Weighted Association Rule Mining without Pre Assigned Weights for Web Log Analysis

A.Gomathi¹, V. Priyadharshini², S. Shanthi³

¹M.E student, Dept of CSE, Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu, India.
²M.E student, Dept of CSE, Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu, India.
³Asst. Professor, Dept of CSE, Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu, India.

Abstract-The Aim of the Association rule mining is to analyze a large transaction database in order to find the relationships among the data attributes or items. Moreover the Association rule Mining is one of the thriving research topic in Data Mining. The traditional model of Association rule Mining aids with the support measure. Also the measure considers every transaction alike. Conversely, each transaction has variable weights in real time. Pre assigned weight plays a vital role in Association rule mining. The proposed system introduces an innovative measure w-support, which does not in requirement of pre assigned weights. It focuses the quality of transactions into consideration using link-based models. A fast mining algorithm is designed, and a large amount of experiments are conducted. The system can also enhance for web log analysis.

Keywords- HITS, W-support, W-confidence.

1. INTRODUCTION

The link based weighted rule mining system for web user logs is designed to handle the association rule mining process for the web user logs. The system does not require any pre assigned weights. The weight based rule mining uses the W-Support and W-Confidence. The links in the transaction are used for the weight extraction process. The system is divided into four major modules. They are log preprocess, ranking process, weight estimation and rule mining process. The log preprocess module is designed to perform web user log cleaning process. The web user logs are collected as data files. The data values are converted into database tables. The data values are analyzed into database tables. The preprocess also cleans the web access log by eliminate the redundant and irrelevant records. The transactions are updated with all item sets into the database. The ranking process uses the transaction and item set values with its link nature. The HITS algorithm is used to rank the transactions. The transaction and item sets are represented into a table. The relationship between the transaction and item sets are analyzed. The transactions are identified as hub and the item set values are referred as authority values. Each transaction is ranked with reference to its link values. The system uses two types of ranking process. They are ranking based on architecture and rank based on transactions. The architecture based ranking is done statically. The transaction based ranking is differing from transaction to transaction. The combined ranking scheme is used in this system. The transaction and its rank values are updated into the database. The weight estimation process is done after the ranking process. The transactions are fetched with its ranks and its links values are analyzed. The weight values assigned with reference to the rank values that assigned by the HITS algorithm. The rule mining process is applied after the weight estimation process. The weight based support and weight based confidence values are estimated. The minimum support and minimum confidence values are used to find out the best rules from the transactions.

2. WEB ACCESS LOGS

The proposed system is designed to perform weighted rule mining without pre assigned weights for web access logs. W-support is a new measure of item sets in databases with only binary attributes. The basic idea behind w-support is that a frequent item set may not be as important as it appears, because the weights of transactions are different. These weights are completely derived from the internal structure of the database based on the assumption that good transactions consist of good items. In weighted rule mining items are assigned with weights. Support is distinct from weighted support and confidence is also different from weighted confidence. The system is designed by extending Hyperlink Induced Topic Search (HITS) model to bipartite graphs. The weight estimation is done with page links and access log links. The system is designed with link-based model for weight estimation process. The integrated weight estimation model combines the
weight estimation using the site architecture and access logs. The site architecture based weight estimation uses the site internal links. The access log based weight estimation process uses the links derived from access logs. The integrated weight values are used for the mining process.

3. SYSTEM DESIGN

The link based weighted rule mining system for web user logs is designed to handle the association rule mining process for the web user logs. The system does not require any pre assigned weights. The weight based rule mining uses the W-Support and W-Confidence. The links in the transaction are used for the weight extraction process. The system is divided into four major modules. They are log preprocess, ranking process, weight estimation and rule mining process. The log preprocess module is designed to perform web user log cleaning process.

The web user logs are collected as data files. The data values are converted into database tables. The preprocess also cleans the web access log by eliminate the redundant and irrelevant records. The transactions are updated with all item sets into the database. The ranking process uses the transaction and item set values with its link nature. The HITS algorithm is used to rank the transactions. The transaction and item sets are represented into a table. The relationship between the transaction and item sets are analyzed. The transactions are identified as hub and the item set values are referred as authority values. Each transaction is ranked with reference to its link values. The system uses two types of ranking process. They are ranking based on site architecture and rank based on transactions. The site architecture based ranking is done statically. The transaction based ranking is differing from transaction to transaction. The combined ranking scheme is used in this system.

The transaction and its rank values are updated into the database. The weight estimation process is done after the ranking process. The transactions are fetched with its ranks and its links values are analyzed. The weight values assigned with reference to the rank values that assigned by the HITS algorithm. The rule mining process is applied after the weight estimation process. The weight based support and weight based confidence values are estimated. The minimum support and minimum confidence values are used to find out the best rules from the transactions.

4. INPUT DESIGN

The weighted rule mining system for web logs is designed as a standalone graphical user interface based application. The system uses the access logs and site information as the major input. The access log server collects from web servers. The access logs are created by the web sites. The system uses the log input data and user input data values. The page details are collected from the users. The page links are also collected from the users. The input forms are designed with Java swing. The java.io package is also used for the input process.

Fig.1 Data flow Diagram

The system uses five input forms. They are page entry, link entry, log entry, clean process and session conversion. The page entry form is designed to collect the page details in the web site. The page URL, page type and hosted time details are collected from the form. The link information form is designed to show the in link and out link details for the selected page. The log entry form is designed to access the access log details from the user. The user can enter the access log details into the system. Session id, page URL, requested time and IP address details are fetched from the log entry. The clean process is used to clean the logs from noisy records. The session conversion is designed to convert the log entries into the session list.

5. OUTPUT DESIGN

The weighted rule mining on web usage logs is designed with a set of intermediate and final output forms. The system produces summary and tabular results. The system is designed with 14 output forms. The page list form shows the list of hosted pages in the site. The pages are listed with in link and out link count. The link information form is designed to show the in link and out link details for the selected page. The log...
list form shows the access log details. The log list shows the session id, page URL, requested time and requested IP address. The page access sequence form shows the pages that accessed in selected session id. The session details form shows the pages and session details.

The system estimates the page weights in three ways. They are site architecture based weight; access log based weights and integrated weights. The site architecture based weights are listed with page URL and weight details. The access log based weights are listed with the page URL and weights. The integrated weights are produced with access log and site architecture weights. The rule mining details are produced in two ways. They are general mining results and weighted mining results. The general mining results are produced with respect to the frequency values. The general rules are produced in two forms. They rule list and interest rule selection. The interest rule selection is produced with minimum support and minimum confidence values. In the same way the weighted rules are also produced with weighted support and weighted confidence values.

6. DATABASE DESIGN

The database design describes the list of data of database tables used in the system. The system is designed with Oracle back end. The page and log details are maintained in the database tables. The system uses a set of tables and views for page and link details. The system is designed with 7 tables. They are page in for, user logs, out links, session list, attricount1, attricount2 and in weights. The page info table is used to maintain the page details. The out links table is used to store the link details. The session list table is used to store the session details. The attricount1 and attricount2 tables are used to maintain the frequency values. The in weights table is used to store integrated weight values.

7. IMPLEMENTATION

(a) Access Log Analysis

The access log analysis module is designed to tune up the user access log for rule mining process. Page access entries are grouped into session access details. The session access details maintain single entry for each session with multiple page access data values. The access log analysis module is designed to manage the web site architecture and access details. The web site architecture is prepared using the page information and link information. The page entry sub module is designed to register the page details in a web site. The site architecture is represented with internal page links and hierarchy level. The access log entry sub module is designed to enter the page access details. The session id, page URL, access time, IP address, inlink and outlink details are updated into the access logs. The access log list shows the page access information. The same session id values can be used for multiple page access with reference to the time values. The session conversion sub module is designed to convert the multiple page access entry into single session entry. The session id is used for the session conversion process. The access log is reproduced as access session information. The session list is used to display the session information with page access details.

(b) Weight Estimation Process

The weight estimation is performed to automatically evaluate the importance of the pages. The weight estimation process is done with two methods. The web site architecture and access log links are used for the weight estimation. Both weights are integrated to assign actual weight for the data sets. The weight estimation is done in three levels. They are site architecture based weight; access link based weight and integrated weight estimation. The site architecture based weight estimation is done using the internal page links for a web site. The page links and its hierarchy are used for the site architecture based weight estimation process. The access log details are used to extract access link hierarchy information. The access log based weight is estimated using the access link hierarchy values. The access link based weight and site architecture based weight are used to estimate integrated weight estimation process. The weight details are displayed in separate forms.

(c) Rule Mining Process

The rule mining process is carried out to detect association rules for the web logs. The weighted support and weighted confidence values are calculated using automated weights. The minimum support and minimum confidence values are used to extract strong rules. The weighted rule mining process is compared with general rule mining process. The rule mining process is done with two levels general rule mining and weighted rule mining. The general rule mining process is done with the frequency values. The support and confidence values are used in the general rule mining process. The interested rules are extracted from the rule mining process. The weighted rule mining is done in three levels. The site architecture based mining is applied with web site link weights. The access link based rule mining is applied with access link weight values. The integrated weight is used for integrated rule mining process. The weighted support and weighted confidence values are used in the weighted rule mining process. The performance analysis done with in weighted rule mining and general rule mining techniques.
8. FAST MINING ALGORITHM

1) Initialize auth (i) to for each item i
2) For (l =0; l<num _ it; l++) do begin
3) Auth’ (i) = 0 for each item i
4) For all transactions t D do begin
5) Hub (t) = I: it auth (i)
6) Auth’ (i) += hub (t) for each item i T
7) End
8) Auth (i) = auth’ (i) for each item I, normalizeauth
9) End
10) L1 = {(i): wsupp (i) >= minwsupp}
11) For (k =2; Lk;i; K++) do begin
12) Ck = apriori-gen (Lk-1)
13) For all transactions t D do begin
14) Ct = Subset (Ck, t)
15) For all candidates c Ct do
16) C.wsupp += hub (t)
17) H += hub (t)
18) end
19) Lk = {c Ck | c.wsupp/H >= minwsupp}
20) end

9. CONCLUSION

A novel framework is presented in association rule mining. First, the HITS model and algorithm are used to derive the weights of transactions from a database with only binary attributes. Based on these weights, a new measure w-support is defined to give the significance of item sets. It differs from the traditional support in taking the quality of transactions into consideration. Then, the w-support and w-confidence of association rules are defined in analogy to the definition of support and confidence. An Apriori-like algorithm is proposed to extract association rules whose w-support and w-confidence are above some given thresholds.

Experimental results show that the computational cost of the link-based model is reasonable. At the expense of three or four additional database scans, the system can acquire results different from those obtained by traditional counting-based models. Particularly for sparse data sets, some significant item sets that are not so frequent can be found in the link based model. Through comparison this model and method address emphasis on high-quality transactions.

The link-based model is useful in adjusting the mining results given by the traditional techniques. Some interesting patterns may be discovered when the hub weights of transactions are taken into account. Moreover, the transaction ranking approach is precious for estimating customer potential when only binary attributes are available, such as in Web log analysis or recommendation systems.

REFERENCES

• Chris Cornelis and Peng Yan, Xing Zhang, Guoqing Chen (2006) “Mining Positive and Negative Association Rules from Large Databases” IEEE International Forum on Information and Documentation
• Joshua Zhexue Huang, Michael K. Ng, HongqiangRong, and Zichen Li (2005)“Automated Variable Weighting in k-Means Type Clustering” IEEE In Proceedings of SIGIR, pages 504-511.


BIOGRAPHY

A. GOMATHI received her MCA Degree from Anna University, Chennai, Tamil Nadu, India and pursuing M.E Degree from Sri Eshwar College of Engineering, Coimbatore, India. Her field of Interest is Data Mining.

S. SHANTHI Working as an assistant professor at Sri Eshwar engineering college, her work experience is 8 years and 2 months. Her Area of interest is Data Mining.

V. PRIYA DHARSHINI received her B.E Degree from Karpagam University, Coimbatore, Tamil Nadu, India and pursuing M.E Degree from Sri Eshwar College of Engineering, Coimbatore, India. Her field of Interest is Network Security.