logs. The organizations collects very large volumes of data in their daily operations, generated automatically by Web servers collect these data in Web access log files. In Web usage mining we mines this data for web information extraction. There are many kinds of data that can be used in web mining and can be classified into following five types:

- Content of Web Page.
- Inter-Page Structure of Web Page.
- Intra-Page Structure of Web Page
- Usage Data.
- User Profile.

Types Of Web Mining :
Web mining is divided into three types:

A. Web Content Mining

Web content mning deals with the discovering important and the useful knowledge or information from web page contents according to user needs. Web consist unstructured informations like image, audio, text, and video hence pattern recognitions are used according to type of information. [3] A lot of techniques and tools like statistical, neural network approaches, rapid miner, web data extractor etc. may be use for this purpose.

B. Web Structure Mining

Web Structure Mining deals with modeling and discovering the link structure of the web and discovers important information hidden in them. Web information retrieval tools not only make use of text available on web pages but also ignoring the valuable information contained in web links hence main focus of it is on link information . It is also used for the purpose of ranking of a webpage on the web. Algorithms like Page Rank, Weighted Page Rank and HITS (Hyper-link Induced Topic Search) are available for page ranking over the web.

C. Web Usage Mining

Web Usage Mining understands the user behaviour in interacting with a particular web site. Web usage mining uses web logs to record user access patterns. Log files are created by web servers and filled with information about user requests on a particular Web site. Web usage mining is used to know the importance of a webpage over others hence important information may be extracted according to the importance of the webpage. According to importance of the webpage web page ranking and content quality may also be improved i.e. integration of web content mining and web usage mining may provide very important results.

There are so many approaches available for mining the information from websites, but all the techniques may have some flaws, like complexity and use of synthetic data. So we are proposing a proficient Web Mining Algorithm for Web Log data classification and information extraction on web by removing all these flaws in our approach, using Distributed association rules Mining for correct web page prediction. In our research we focus on distributed association rules mining method. The proposed system will use visiting frequency of a page, time spent on a page to assign the quantitative weights to each page for a user. The other systems give user recommendations related to a navigation session or the user profile stored in the system. But The instinct of our approach is that, time spent and visiting frequency of pages is biasing factors to illustrate the interest on a page.

II. RELATED WORK AND TECHNOLOGIES USED:

Ajith Abraham in the paper in at al[2] introduces a novel approach 'Intelligent-Miner' to optimize and to analyses the trends of a concurrent architecture of a fuzzy inference system and fuzzy clustering algorithm (to discover data clusters). To segregate similar user interests they presents a hybrid evolutionary FCM approach. Then they use the clustered data to analyze the trends using a fuzzy interference system. In this paper, focus of the author was to develop accurate trend prediction models to analyze the hourly and daily web traffic volume, for this they presents the structural framework of the proposed hybrid model and some theoretical concepts of the Takagi-Sugeno fuzzy inference system and optimization of the fuzzy clustering algorithm.

S.Veeramalai in at al[3] uses different algorithms like Hash tree , Apriori , and Fuzzy to analyse the pattern and then to give the solution for Crisp Boundary problem with higher optimized efficiency, they used enhanced Apriori algorithm while comparing to other algorithms. In this paper the author proposed an algorithm which is based on rule generation phases, the Hash tree Algorithm and steps of frequent item sets. In this paper the author focuses on content and link filtering to eliminate the duplicate items from the search results. They proposed an algorithm which is based on the Hash tree Algorithm in which databases are scanned multiple times. A Hash tree stores all candidate K- item sets and their counts. In this process the authors by using their modified algorithm, can overcomes the crisp boundary the problem and also improved the efficiency by that algorithm. For tracking the facts of the summary, it has knowledge based summarizer on synonyms and keywords and provides a link for back reference.

Jiyi Xiao, Lamei Zou and Chaunqi Li in at al[4] demonstrates a new training method based on GA and Baum-Welch algorithms to obtain an HMM model for extracting the web information with optimized number of states. The author's method not only finds the better numbers of state in the HMM topology and its model parameters but also able to overcome the shortcoming of that slow convergence speed of the HMM approach. In this paper, the authors build a hybrid-GA to improve the quality of their results and the runtime behavior and they do this by combining the GA with the Baum-Welch algorithm. To build the HMM model for web information extraction, they firstly decide how many number of states the model should contain and what transaction or links should be allowed. After selecting the model structure, they estimated the transition and emission parameters. Then they extracts the information using "target" states, and for this information extraction they used the Viterbi algorithm for finding the most likely state sequence.

The Sang TT Nguyen in at al[5] presents a new web usage mining process for finding sequential patterns in web usage data. To enhance mining performance they presents design based on integration of the dynamic clustering based Markov Model with Pre- Order Linked WAP -Tree Mining Algorithm. Markov Model uses for web personalization and it a powerful and probabilistic model to estimate the probability of visiting web pages. They compared the existing work to build novel mining process, they combined the tree algorithm and the Markov model. This process can resolve the drawbacks of the existing methods like complexity problem and also predict user's web navigation patterns more effectively by only using the interesting web access patterns. In this paper the authors proposed a new web usage mining process which is the combination of the dynamic clustering-based Markov model and PLWAP algorithm. It overcomes the drawbacks of Markov model and PLWAP algorithm by omitting uninteresting web pages and also inherits the advantages of the PLWAP-tree and the dynamic clustering-based Markov model. It can predict the most interesting web access patterns from the users' navigation history. It has been tested using the two web log datasets taken from real websites. The testing results show that the new mining process can considerably improve the drawbacks of Markov model and enhances the performance of PLWAP algorithm.

For web log analysis Mahendra Pratap Yadav in at al[1] presents an efficient web mining algorithm for web log analysis and applied the results obtained on this web log analysis to a class of problems for finding out the contexts of website design of a E- commerce web portal which demands security. In this paper the authors compared the algorithm with Improved AprioriAll Algorithm which is its other earlier incarnation . The proposed algorithm, Efficient Web Miner or E-Web Miner can be verified by computational comparative performance analysis and can be traced for its valid results and. This paper intends to show that the E-Web Miner has lower complexity of time and space than Improved AprioriAll Algorithm and confirms the correctness of result obtained by providing a trace back route for candidate set pruning for both the algorithms. E-Web Miner is the proposed web mining algorithm that removes the flaws of Improved AprioriAll algorithm and improve upon the time complexity of the earlier Aprioriall algorithm. It provides an improved candidate set pruning as well. In fact, it has been shown successfully that it mines correct result of candidate set where as the Improved AprioriAll algorithm fails to deliver the correct result. The algorithm has been designed independent of Aprioriall algorithm. In this paper authors work shows that when Improved AprioriAll Algorithm fail to deliver the desired result, the proposed algorithm of web mining in web log analysis presents a cost effective valid result having reduced candidate set pruning of correct order. So in this way the authors proved that E- Web miner is proved to be more time effective than other algorithm.

An algorithm called AIS was proposed for mining association rules in [6] and also the problem of discovering the association rules was very first introduced in this. Many algorithms for rule mining have been proposed for last fifteen years. Most of them follows the representative approach by Agrawal et al.[7], namely Apriori algorithm. For improving the performance and scalability of Apriori various researches were done included parallel computing.

The Apriori Algorithms or an association rules mining algorithm generally contains the following steps [8]:
By 1-extensions of the large (k 1)itemsets generated in the previous iteration, the set of candidate k itemsets is generated. And then by a pass over the database, the supports for the candidate k itemsts are generated. The Itemsets which don't have the minimum support are discarded and the remaining itemsets are called large k itemsets. Until no more large itemsets are found, this process is repeated. The Partition algorithm [9] requires just two database scans to mine large itemsetsand logically partitions the database D into n partitions. In Our Apporch we improved the performance of the conventional Apriori algorithm that mines association rules by presenting fast and scalable algorithm for discovering association rules in large databases. We propose a proficient Web Mining Algorithm for Web Log Analysis using Distributed association rules Mining. We also propose an effective Web log mining system that deals with log preprocessing, sequential pattern mining, and result visualizing.

III. RESEARCH METHODOLOGY:

For Information extraction so many approaches have been used and the latest approach used by Mahendra Pratap Yadav in at al[1] in which they use a technique in which candidate sets have to create for storing the data but this process of candidate set generation creates so much problems related to complexity and felxibilty. So to remove this flaw we are presenting an approach in which candidate set generation has been completely eliminated to make the process more flexible. In our scheme, the process works in parallel on a shared-nothing architecture. The processors begin with counting local support for each item. This counts are then exchanged across the group in order to each processor calculates their sum. After discarding globally infrequent items, each processor constructs the *F-list* structure sorted by descending order of frequent item supports figure 1. In the next phase, local *CFP-trees* are built by scanning local data partitions and considering only local items belonging to *F-list*.

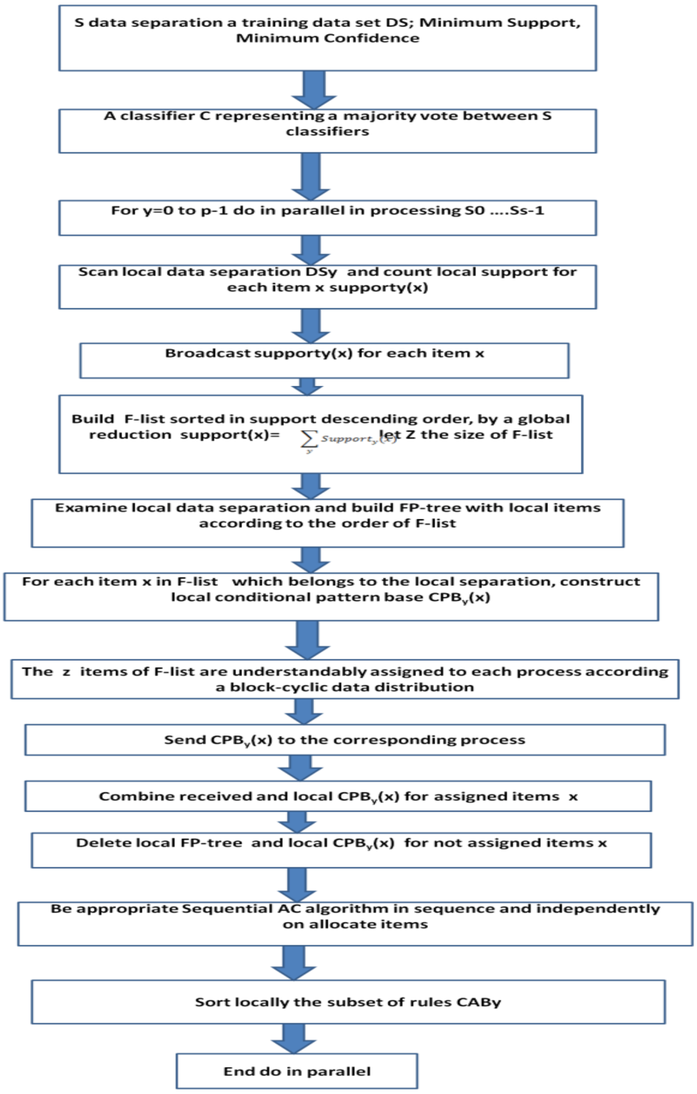
S data separation a training data set DS; Minimum Support, Minimum Confidence
A classifier C representing a majority vote between S classifiers
For y=0 to p-1 do in parallel in processing S0 ....Ss-1
Scan local data separation DSy and count local support for each item x supporty(x)
Broadcast supporty(x) for each item x
Build F-list sorted in support descending order, by a global reduction support(x)= $\sum_{y} Support_y(x)$ let Z the size of F-list
Examine local data separation and build FP-tree with local items according to the order of F-list
For each item x in F-list which belongs to the local separation, construct local conditional pattern base $CPB_y(x)$
The z items of F-list are understandably assigned to each process according a block-cyclic data distribution
Send $CPB_y(x)$ to the corresponding process
Combine received and local $CPB_y(x)$ for assigned items x
Delete local FP-tree and local $CPB_y(x)$ for not assigned items x
Be appropriate Sequential AC algorithm in sequence and independently on allocate items
Sort locally the subset of rules CABy
End do in parallel

Figure 1: DAR Algorithm

The Sang TT Nguyen in at al[5] and Mahendra Pratap Yadav in at al[1] uses only the raw data and Synthetic data in their experiments. But we are applying the propose algorithm to more extensive empirical evaluation and to confirm the experimental results in the real life domain we also use real data like medical transactions and retail sales transaction in our approach. Till now the authors used the association rules only in Local environment but in our approach, we are using rules in distributed environment also. For dealing with the problem of discovering hidden information from large amount of Web log data collected by web servers the distributed association rules mining is performed from transaction databases (these rules involve items at different levels of abstraction). Our contribution is to introduce the web log mining process, and to illustrate how frequent pattern discovery tasks can apply on the web log data for obtaining the useful information about the navigation behavior of users.

IV. OVERALL COMPARISON AND CONCLUSION:

On Surveying the Above Papers, we conclude that, there are so many approaches presented for web information extraction by using different algorithms like, APRIOR Algorithm, APRIORI ALL Algorithm, HMM training method, i-miner, E- Web Miner etc. Among all these the latest technology used is E-Web miner. But in the E-web miner[1], although it has reduced the problem of candidate set generation by providing an improved candidate set pruning but still it cannot removed this problem completely. So, to eliminate this problem and to remove this problem completely for reducing the complexity we presents an advanced approach for web log data classification and information extraction, in which *the generation of the candidate sets is completely eliminated* with the help of a CPB trees, so that the time of the processing will be saved and complexity errors can also be eliminated.

Also in this area, till now the work is being done only on the *Synthetic data*. But in the experiment, we are using the Real data set like retail sales transaction and medical transactions to confirm the experimental results in the real life domain. Also till now the association rules are used only in *Local environment* for information extraction. But we use the association rules in the Distributed environment. Distributed association rules algorithm discovers small frequent item sets in a very quick way, this is the main advantage of the distributed association, thus the task of discovering the longer ones is enhanced as well by it.

REFERENCES

[1] Mahendra Pratap Yadav, Pankaj Kumar Keserwani, Shefalika Ghosh Samaddar," An Efficient Web Mining Algorithm for Web Log Analysis: E-Web Miner" 1st Int'l Conf. on Recent Advances in Information Technology | RAIT-2012 |.

[2] Ajith Abraham and Xiaozhe Wang," i-Miner: A Web Usage Mining Framework Using Neuro-Genetic-Fuzzy Approach"

[3] S.Veeramalai , N.Jaisankar and A.Kannan ," Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy" International journal of computer science & 2 Algorithm for Web Information Extraction" ISKE-2007 Proceedings.

[4] Jiyi Xiao Lamei Zou Chuanqi Li," Optimization of Hidden Markov Model by a Genetic Algorithm for Web Information Extraction" ISKE-2007 Proceedings.

[5] Sang T.T. Nguyen," Efficient Web Usage Mining Process for Sequential Patterns" iiWAS2009, December 14–16, 2009, Kuala Lumpur, Malaysia.

[6] R. Agrawal, T. Imielinski, and A. Swami. "Mining association rules between sets of items in large databases". In Proceedings of the ACM SIGMOD International Conference on Management of Data (ACM SIGMOD '93), pages 207216, Washington, USA, May 1993.

[7] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Prof. 20th Int'l Conf. Very Large Data Bases, pp. 478499, 1994.

[8] K. Sotiris, and D. Kanellopoulos, "Association Rules Mining: A Recent Overview. GESTS International Transactions on Computer Science and Engineering", Vol.32 (1), 2006, pp. 7182.

[9] A. Savasere, E. Omiecinski, and S. Navathe. "An Efficient Algorithm for Mining Association Rules in Large Databases". Proceedings of 21th International Conference on Very Large Data Bases (VLDB'95),

[10] N. Chawla, S. Eschrich, L.O. Hall, "Creating Ensembles of Classifiers," IEEE International Conference on Data Mining, pp. 580-581, 2001.