International Journal of Recent Engineering Research and Development (IJRERD)
ISSN: 2455-8761
www.ijrerd.com || Volume 04 – Issue 04 || April 2019 || PP. 01-05

# Website Classification Based On Data Mining

[1]Kuche Bhavani Priya, [2]Vanishree K
*[1]Department of Information Science and Engineering*
*RV college of Engineering*
*Bengaluru-560060*
*[2]Assistant Professor*
*Department of Information Science and Engineering*
*RV college of Engineering*
*Bengaluru-560060*

**Abstract:**The world wide web is vast. There are many HTML, XML documents, images and other multimedia files available out there. Data mining plays a vital role in classifying the websites. Data mining involves locating patterns in the data sets and involves different methods at the intersection. On the basis of the kind of data to be mined, there are two categories involved in Data Mining which are Descriptive and Predictive functions. There are six data mining class tasks which are Classification, Clustering, Association rule discovery, Sequential pattern discovery, Regression and Deviation detection. Out of which only classification and Regression belong to Predictive function and remaining tasks belongs to Descriptive function. Data mining is the process to extract information from the data set and transform it into a required and understandable structure. In this paper, we will discuss the techniques. A web search engine is a software system that is designed to carry out web search. The operations of the web search engine are Web-crawling, Indexing, Searching and Ranking. The goal of this study is to provide a comprehensive review of different techniques in data mining.
**Keywords:**Web content mining, Web structure mining, web usage mining, Apriori algorithm, FP growth algorithm.

## Introduction

There are two categories involved in data mining which are Predictive and Descriptive. Predictive tasks predict unknown or future values of other variables. Basically they determine what might happen in future. Descriptive tasks find human understandable patterns that describe the data. Basically they determine what happened in the past. With the help of descriptive tasks, we can frame predictive tasks. Out of all the six data mining class tasks only Classification and Regression belong to predictive tasks because we need to know about the past relation and then classify or create a linear relationship. Web mining is a data mining technique where the web search engine performs mining on the databases. In this paper, we will discuss about two algorithms which are Apriori algorithm and FP growth algorithms.

**Why data mining?**

Data mining plays a vital role in discovering patterns and relationships in the given large data sets which helps tomake the business better. It can help the business people to get an analytical result about their products in market. Some of the data mining uses are:

- Direct marketing
- Fraud detection
- Market segmentation
- Interactive marketing
- Analyzing the trend

FP growth technique has two phases. The first phase is a training phase where experts determine the categories of the web pages and the supervised Data mining algorithm will combine these categories with appropriate weighted index terms among the most frequent words. The second phase is the categorization phase where a web crawler which is also called spider 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.

## Web content mining

Web content mining is related to data mining and text mining but it different when compared with those two. It is kind of related because its text almost everywhere in web contents. Web content mining is also different from text mining because of semi structure nature of the web, while text mining focuses on unstructured texts. In the past few years, there was a rapid expansion of activities in the area of web content mining.

## Web structure mining

The main motto of the web structure mining is to produce a structural summary about the web pages and web sites. Basically, Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level.

Web page contains a link to other web page, if they are linked directly or if they are neighbors. Then we can easily discover the relationship among those web pages. The main task of web structure mining is to locate the nature of the hierarchy of a particular domain. Because of this it is easy to generalize the flow of information in Web sites that may represent some particular domain, therefore the query processing will be easier and more efficient.

## Web usage mining

Extracting useful usage patterns from the web data this process is called web usage mining. As a sub-field of data mining, Web usage mining focuses specifically on finding patterns relating to users of a Web based system: who they are, what they tend to do, etc.

We usually use association rules to discover interesting relations between variables in large databases. It is a rule based machine learning method. Apriori and FP growth algorithms comes under association rule. Let's discuss both the techniques, their advantages and disadvantages.

Given records each of which contain number of items from a given collection. Generate dependency rules which will predict occurrence of an item based on occurrences of other items. Given support 50% and confidence 75%.

For example,

| Transaction Id | Items Purchased |
|---|---|
| 1 | Bread, Cheese, Egg, Juice |
| 2 | Bread, Cheese, Juice |
| 3 | Bread, Milk, Yogurt |
| 4 | Bread, Juice, Milk |
| 5 | Cheese, Juice, Milk |

Table 1

We need find out the support where

$$\text{Support} = \frac{\text{Number of occurrences of Bread (Frequency)}}{\text{Number of transactions}}$$

| Item | Frequency | Support |
|---|---|---|
| Bread | 4 | 4/5 = 80% |
| Cheese | 3 | 3/5 = 60% |
| Egg | 1 | 1/5 = 20% |
| Juice | 4 | 4/5 = 80% |
| Milk | 3 | 3/5 = 60% |
| Yogurt | 1 | 1/5 = 20% |

Table 2

Bread occurs 4 times in Table 1 and the total number is transactions is 5. Hence the support of bread is 4/5 which is 0.8 or 80%. And it is calculated similarly for all the items.

As the given support is 50% any item whose support is less than 50% has to be eliminated. Hence, Egg and Yogurt will be eliminated as their support percentage is less than 50. Now we need to make pairs of the rest items and again calculate their support.

Now we need to look for the number of occurrences of the pair formed and then divide it by number of transactions.

| Items pair | Frequency | Support |
|---|---|---|
| Bread, Cheese | 2 | 2/5 = 40% |
| Bread, Juice | 3 | 3/5 = 60% |
| Bread, Milk | 2 | 2/5 = 40% |
| Cheese, Juice | 3 | 3/5 = 60% |
| Cheese, Milk | 1 | 1/5 = 20% |
| Juice, Milk | 2 | 2/5 = 40% |

Table 3

As the given support is 50% any item whose support is less than 50% has to be eliminated. Hence, we eliminate the following pairs – Bread, Cheese

Bread, Milk
Cheese, Milk
Juice, Milk

The remaining pairs are (Bread, Juice) – (1)
(Cheese, Juice) – (2)
Rules – Bread, Juice

The customer might first buy bread and then juice or vice-versa. we can find that by confidence.

Confidence = Support (A U B) / Support (A)
Here,
Confidence (Bread → Juice) = Support(Bread U Juice) / Support (Bread)
= 60/ (4/5) = (60*5)/4 = 75

Confidence (Juice → Bread) = Support(Juice U Bread) / Support (Juice)
= 60/ (4/5) = (60*5)/4 = 75

Both rules are good to be implemented because both the results are more than or equal to the given confidence that is 75.

Now we apply the same method for the (2).
Confidence (Cheese → Juice ) = Support(Cheese U Juice) / Support (Cheese)
= 60/ (3/5) = (60*5)/3 = 100

Confidence (Juice → Cheese ) = Support(Juice U Cheese) / Support (Juice)
= 60/ (4/5) = (60*5)/4 = 75

Both rules are good to be implemented because both the results are more than or equal to the given confidence that is 75.

Hence, we got 4 dependency rules.

Now let's discuss about FP growth algorithm.

Given a set of records each of which contain some number of items from a given collection. Generate FP tree for the following transaction data set. Given minimum support=30%.

| Transaction ID | Items |
|---|---|
| 1 | E, A, D, B |
| 2 | D,A,C,E,B |
| 3 | C,A,B,E |
| 4 | B,A,D |
| 5 | D |
| 6 | D,B |
| 7 | A,D,E |
| 8 | B,C |

Table 4

Minimum number of transaction =

$$\frac{\text{Support}}{\text{Number of transactions}}$$

= 30/8 = 2.4

We calculate the frequency and decide the priority of each item. Lower priority means Higher priority.

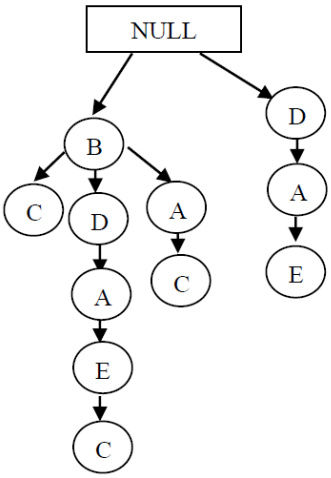
| Item | Frequency | Priority |
|---|---|---|
| A | 5 | 3 |
| B | 6 | 1 |
| C | 3 | 5 |
| D | 6 | 2 |
| E | 4 | 4 |

Table 5

As 6 is the highest number we gave it the highest priority that is the lowest number (1) and as 3 is the lowest frequency number we gave it the least priority that is the highest number (5). The order according to priority is B,D,A,E,C.

| Transaction ID | Items | Ordered items |
|---|---|---|
| 1 | E, A, D, B | B,D,A,E |
| 2 | D,A,C,E,B | B,D,A,E,C |
| 3 | C,A,B,E | B,A,E,C |
| 4 | B,A,D | B,D,A |
| 5 | D | D |
| 6 | D,B | B,D |
| 7 | A,D,E | D,A,E |
| 8 | B,C | B,C |

We wrote the ordered items using the priority order. Now we draw a FP tree. FP-Growth simplifies all the problems present in Apriori by using a structure called an FP-Tree. In an FP-Tree each node represents an item and its current count, and each branch represents a different association.




By the end when we count the number of each item they will correspond to the respective Item frequency.

## Conclusion

Apriori algorithm or technique suffers from a number of inefficiencies. The algorithm has to scan the whole database many times, which reduces the performance. Due to this, the algorithm assumes that the database will be in the Permanent memory. The space and time complexity of this algorithm is very high. These techniques help us to generate rules and classify the web pages accordingly. FP tree helps us to traverse easily and get know the location of the item very easily.

## References

[1]. Anurag Kumar and Ravi Kumar Singh, "Web Mining Overview, Techniques, Tools and Applications: A Survey," International Research Journal of Engineering and Technology (IRJET), vol. 03, no. 12, pp. 1543-1547, December 2016.

[2]. Simranjeet Kaur and Kiranbir Kaur, "Web Mining and Data Mining: A Comparative Approach," International Journal of Novel Research in Computer Science and Software Engineering, vol. 2, no. 1, pp. 36-42, January - April 2015.

[3]. Ahmad Tasnim Siddiqui and Sultan Aljahdali, "Web Mining Techniques in E-Commerce Applications," International Journal of Computer Applications, vol. 69– No.8, pp. 39-43, May 2013.

[4]. Kshitija Pol, Nita Patil, Shreya Patankar, and Chhaya Das, "A Survey on Web Content Mining and extraction of Structured and Semistructured data," Emerging Trends in Engineering and Technology, pp. 543-546, July 2008.

[5]. R. Malarvizhi and K Saraswathi, "Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study," International Journal of Computer Trends and Technology (IJCTT), vol. 4, no. 8, pp. 2940-2945, Augest 2013.

[6]. Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey," SIGKDD Explorations, vol. 2, no. 1, pp. 1-15, July 2000.

[7]. Faustina Johnson and Kumar Santosh Gupta, "Web Content Mining Techniques: A Survey," International Journal of Computer Applications (0975 – 888), vol. Volume 47– No.11, pp. 44-50, June 2012.

[8]. R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web data. PhD thesis, Dept. of Computer Science, University of Minnesota, May 2000.

[9]. R. Cooley, B. Mobasher, and J. Srivastava. Data preparationfor mining World Wide Web browsing patterns. Knowledge and Information Systems, 1(1), 1999.