Data Preprocessing in frequent item-sets of webpage traversals

[#1] R.Pradeepa, [*2] A.Kumaravel,
#1 PG Student, Department of Computer Science and Engineering.
Bharath University, Selaiyur, Chennai-600073, India
pradimca@gmail.com
*2 Dean & Professor,Department of Computer Science and Engineering.
Bharath University, Selaiyur, Chennai-600073, India
drkumaravel@gmail.com

Abstract

Data Preprocessing plays a main role in Data mining projects. Preprocessing is the ground work of data mining. Before performing the mining process the raw data has to be preprocessed in order to improve the quality of data to be mined. This mining involves the automatic filtering of patterns from various web pages of a Web site. This helps the web developers to determine the value of specific web pages, this paper discusses the importance of data preprocessing methods and various steps involved in getting the required web page effectively. A complete preprocessing technique is being proposed to preprocess the web site for extraction of web page hits. Data cleaning algorithm removes the irrelevant entries and filtering algorithm discards the uninterested attributes from database. At first, scans the sequence database once, finds all the weighted frequent items and make a sparse matrix to identify significant item set and satisfying minimum support and the minimum weighted support count. The experimental results show that this method is more efficient in mining sequential pattern and satisfying the requirement of users.

Keywords: Web mining, Sparse Matrix, Preprocessing, Search methods.

1. INTRODUCTION

The World Wide Web provides a platform for exchanging various kinds of facts. The information available on the internet is rapidly increasing with the volatile growth. A common classification of web mining defines three main researches: content, structure and usage mining. Web Usage mining is the application of data mining techniques to discover usage patterns from web data. Data is usually collected from user's interaction with the web, e.g. various page hits of a web site by different users. Usage mining tools discover and predict user behavior, in order to help the designer to improve the web site, to attract visitors, or to give regular users a personalized and adaptive service [1].

In this paper, we describe a solution to Web developers to discover a web site with best web pages to attract visitors. We demonstrate how web mining technology can be effectively applied. The structure we propose takes the outcome of the web mining process as input, and converts these results for future development.

2. PROPOSED METHOD

The main procedure of using Web usage mining for Web recommendation consists of three steps, i.e. data collection and pre-processing, pattern mining as well as knowledge application [3].

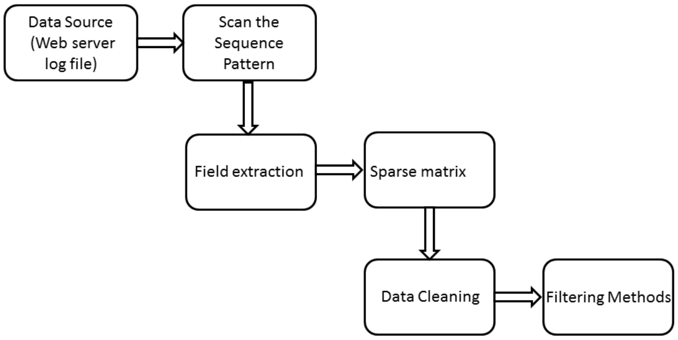

Fig.1 Flow of Preprocessing

The Data source consists of a sampling and processing the www.microsoft.com logs. Using the field extractor the attributes are separated using the delimiters and converted into sparse matrix. Then the data cleansing process is applied for filtering the unwanted and irrelevant data entry to increase the quality of data. The cleaned data are then used for future process.

This paper shows the various search method for data preprocessing to solve some of the existing problems in traditional web based method. The Distinct attribute Identification strategy of identifying the frequent item sets from the referred web page is differing from that of the traditional web based method.

3. EXPERIMENTAL ANALYSIS AND RESULTS

In this section, we test the implementation efficiency of various search methods and compare with whole sequence and the selected sequence. Weka tool is used to select the attributes from the sparse matrix. The datasets for these experiments are from www.microsoft.com logs.

3.1 Dataset

3.1.1 Dataset Information

The data is in an ASCII-based sparse-data format called "DST". Each line of the data file starts with a letter which tells the line's type. The three line types of interest are:

-- Attribute lines
-- Case Lines and
-- Vote Lines

For example,

'A,1277,1,"NetShow for PowerPoint","/stream"'
C,"10164",10164
V,1123,1
V,1009,1
V,1052,1

Where:

'A' marks this as an attribute line,
"'/stream"' is the URL relative to web site
'C' marks this as a case line,
'10164' is the case ID number of a user,
'V' marks the vote lines for this case,
'1123', 1009', 1052' are the attributes ID's of Vroots that a user visited.

Training Instances	32711
Testing Instances	5000
Attributes	294
Mean vroot visits per case	3.0

3.1.2 Attribute Information:

Each category is associated--in order--with an URL id starting with "1000". For example, "/regwiz" is associated with 1000, "/support" with 1001, and " /athome" with 1002 each row below "% Sequences:" describes the hits--in order of single user. For example, the first user hits /regwiz" twice, and the second user hits "/support" once.

3.1.3 Data Preprocessing:

Web page personalization to perform web usage mining preprocessing is necessary, because Log file encloses noisy and unclear data which may affect results of the mining process. Some of the web log file data are unnecessary for analytical process and could affect the performance.

Data preprocessing is an important step to filter and organize the appropriate information before applying any web mining algorithm. Preprocessing increases the quality of available data by reducing the log file size. The primary use of data preprocessing is to improve data quality and increase mining accuracy. The Preprocessing consist of following steps

- ➢ Field extraction
- ➢ Data Cleansing
- ➢ Identifying attributes

3.1.3.1Field Extraction:

Extract the log file into sequence format to find the equivalent sparse matrix. Table 1 Shows the sample extracted sequence from web log file. A sequence *s* is set of visitors of a webpages, denoted as <*s1*, *s2*,…, *sl*>, where *sj* (*1≤j≤l*) is an itemset. Each itemset of sequence is also called an element of sequence, denoted as (*x1*, *x2*,…, *xm*), where *xk* (*1≤k≤m*) is an item. For brevity, the brackets are omitted if an element has only one item. A sequence database *SDB* is a set of sequence. The total number of items in a sequence is called the length of the sequence and a sequence with length *l* is called an *l*-sequence.

TABLE 1
SAMPLE EXTRACTED SEQUENCE

SID	Sequence
V1	1038 1026 1034
V2	1008 1056 1032
V3	1064 1065 1020 1007 1038 1026 1052 1041 1028
V4	1004
V5	1017 1156 1004 1018 1008 1027 1009 1046 1038 1006
V6	1052 1060 1001 1041 1034
V7	1034 1004
V8	1008 1000 1035 1016 1031 1003 1034 1018
V9	1065 1123 1009 1007
V10	1017 1043

The above extracted sequence is converted into sparse matrix [3] for a data cleaning process. A useful application of linear list is the representation of matrices that contain a preponderance of zero elements. These matrices are called sparse matrices.

3.1.3.2 Data Cleansing:

This stage consists of removing the entire data track in Web logs that are ineffective for mining purposes. Graphic file requests, agent/spider crawling etc. could be simply removed by only seeming for an HTML file requests. Normalization of URL's is regularly required to make the requests consistent.

3.1.3.2 Identifying attributes:

Our analysis was to reduce the high data dimensionality. For this purpose we used Weka tool for attribute selection based on various search methods made in the attribute space as shown in table 2.We used factors which are selected after preprocessing as new predictors.

Table 2: Identified Attributes

S.No	Selected Page	Page id	Attribute No
1	regwiz	1000	2
2	ie	1034	36
3	hardwaresupport	1205	206
4	centroam	1297	295

Methods used for identifying attributes are,

Best First Search:

This method [6] searches the attribute subset space by best first search. To specify a starting set of attributes by -P start set. Eg 1,4,7-9. The direction of search is defined by the options -D 0 = backward | 1 = forward | 2 = bidirectional default is 1.Number of non-improving nodes to consider before terminating search is denoted by –N default is 5.Size of lookup cache for evaluated subsets. Expressed as a multiple of the number of attributes in the data set.

Greedy Stepwise Algorithm:

This method [7] performs a greedy forward or backward search through the space of attribute subsets. It may start with no attribute or all attributes or from an arbitrary point in the space.Stops when the addition or deletion of any remaining attributes results in a decrease in evaluation. It can also produce a ranked list of attributes by traversing the space from one side to the other and recording the order that attributes are selected.

Ranker:

This method [8] performs a ranks attributes by their individual evaluations. Use in conjunction with attribute evaluators (ReliefF, GainRatio, Entropy etc). Method used in the ranker search is getNumToSelect() to gets the number of attributes to be retained. Kind of a dummy search algorithm. Calls an Attribute evaluator to evaluate each attribute not included in the start Set and then sorts them to produce a ranked list of attributes.

4. CONCLUSIONS

An important task in this preprocessing application is the creation of a suitable target data set to which data mining and statistical algorithms can be applied. This is particularly important in Web mining due to the characteristics of click stream data. The data preparation process is often the most time consuming and computationally intensive step in the Web usage mining process, and often requires the use of search algorithms and heuristics, not commonly in other domains.

Table 3: Summary of preprocessed data

Records in original training log file	32711
Records in original testing log file	5000
Records in cleaned log file	4999
Attributes in original log file	294
Identified attributes	4

ACKNOWLEDGEMENTS

The authors would like to thank the management of Bharath University for the support and encouragement for this research work.

REFERENCES:

1 Magdalini Eirinaki and Michalis Vazirgiannis , "Web Mining for Web Personalization" Communications of the ACM, vol. 3, No. 1, pp.2-21, Feb. 2003.

2 http://www.mec.ac.in/resources/notes/notes/ds/sparse.htm

3 http://research.ijcaonline.org/volume58/number3/pxc3883438.pdf

4 Mining web frequent multi-dimensional sequential patterns by Guoyan Huang ,Na zuo and Jiadong Ren. Information Technology journal 10(12):2434-2439,2011.

5 Yong Chen, Rongfang Bie, Chuan Xu, “A New Approach for Maximal Frequent Sequential
Patterns Mining Over Data Streams”, International Journal of Digital Content Technology and its
Applications, Vol.5, No.6, pp.104-112,2011.

6 http://weka.sourceforge.net/doc/weka/attributeSelection/BestFirst.html

7 http://weka.sourceforge.net/doc.dev/weka/attributeSelection/GreedyStepwise.html

8 http://weka.sourceforge.net/doc.dev/weka/attributeSelection/Ranker.html

9 Data Preprocessing in Web Usage Mining Vijayashri Losarwar, Dr. Madhuri Joshi International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore

10 J. Vellingiri and S. Chenthur Pandian, “A Novel Technique for Web Log Mining with Better Data Cleaning and Transaction Identification”,Journal of Computer Science, pp. 683-689, 2011.