WEB PAGE CATEGORIZATION USING WEB MINING

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 primary goal of the web site is to provide the relevant information to the users. Web mining technique is used to categorize users and pages by analyzing users behavior, the content of pages and order of URLs accessed. In this paper, proposes an auto-classification algorithm of web pages using data mining techniques. The problem of discovering association rules between terms in a set of web pages belonging to a category in a search engine database, and present an auto – classification algorithm for solving this problem that are fundamentally based on FP-growth algorithm.**

***Keywords*- Association Rules, Auto-Classification FP – growth, Web Mining.**

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.

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.

The proposed technique has two phases. The first phase is a training phase where human experts determines the categories of different Web pages, and the supervised Data mining algorithm will combine these categories with appropriate weighted index terms according to the highest supported rules among the most frequent words. The second phase is the categorization phase where a web crawler will crawl through the World Wide Web to build a database categorized according to the result of the data mining approach. This database contains URLs and their categories [2].

This paper is organized as follows: Web Mining is introduced in Section 2, operation of web search engine described in section 3, the existing works are described in section 4, the proposed works are described in section 5, and in section 6 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 [3]. 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 consists 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

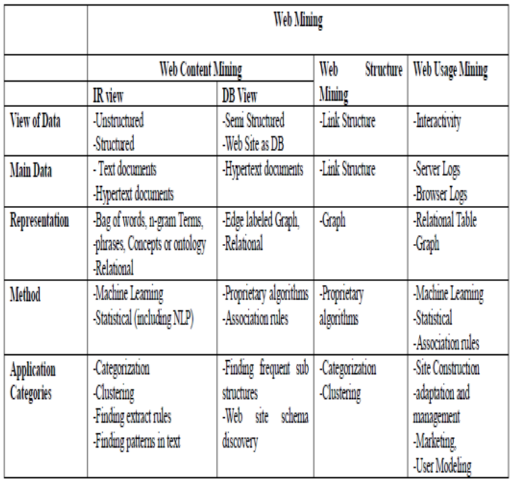
Web structure mining is concerned with discovering the model underlying the link structures of the web. Web structure mining used to categorize web pages and is useful to generate information such as the similarity and relationship between web sites. Web structure mining describes the study of topology 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. Table 1 gives an overview of above mining categories [4].

I Web Mining Categories

TABLE 1: Web Mining Categories



	Web Mining			
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR view	DB View		
View of Data	-Unstructured -Structured	-Semi Structured -Web Site as DB	-Link Structure	-Interactivity
Main Data	- Text documents -Hypertext documents	-Hypertext documents	-Link Structure	-Server Logs -Browser Logs
Representation	-Bag of words, n-gram Terms, -phrases, Concepts or ontology -Relational	-Edge labeled Graph, -Relational	-Graph	-Relational Table -Graph
Method	-Machine Learning -Statistical (including NLP)	-Proprietary algorithms -Association rules	-Proprietary algorithms	-Machine Learning -Statistical -Association rules
Application Categories	-Categorization -Clustering -Finding extract rules -Finding patterns in text	-Finding frequent sub structures -Web site schema discovery	-Categorization -Clustering	-Site Construction -adaptation and management -Marketing, -User Modeling

III. OPERATIONS OF WEB SEARCH ENGINE

A search engine work by storing and retrieving information from many web pages [8]. The operations are classified in the following order:

1. Web Crawling

2. Indexing

3. Searching

4. Ranking

A. *Web Crawling*

A web crawler fetches the web pages from the World Wide Web for parsing and indexing. It provides up-to –date data of web sites. Web crawlers are mainly used to create a copy of all the visited pages, and it's finally visit all the Web pages through the internet [9].

B. Indexing

Search engine indexing collect the parsed and stored data from the database for the fast and efficient information retrieval. This index contains a copy of each crawled page. The purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query [10].

C. Searching

A web search query is a query that accepts queries from the user, to perform concept based searching. The search queries are providing plain text or hypertext [11].

D. Ranking

Web search engine receives a search query and search through its index and finds the all web pages that are relevant to the search query. The indexed web pages are sorts, and the resulted links are displayed to the user.

IV. EXISTING METHOD
V.

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.
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.

A. Limitations of Apriori Algorithm

Apriori algorithm, in spite of being simple, has some limitation. They are,

- It is costly to handle a huge number of candidate sets.
- It is difficult to repeatedly scan the database and check a large set of candidates by pattern matching, which is especially true for mining long patterns.

In order to overcome the drawback inherited in Apriori, an efficient FP-growth based mining method is used. FP-growth, which contains two phases, where the first phase constructs an FP tree and the second phase recursively, researches the FP tree and outputs of all frequent patterns.

V. 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.

A. FP Growth Algorithm

FP-Growth algorithm allows mining frequent item set without candidate generation.
The algorithm consists of two steps:

- Compress a large database into a compact, for mining frequent set to build a FP-tree.
- Develop an efficient, FP-tree-structure, extracts the frequent itemsets directly. It divide into smaller ones to avoid candidate generation [7].

The example illustrates how to find frequent items from the FP-tree. The following databases have the following transaction to find all frequent itemsets using FP-Growth algorithm. Table 2, consists transactions and items, in table 3, the frequency of occurrence of each item in the databases are shown and in table 4, all the items are ordered according to its priority. Finally, Complete FP-tree structures are illustrated in figure 1.

II

TID	Items
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B,C
7	A,C
8	A, B, C, E
9	A, B, C

III

Item	Frequency	Priority
B	7	1
A	6	2
C	6	3
D	2	4
E	2	5

IV

TID	Items	Ordered Items
1	A, B, E	B, A, E
2	B, D	B, D
3	B, C	B, C
4	A, B, D	B, A, D
5	A, C	A, C
6	B,C	B, C
7	A,C	A, C
8	A, B, C, E	B, A, C, E
9	A, B, C	B, A, C

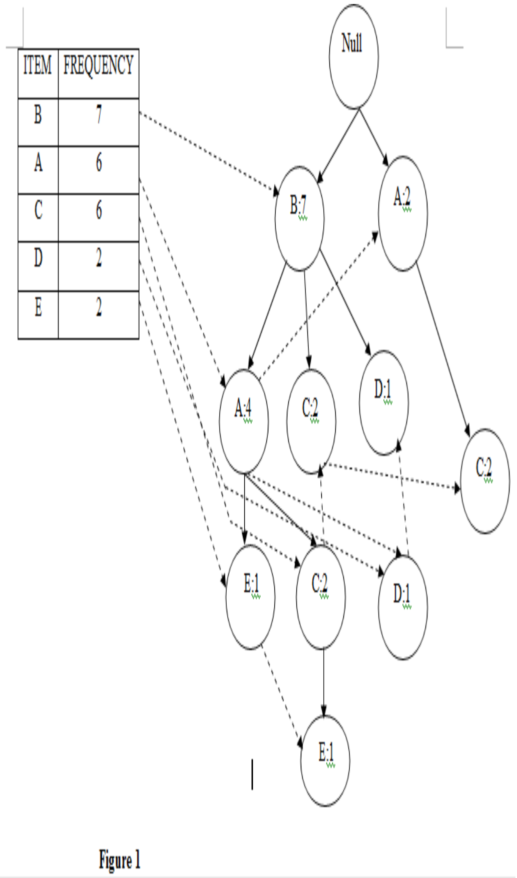


Fig 1: Complete FP Tree Growth

B. Working of FP-Growth

The proposed technique uses FP-growth to categorize the web pages. For instance the university website may contain the web pages for student, staff, and others.

Step 1: In the First step, the support items are scanned for the transaction. If the items are minimum support they are eliminated and it will be considered as inadequate for transactions.

Step 2: Then the items are sorted in descending order for the frequency to find the frequent items. The sorted items are indexed for the transaction purpose.

These sorted items are stored in array, which contains a pointer to the head of the list. It is easy to read the frequent items.

The sorted items are arranged in a tree format to fetch the relevant web pages.

In this proposed technique the FP Growth algorithm operates in the following four steps to categorize the web sites.

- Preprocessing
- FP Tree an FP Growth
- Association Rule Generation
- Results

The preprocessing techniques include data selection, cleaning and transformation which of them general steps for clean, correct and complete the input data for web mining requirements.

The second step is performed in two steps.

- FP Tree generation
- Applying FP Growth to generate association rules. FP tree is a compact data structure that stores important, information about frequent patterns.

C. Advantages of FP-growth algorithm

The major advantages of FP-growth algorithm is,

- Uses compact data structure.
- Eliminates repeated database scan.

FP-growth is an order of magnitude faster than other association rule mining algorithms and is also faster than FP-tree [5, 6].The algorithm reduces the total number of candidate item sets by producing a compressed version of the database in terms of an FP-tree. The FP-tree stores relevant **i**nformation and allows for the efficient discovery of frequent item sets [5, 6].

VI. CONCLUSION

Web mining is the Data Mining type that automatically discovers or extracts the information from web documents. In this paper, An Enhanced Novel approach on Web page Categorization is discussed. Our goal of research is finding the frequent web pages as efficient as the other frequent mining rules.

REFERENCES

[1] Chekuri, M. Goldwasser, P. Raghavan, and E. Upfal, "Web Search Using Automatic Classification," Proceedings of the 6th International World Wide Web Conference, April 1997.

[2] J. Hou and Y. Zhang, Effectively Finding Relevant Web Pages from Linkage Information, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 4, 2003.

[3] 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.

[4] Arun K Pujari, “Data Mining Techniques”.

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

[6] 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.

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

[8] http://en.wikipedia.org/wiki/ Web_search_engine.

[9] http://en.wikipedia.org/wiki/ Web crawling.

[10] http://en.wikipedia.org/wiki/ Index_ (search engine)

[11] http://en.wikipedia.org/wiki/ Web_search_query.