Analytical Study for Pattern Mining In E-Commerce Data

Prof. S. P. Godse¹, Shilpi Singh², Anushree³, Shaikh Nafisa⁴ & Nikita Medidar⁵
¹Internal Project Guide
¹²³⁴⁵SAE, Kondhwa, Pune (India)

Abstract: In today’s digital world data is been increasing like anything largely in unstructured file format. There are many data mining techniques are there to mine these files. But it is bit challenging affair to mine semi structured data like XML. Many researchers are engaged in mining rules from XML data using popular algorithms like Apriori. As Apriori creates large amount of candidate sets so it needs more processing space and execution time is also high as it keep creating the candidate sets at every time ,this creates confusion in many researchers to choose Apriori for large data sets. So our idea of extraction of rules from large datasets like Reuters using Eclat rule mining algorithm which uses intersection of transactions for generating candidate item sets. Our approach enhances the Eclat performance by enforcing comparative vertical power sets for creation of candidate item sets to enhance the quality of the rules with less processing space and also with less execution time.

1. Introduction

The Mining of data provide an efficient way to extract summary of data which is substitution of original content. This can be carried out by using following conditions. Preferably huge amount of documents is given as input for data mining or rule mining techniques input as the documents in huge number and filtration of these data is been carried out by many information retrieval methods. Most widespread job in data mining is a classification of data that automatically forecast the class for an unseen instance as precisely as possible. While in single label classification that assigns each rule as a classification has been widely used as a most obvious label, moreover discovery of all association rule is another important task in data mining. It has been depicted that classification and association rule are similar, where association rule can analyze any given attribute in the data. Wherein, classification there is only one answer to guess in classification. At present prediction of multiple labels using associative classification technique are not suitable using rules, since only one class label is associated with each derived rule. Moreover, multi-label classification is often useful in practice.

2. Architecture

Fig.1: System Architecture

3. Modules

3.1 Preprocessing
Following is the three steps of preprocessing the result obtained are preprocessed. For that we have to follow some preprocessing steps. These steps are:
3.1.1 Stop words Removal
3.1.2 Tokenization
3.1.3 Stemming

3.2 TF-IDF:
This module will help to calculate tf-idf where df, - document frequency for term t idf, - inverse document frequency for term Idf, =log ( N / df, ) where N – Total number of document tf-idf- a combined weight for term t in document d.

3.3 Term Weight:
This model eliminates the most repeated word in the document. It works as tf-idf. This model is applied on the whole database.
3.4 Shanon Info-Gain:
This method extracts the similarity structure among a set of documents through a hierarchical clustering. It gives higher weights to words that contribute to forming the structure. Important words are calculated based on IGR as follows:

\[ \text{IGR}(C) = -\sum \left( \frac{|C_i|}{|C|} \log \left( \frac{|C_i|}{|C|} \right) \right). \]

Where \( C_i \) is the frequency of the word \( w \) in Cluster \( C \).

3.5 Association Rule Generation
For each frequent itemset \( L \), generate all non-empty subsets of \( L \).
For every non-empty subset \( S \) of \( L \), output the rule:

\[ S \Rightarrow (L - S) \]

If \( \frac{\text{support_count}(L)}{\text{support_count}(S)} \geq \text{min_conf} \)

3.6 Frequent ItemSets:
This module takes care of generating frequent item sets of the pattern.

4. Algorithms
Input: frequent Item sets \( p \subseteq P(A) \) called Prefix, incidence matrix \( C \) of frequent 1-item extensions of \( p \)
Output: add all frequent extensions of \( p \) to global variable \( F \).
for \((x, T_x) \in C\) do
\( q := p \cup \{X\} \)
\( C_q := \{ (y,T_y \cap T_x) \mid (y,T_y) \in C, y > x \} \)
\( C_q' := \text{freq}(C_q) := \{ (y,T_y) \mid (y,T_y) \in C_q, |T_y| \geq \text{minsup} \} \)
If \( C_q' \neq \emptyset \) then
Add frequent supersets \((q,C_q')\)
End if
\( F := F \cup \{(q, |T_q|)\} \)
End for

5. Implementation

![Application Screenshot 1](image1)
![Application Screenshot 2](image2)
![Application Screenshot 3](image3)
![Application Screenshot 4](image4)
![Application Screenshot 5](image5)
6. Discussions

To evaluate the effectiveness of the proposed approach, we examined how many relevant rules can be extracted for both Apriori and Eclat Algorithm. The retrieval effectiveness can be defined in terms of precision and recall rates. So precision can be defined as the ratio of the number of relevant rules retrieved to the total number of irrelevant and relevant rules retrieved. It is usually expressed as a percentage. This gives the information about the relative effectiveness of the system. Whereas Recall is the ratio of the number of relevant rules retrieved to the total number of relevant rules in the database. It is usually expressed as a percentage. This gives the information about the absolute accuracy of the system.

7. Conclusion and Future Scope

In the proposed approach of mining association rules system efficiently enhance the feature of Eclat algorithm with comparative power set. Comparative power set extract the maximum frequent itemsets from important words which are been decided by tf-idf and Shannon information gain. Proposed system enforces the powerset with multi recursion methodology to get as maximum as possible of intersection transactions. This method actually enhances the Eclat algorithm to create frequent itemsets on intersection and thereby to reduce the space and time complexity efficiently.

8. Acknowledgements

The survey paper on the topic “Analytical Survey For Pattern Mining in E-Commerce Data” was published in Imperial Journal of Interdisciplinary Research (IJIR) Vol.2, Issue-2, 2016 ISSN :2454-1362, www.onlinejournal.in

9. References

[1] IEEE paper titled 'Extracting Interesting Patterns from E-commerce Databases to Ensure Customer Loyalty


