# Web Mining Using Behavioral Modeling

Syed Muhammad Tajdar Khalid[1], Muhammad Nadeem[2]

[1]*MSCS SZABIST, Karachi*
Taji_tool@hotmail.com

[2]*SZABIST, Karachi*
nadeem@szabist.edu.pk

**Abstract:** *In past few years the rapidly expansion of the World Wide Web go beyond all expectations. These days there are various numbers of HTML documents, images and other multimedia files available via internet and the number is still growing. To examine such data can help any organizations to evaluate the life time value of consumer, design cross marketing strategies across products and services, analyze the efficiency of promotion campaign. So, Web mining implies in this sort of situation. Web mining can be classified into three wide areas, Web usage mining, web structure mining and web content mining. Web mining methodology is utilizing these techniques to get the valuable data in no time for the customer. This paper addressed the customer behavior incorporate in web mining to enhances the effectiveness, improve the growth of the business and provide the comfort to the customer. Customer behavior can be extracted for various Web mining tasks such as the discovery of association rules, frequency and sequential patterns from the Web data.*

**Keywords:** *Web mining, web content mining, web usage mining, Web structure mining, Market basket analysis, Apriori algorithm*

## 1. INTRODUCTION

The involuntary sighting of interesting and helpful information from the Web is always been the major apprehension. Web is a world of copious data and information, where user's major task is to discover data that is valuable and relevant according to his requirements. Furthermore the bundle of information stored on the Web is haphazard (i.e. not in any specific sequence), causing real challenges for those searching for high quality information and to reveal the knowledge masked in heaps of web pages.

To overcome this major challenge a progressively essential research field called as web mining was introduced. Web mining is a mechanism which makes use of advanced machine learning techniques to understand the complex structure of web data. Web data are primarily based on multimedia streams that include text, sound, images, and a range of database information.[1] Whereas useful information extraction, direction-finding or organization involve mining of all media methodologies, this report focuses on web mining with comprise of user behavior modeling.

Behavioral Modeling is to concern about the understanding the customer way to acquire and utilize the products. Its important for the growth of business to know that what are the things that customer looking is for by using different techniques. There are various ways to influence consumer such as quality, discount price, new products etc.

This research paper illustrate web mining, techniques used in web mining, Behavioral Modeling, Algorithm for Behavioral modeling, its example and implementation.

## 2. WEB MINING

Web mining aims to discover interesting patterns in the structure, the contents and the usage of web sites. An indispensable tool for the webmaster, it has, nevertheless, a long road ahead in which visualisation plays an important role [2], whereas excess of data on the web is one of the major challenges face by web mining.

### 2.1 Classification of Web Mining

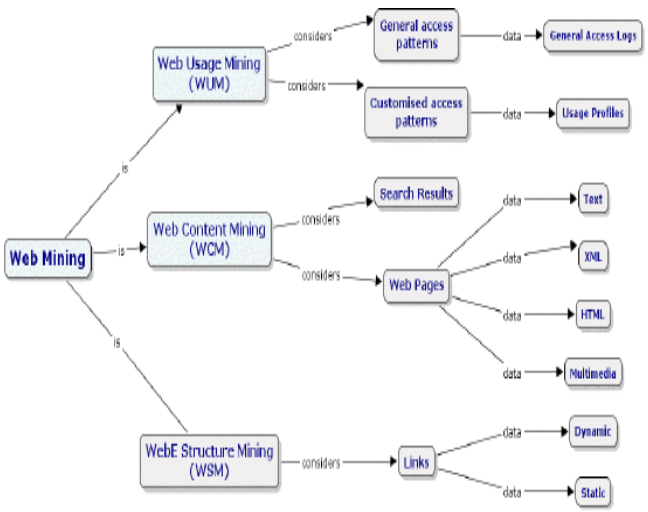


Figure 1: Conceptual Map of Web Mining [2]

## 2.2 Web Mining Techniques

*2.2.1 Web Content Mining (WCM)*

Web content mining is defined as the extraction for valuable information from web data. User retrieves most of the data for the desired topic in a systematic manner. Web content mining emphasize on the fact that data present on the web has no machine readable format and thus goes ahead of simple keyword extraction.

Web content mining can be divided into two separate approaches:

i) **An agent based approach** in which the data on the web are directly mined.
ii) **A database approach** It is based on the enhancement of web mining used by the search engine with regard of backend database it utilizing.

*2.2.2 Web Usage Mining*

Web usage mining is a web mining technique in which Data Mining are applied to Web usage data. In general Web usage mining is defined as automatic recognition of user access behavior from the server. Web servers are responsible to automatically collect the data generated and routine operation in their server access logs.

If we look at the previous researched area in web usage mining we found out a Personalization is the one who has been mostly researched. Adaptive web sites as an early effort result are the direction of personalization, which automatically change their presentation and arrangements with respect to the user using them. One of the examples of it is Amazon web site which is using this technique and provides the recommended links with the same property to the user buying any thing.

There are two approaches of WUM:

i) **General access pattern tracking** The general access pattern examines the web server logs to recognize usage pattern and style. This can result in improvement of structure and aligning of resource providers. Executing the data mining methodology on we server access logs reveal useful access pattern that can be further used to streamline sites in more competent manner and provide the user what he is looking for.
ii) **Customised access pattern tracking** Customized usage tracking examine individual styles. It is used to personalize web sites according to users. The information view, the structure of site and the pattern of the resources can be dynamically adapted for every user over time based on their usage patterns

*2.2.3 Web Structure Mining*

Web structure mining is defined as the system which examines the link structure of web site in order to improve routing between the pages. Generally web content mining primarily dealing with structure of inner – document whereas structure mining determine link structure of hyper links used for the navigation between pages. Hyperlinks anatomy helps to organize the pages and generate the valuable information like the association among the different sites.

The two approaches of WSM are:

1. Static The content in the page remains unchanged until any changes are made in Hypertext Mark-up language (HTML). URL of the static page does not change and it does not have any variable to be evaluated during rending the page. Example of static link is: http://www.anysite.com/anytopic/anypage.htm
2. Dynamic The content of the page are displayed as a result of search from the database against the query string (variable containing value to be executed). The dynamic pages have only a template and design in which the data is going to be displayed. The change affect the database rather then HTML code. Example of dynamic link is: http://www.anysite.com/anytopic/anypage.aspx?anyid=1005.

## 3. BEHAVIORAL MODELING

Behavior Modeling is a method which identifies the association among the pair of products Order or purchased together. The method can be used to figure out the related product. Behavior modeling often called Market Basket Analysis

Market Basket Analysis looks at your orders for products have been buying mutually. Like by using market basket analysis you might discover the fact that customers be likely to buy bread and butter / Jam together. Using this kind of information, supplier should arrange the store item in such a manner that bread and butter / Jam are next to each other. And also when talking about an e-commerce site you might build up a cross-sell methodology to offer the buyer Jeans and belt whenever they place Shirt in their shopping cart and to the vice versa .

### 3.1 Some Examples Sites

Example 1 (Figure 3) displaying the recommended items list (Diamond rings) to user after adding the diamond solitaire to their shopping cart.

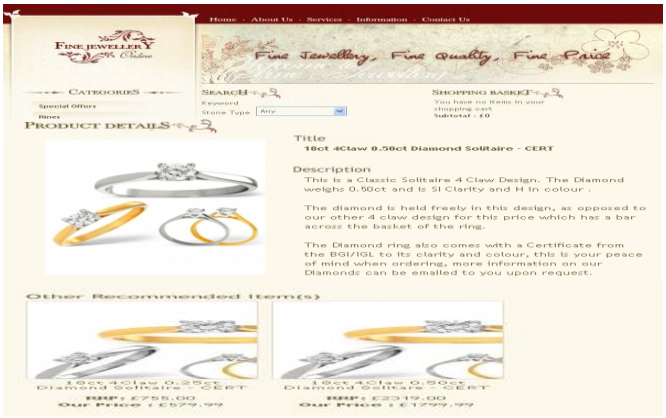


Figure 2: 18ct 4Claw 0.50ct Diamond Solitaire - CERT [3]

Example 3 (Figures 5) Displaying recommended links of books purchased together in most of the time after customer add the book to the shopping basket he wants to purchased


Figure 3: Social Computing, Behavioral Modeling, and Prediction [4]

### 3.2 Techniques to Measure Market Basket Analysis

*3.2.1 Frequency*

The frequency is defined as the number of times that two Items were procured together. Like if mobile phone and sim card, two Items are together in 940 baskets then this would be showing its frequency.

Frequency on its own doesn't explains the complete scenario for this fact consider the above example that if mobile phone and sim card were purchased 940 times together how would any one knows that was related to each other or not.

*3.2.2 Association Rules*

Association and sequencing tools analyze data to discover rules that identify patterns of behavior, e.g. what products or services customers tend to purchase at the same time, or later on as follow-up purchases. The process of using an association or sequencing algorithm to find such kinds of rules is frequently called market basket analysis. Detailed description of this method can be found in the book by M. Berry and G. Linoff [5].

**Example of rule:**

When people purchase a mobile phone they also purchase a sim card 20% of the time.

Association rule has two measuring criteria such as support and confidence.

**Support (prevalence)** indicates the frequency of a pattern, i.e. how often items occur together. In the example above, the confidence is 20%. Rules with a low value for support might simply be due to a statistical anomaly. A minimum support is necessary if an association is going to be of some business value [5].

If X and Y then Z with support s. The rule holds in s% of all transactions.[5]

Support is computed as follows:

s(A=>B) = P(A U B).[5]

**Confidence (predictability)** denotes the strength of an association, i.e. how much a specific item is dependent on another [5].

If X and Y then Z with confidence c [5].

If X and Y are in the basket, then Z is also in the basket in c% of the cases [5].

c(A=>B) = P(B|A) = P(AUB)/P(A) [5]

*3.2.3 Sequential Patterns Discovery*

Sequential patterns discovery (sequencing) make use of time comparisons between transactions in order to broaden association rule.

This shows that Sequential patterns discovery, not only focuses on items existing together within a transaction, but also on the sequence of the items emerge against ordered transactions and the total time between transactions being executed.

This is obtained by switching every time series into multi-item transaction and separating the replicas of items

afterwards association rule can be implemented to the transactions [5].

## 4. ALGORITHM AND IMPLEMENTATION

### 4.1 APriori Algorithm

The Apriori Algorithms an influential algorithm for mining frequent itemsets for Boolean association rules [6].

#### 4.1.1 Pseudo-code

Join Step: CK is generated by joining Lk-1 with itself
Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset [6]

CK: Candidate itemset of size k
LK: frequent itemset of size k

L1= {frequent items};
for(k= 1; LK!= Ø; k++) do begin
CK+1= candidates generated from LK;

for each transaction tin database do increment the count of all candidates in CK+1that are contained in t
LK+1= candidates in CK+1with min_support
end
return ỤK LK; [6]

### 4.2 Implementation of Algorithm

*4.2.1 Background*

Itemset can be defined as the set of items present in dataset. Example of itemset can be {Shirt, Tie, Jeans, Shoes, Socks}

The dataset can be build up from the real world database containing the transaction carried out on e- commerce site on daily bases, Sale of Items, Purchase of items, Item orders etc. Association rule is applied on this type of data mining where the relation between the items is identified in dataset.

Association rule example can be a Market basket analysis. Market basket analysis figure out the relevant items purchased together by user. Based on the market basket analysis of customer shopping cart one may finds the relation among shirt and jeans showing that customer that purchased shirt be likely to purchased jeans also. This can direct to a tactical placement of item like in our case it is shirt and jeans so that more jeans will be sold and increasing the overall sell when shirt will be bought.

#### 4.2.2 Working of APriori Algorithm

Apriori algorithm is used to find out number of transaction in which the items often buy together.
Considering the example discussed above in which the sale of jeans is increased by the sale of shirts. Support and Confidence are two measures applied for the market basket analysis while implementing the association rules formed with an APriori algorithm.

Lets assume that; Support = 8%, Confidence = 35%
This shows that 8% of transaction performed by the customer contains shirts and jeans and 35% of customer who purchased shirt also purchased jeans

Itemset having set of item which contains k items is called k item set. For e.g. a set of item {Shirt, Jeans, Tie} has 3 itemset.

The number of transaction that includes itemset is called frequency or Support count and Itemset is said to be frequent itemset if it attains the minimum support count.

*4.2.3 Implementation*

The application from [7] is used for evaluating the real time live example of market basket analysis. The xml file which contains the 9 database transaction is executed. Sample xml file showing transactions are as:

```
- <SourceCode name="StartingSampleC">
  - <summary>
    - <para>
      - <example>
        - <TransactionTable>
            <TransactionID>1</TransactionID>
            <Transactions>Shirt, Jeans, Belt</Transactions>
          </TransactionTable>
        - <TransactionTable>
            <TransactionID>2</TransactionID>
            <Transactions>Jeans, Tie</Transactions>
          </TransactionTable>
        - <TransactionTable>
            <TransactionID>3</TransactionID>
            <Transactions>Jeans, Shoes</Transactions>
          </TransactionTable>
        - <TransactionTable>
            <TransactionID>4</TransactionID>
            <Transactions>Shirt, Jeans, Tie</Transactions>
          </TransactionTable>
        - <TransactionTable>
            <TransactionID>5</TransactionID>
            <Transactions>Shirt, Shoes</Transactions>
          </TransactionTable>
        - <TransactionTable>
            <TransactionID>6</TransactionID>
            <Transactions>Jeans, Shoes</Transactions>
          </TransactionTable>
        - <TransactionTable>
            <TransactionID>7</TransactionID>
            <Transactions>Shirt, Shoes</Transactions>
          </TransactionTable>
        - <TransactionTable>
            <TransactionID>8</TransactionID>
            <Transactions>Shirt, Jeans, Shoes, Belt</Transactions>
          </TransactionTable>
        - <TransactionTable>
            <TransactionID>9</TransactionID>
            <Transactions>Shirt, Jeans, Shoes</Transactions>
          </TransactionTable>
        </example>
      </para>
    </summary>
</SourceCode>
```

Figure 4: Sample Transaction Table

The example displayed above {Shirt, Jeans} 2 – items set has a support count of 4 referring to figure 4.

Apriori algorithm create an association data mining rule between {Shirt} and {Jeans} using the market basket analysis.

**Data:**
Considering the 2-itemset {Shirt, Jeans} has the number of transaction is 4, and
Number of transaction having the itemset {Shirt} is 6.

**Measuring Support**

- Support for the 2-itemset {Shirt, Jeans} is (4/9) * (100%) = 44.44%.

**Measuring Confidence**

- The confidence for the 2-itemset {Shirt, Jeans} can be calculated as = (Support Count ({Shirt, Jeans}) / Support Count ({Shirt}) * (100%).
- So, the confidence for the 2-itemset {Shirt, Jeans} is = ((4/6) * 100%) = 66.67%.

Support count and confidence of itemset can be found using Appriori algorithm by removing those itemset that do not fulfill a minimum support count and confidence measure through rules created.

Following are the result generated with Support = 44% and Confidence = 66%

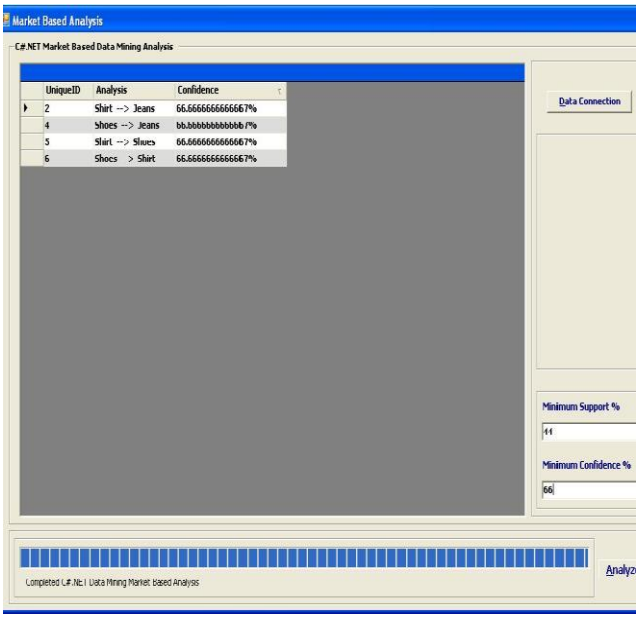


Figure 5: Market Basket Result With Support 44% and Confidence 66%

Support count and Itemset is said to be frequent itemset if it attains

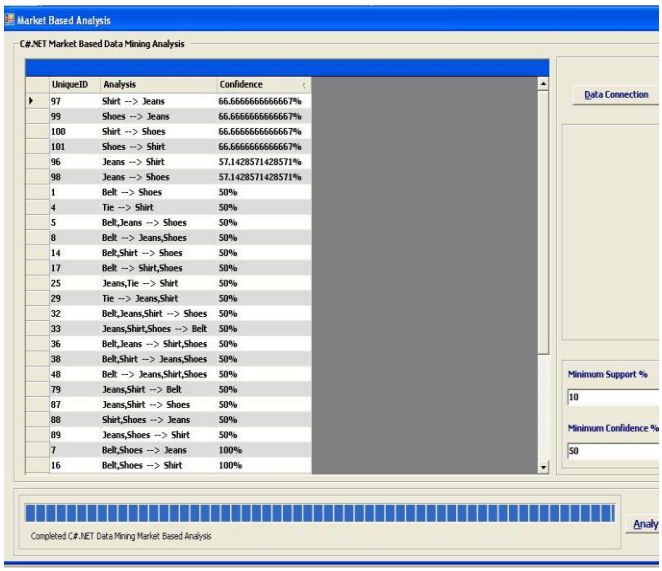


Figure 6: Market Basket Result with Support 10% and Confidence 50%

### 4.2.4 Result Description

i) The APriori algorithm makes a list of distinctive items in a 1 itemset Candidate Itemset equivalent to {Shirt, Jeans, Shoes, Tie, Belt}

ii) The support count of each item is acquire and any item that does not fulfill the minimum support count is removed from more analysis making a 1-itemset frequent itemset

iii) 2 itemset candidate itemset is created after coupling the 1 itemset frequent itemset with itself.

iv) Repeat the steps for the 2 itemset candidate itemset taken for the1 itemset candidate itemset.

v) Repeat the above step till frequent itemset become empty and no new candidate itemset produced

## 5. CONCLUSION

Web is approaching towards globalization and we need to respect user. Repository of Web is increasing day by day in no time, responsible of the availability of data for every aspect of life whether it's was past, is present or forth coming future. Web mining is the technique to extract the useful knowledge from the huge amount of data present on world wide web and after combining with the behavior modeling will result in undoubtedly increases in sale of organization. This can be achieved from the knowledge of user, his past interest, interactions, how user is taking their decision, what factors he kept in his mind while purchasing any item.

## REFERENCES

[1] Jan Larsen, Lars Kai Hansen, Anna Szymkowiak, Torben Christiansen and Thomas Kolenda. "Web mining Learning from the World Wide Web." Internet: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.38.62. [April 25, 2009].

[2] Juan C. Dürsteler. "Web Mining." The digital magazine of InfoVis.net. Internet: http://www.infovis.net/printMag.php?lang=2&num=172. September 18, 2005. [April 25, 2009].

[3] Fine Jewellery Online: Fine Jewellery, Fine Quality, Fine Price. Internet: http://finejewelleryonline.com/details.php?id=ML1637HS. [April 25, 2009].

[4] Amazon.com: Online shopping for Electronics, Apparel, Computer, Book, DVS & more. Internet: http://www.amazon.com/gp/product/handle-buy-box/ref=dp_start-bbf_1_glance. 2009. [April 25, 2009].

[5] Anita Wasilewska. "Aprior Algorithm." Internet: http://www.cs.sunysb.edu/~cse634/lecture_notes/07apriori.pdf. [April 25, 2009].

[6] Kingsley Tagbo. "C# Apriori Algorithm source code Release – 2001 Vetsion For Market Basket Analysis." Internet: http://www.kdkeys.net/forums/thread/2043.aspx. August 21, 2004. [April 25, 2009].