ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 7, September 2012

Enhanced Approach on Web Page Classification Using Machine Learning Technique

S.Gowri Shanthi
Research Scholar, Department of Computer Science,
NGM College, Pollachi, India.

Dr. Antony Selvadoss Thanamani,
Head & Associate Professor , Department of Computer Science,
NGM College, Pollachi, India.

Abstract **- The data set contains WWW-pages collected from computer science departments of various universities in January 1997 by the World Wide Knowledge Base project of the CMU text learning group. The 8,282 pages were manually classified into 7 classes: 1) student, 2) faculty, 3) staff, 4) department, 5) course, 6) project and 7) other. For each class the data set contains pages from the four universities: Cornell, Texas, Washington, Wisconsin and 4,120 miscellaneous pages from other universities. The files are organized into a directory structure, one directory for each class. Each of these seven directories contains 5 subdirectories, one for each of the 4 universities and one for the miscellaneous pages. These directories in turn contain the Web-pages. The proposed work performs the data preprocessing to clean the dataset and transform it in to the pattern for classification. Then the feature extraction is performed for extracting only minimum number of representative features or terms extracted from it without using the entire Web page. After that the classification algorithm is used to classify the dataset into one of the seven classed using FP-Growth algorithm. The proposed approach is compared with the existing system apriori algorithm.**

Keywords- **Apriori algorithm, Classification, Data Preprocessing, Data Set, FP – Growth.**

I. INTRODUCTION

A web search engine is designed to search information on the World Wide Web. Web search engine also mine data available in databases. The existing search engines, Google, Bing, and Yahoo. Interacting with the web for the following purposes: Finding Relevant Information, Discovering New Knowledge, Personalized Web Page Synthesis, Learning about Individual Users.

With more than two billion pages created by millions of Web page authors and organizations, the World Wide Web (WWW) is a tremendously rich knowledge base. The knowledge comes not only from the content of the pages themselves, but also from the unique characteristics of the Web, such as its hyperlink structure and its diversity of content and languages. Analysis of these characteristics often reveals interesting patterns and new knowledge. Such knowledge can be used to improve users' efficiency and effectiveness in searching for information on the Web, and also for applications unrelated to the Web, such as support for decision making or business management. The Web's size and its unstructured and dynamic content, as well as its multilingual nature, make the extraction of useful knowledge a challenging research problem. Furthermore, the Web generates a large amount of data in other formats that contain valuable information. For example, Web server logs' information about user access patterns can be used for information personalization or improving Web page design. Machine learning techniques represent one possible approach to addressing the problem. Artificial intelligence and machine learning techniques have been applied in many important applications in both scientific and business domains, and data mining research has become a significant subfield in this area.

The Apriori algorithm has some drawbacks to classify the web pages. In this paper, FP-Growth algorithm is used to categories the different web page.

This paper is organized as follows: Web Mining is described in Section II, the existing works are described in section III, the proposed works are described in section IV, the methodology and implementation are described in section V, and in section VI present conclusion of the paper.

II. WEB MINING

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents [1].Mining techniques in the web can be categorized into three areas of interest.

A. Web Content Mining

Web content mining describes the discovery of useful information from the web contents. The web contents could include a very broad range of data. Web Content Mining is related to Data Mining because many Data Mining techniques consist can be applied in Web Content Mining. The web content of several types of data such as textual, image, audio, video, metadata, as well as hyperlinks. The textual part of web content data consists of unstructured data and semi-structured data.

B. Web Structure Mining

Web structure mining is concerned with discovering the model underlying the link structures of the web. Web structure mining used to categorize web pages and is useful to generate information such as the similarity and relationship between web sites. Web structure mining describes the study of topology of the hyperlinks. Page Rank and Hyperlink analysis also falls in this category.

C. Web Usage Mining

Web usage mining deals with studying the data generated by the web surfer's sessions of behaviors. Web usage mining mines the secondary data on the web. This activity involves the automatic discovery of user access patterns from one or more web servers. Through this mining technique we can find out what users are looking for on Internet. Some might be looking for only technical data, where as some others might be interested in multimedia data.

III. EXISTING METHOD

The research work was initiated through a study and analysis phase, where significant study was conducted to understand the existing system. Using Apriori algorithm for web log mining is a novel technique [2].

The explosive growth of the World Wide Web (WWW) in recent years has turned the web into the largest source of available online data.

- Situations like several unrelated topics in a single web page may lead to satisfy the visitors are too rigid to reach the information.
- Understand the way user browses the site and find out which is the most frequent used link and pattern of using the features available in the site.

All these information are available on the online, but hidden for the users. Presently, Apriori based approach extract this hidden information for analyzing the visitor browsing behavior.

IV. PROPOSED METHOD

The exponential increase in the amount of Web data, automatic classification of Web pages into predefined categories is essential to support the Web directories. Web pages can be classified by two methods: syntactic and semantic. This proposed work emphasizes syntactic classification, which uses a set of words or patterns in a web page to classify it.

This article suggests a method to classify a Web page with only minimum number of representative features or terms extracted from it without using the entire Web page. The optimum number of features is selected using a three step procedure, by filtering the features in each subsequent step.

V .METHODOLOGYAND IMPLEMENTATION

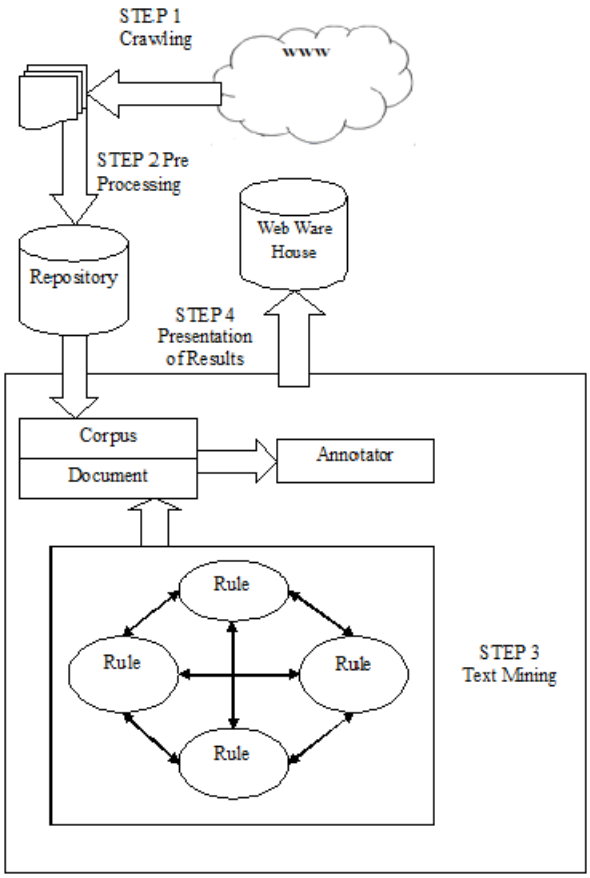

Figure 3.4 Web Text Mining Process

A. The FP-Growth Approach

The FP-growth algorithm consists of two major steps. First we must create an FP tree which is a condensed representation of the dataset. Then we need to create a mining algorithm that would generate all possible frequent patterns from the FP-tree by recursively traversing the conditional trees generated from the FP-tree [3]-[5].

B.Creating the FP-Tree

This algorithm first scans the DG database (DB) to find the edges that appear in the DG database. These edges and their corresponding supports are added to a list called F. Then we create a hash table called transaction DB for all the DGs. Hash table is a data structure that associates keys with values. Given a key it finds the corresponding value without linearly searching the entire list. The document ids are stored as the keys of our hash table. A value against a key stores a vector consisting of all the DFS codes of the edges for the DG in the corresponding key. Based on the minimum support (min_sup) provided by the user, the list F is pruned and sorted in descending order [3, 4 and 5].

C. Algorithm for FP-Tree Construction

Input DG database DB and
Minimum support threshold, *min_sup*.

Output Frequent Pattern Tree (T) made from each DGi in the DB.

1. Scan the DB once.

2. Collect *F*, the set of edges, and corresponding support of every edge.

3. Sort *F* in descending order and create *FList*, the list of frequent edges.

4. Create *transactionDB*

5. Create the root of an FP-tree *T*, and label it as "*null*".

6. For each DGi in the *transactionDB* do the following:

7. Select and sort the frequent edges in DGi according to *FList.*

8. Let the sorted frequent-edge list in DGi be *[p|Pi]*, where *p* is the first

element and *Pi* is the remaining list.

9. Call ***insert_tree([p|Pi], T)*** which performs as follows:

10. If *T* has a child *N* such that *N.edge_dfs = p.edge_dfs*,

then N.count++;

Else

create a new node *N*, with N.count=1

Link N to its parent *T* and link it with the same edge_dfs via the node-link structure.

If *Pi* is nonempty, call ***insert_tree([p|Pi], N)*** recursively.

D. Testing

To test the system, entered the universities' names into the system, and then measured the accuracy, precision and recall of the results returned by the system. In this way, the accuracy of the system is measured. The results are shown in a contingency table (confusion matrix). An example of a contingency table is shown in Table 1. Other than accuracy, Precision, Recall, and F-Measure of each of the different categories can be calculated. Precision is a ratio or percentage showing how many web pages was correctly identified out of the total number of pages available. Recall is the ratio or percentage showing how many pages in a class was correctly matched with respect to the total number of pages in that class. F- Measure is the harmonic mean of precision and recall.

This proposed work tested the proposed implementation of this approach on a sample space of about 4000 pages. These pages are gathered from four different universities. The results and various parameters used were shown in the below table.

Details	**Results**
No. of pages on which the proposed work have tested this implementation	4000
Pages categorized	3700
Pages Categorized correctly	3500
Percentage categorized correctly	94.59%

TABLE I PARAMETERS

Classified as	A	B	C	D	E	F	G
A = Course	N1	N2	N3	N4	N5	N6	N7
B = Department	N8	N9	N10	N11	N12	N13	N14
C = Faculty	N15	N16	N17	N18	N19	N20	N21
D = Other	N22	N23	N24	N25	N26	N27	N28
E = Project	N29	N30	N31	N32	N33	N34	N35
F = Staff	N36	N37	N38	N39	N40	N41	N42
G = Student	N43	N44	N45	N46	N47	N48	N49

TABLE I EXAMPLE OF THE EXPERIMENTS' RESULTS

E. Explanation

The highlighted cells in Table II are the ones used to calculate the overall accuracy of the model. Accuracy is the ratio or percentage of pages that were correctly classified. Using course category as example:

Accuracy = (N1+N9+N17+N25+N33+N41+N49) / $\sum_{k=1}^{49} N_k$

Precision course =N1/ (N1+N8+N15+N22+N29+N36+N43)

Recall course = N1 / (N1+N2+N3+N4+N5+N6+N7)

Algorithm	**Accuracy**	**Precision**	**Recall**
Apriori	80%	75%	50%
Fp Growth	95%	92%	91%

TABLE III COMPARISION TABLE OF ALGORITHMS

Details	**Total Pages**	**Correctly Classified**
Student	1641	1200
Faculty	1124	1000
Staff	137	120
Department	182	150
Project	540	400
Course	930	850
Others	3764	3500

TABLE IV PERFORMANCE OF FP-GROWTH WITH ALL THE FOUR UNIVERSITY DATASET

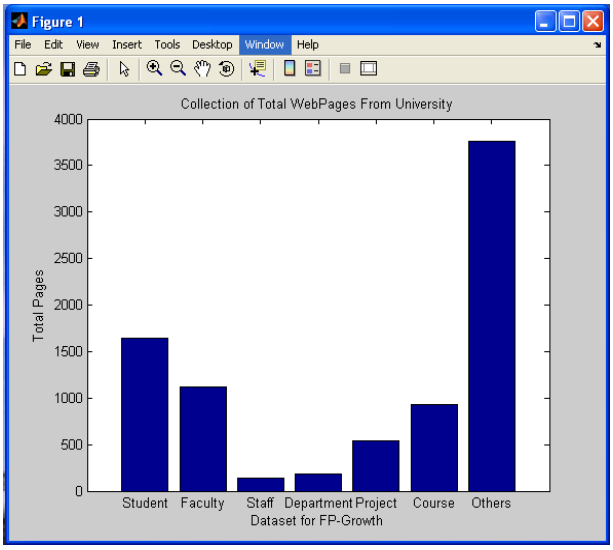

Fig 4.2 Data Set of FP-Growth

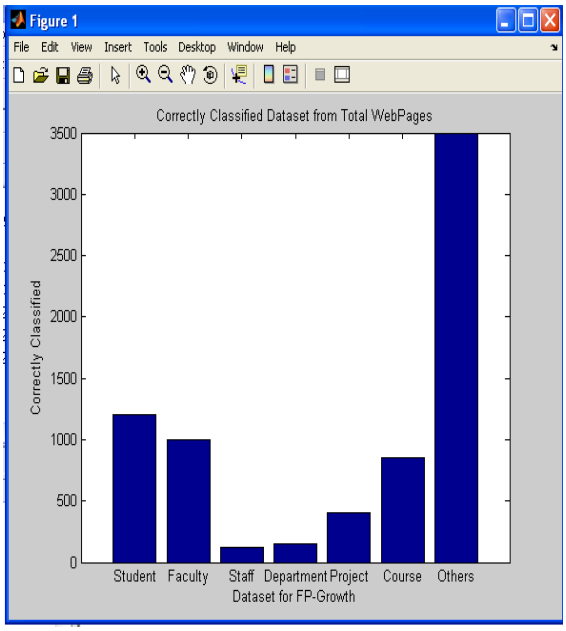

Fig 4.2 Performance of FP-Growth

VI CONCLUSION

The overall accuracy of the program should be higher. However, the training set provided was already stemmed, and the stemming algorithm was not mentioned. In this case, the test set stemming algorithm did not work as well as that used in the training set. As a result, there was a loss of words during the standardization step, i.e., when the testing set was being edited to be compatible with the model and the stemming step. These words were very important because they were what the model relied on to make a

prediction and classify the page. The ways in which this could be solved was either by making sure the training set and test set use the same stemming algorithm, or improve the stemming algorithm to be as good as the one which was used on the training set. This research work presented how to build a Web-based university search portal, i.e. by mining a benchmark university's Web portal structures, classifying the Web structure and using the Web structure as a model. This research work is also mined the Web portal of other universities and integrated these structures into the benchmark university's Web structure. Accuracy, precision and recall evaluation scores are promising.

REFERENCES

[1] R. Kosala, and H. Blockeel, Web Mining Research: A Survey, SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.

[2] Lilac A.E. Al-Safadi, Auto Classification for Search Intelligence, World Academy of Science, Engineering and Technology 49 2009.

[3] Pramod S. and O.P Vyas, Survey on frequent item set mining algorithms, IJCA, volume 1.

[4] C. Borgelt. An Implementation of the FP- growth Algorithm. Proc. Workshop Open Software for Data Mining (OSDM'05 at KDD'05, Chicago, IL), 1–5.ACMPress, New York, NY, USA 2005.

[5] Prateek Gupta and Surendra Mishra,"Improved FP Tree algorithm with customized web log preprocessing" IJCST Vol.2.