Study of Association Rule Mining in XML Data

Balaji Ganesan (balaji*)
Richard Nikhil Martin (rnmartin*)
Satya Narayan Thiruvallur Selvakumar (satya*)

INTRODUCTION:

Most web applications today use XML as the standard for describing and storing data over the internet. This has lead to an increasing necessity for XML data mining techniques. Association Rule Mining is one of the most popular mining techniques used for traditional databases and it is natural that the methodology should be extended to XML data.

Our project aims to study various issues in the design and implementation of an Association Rule Mining system for XML data. The goals of this project are mentioned below

1. To study the basics of Association Rule Mining and its application to XML Data.
2. To evaluate the implementation of the Apriori algorithm for association rule mining using XQuery on some existing XML dataset as described in the paper **"Extracting Association Rules from XML Documents using XQuery".**
3. To explore and implement a BFS intersection set association rule mining algorithm and compare its performance with the Apriori system.

TASKS COMPLETED:

In order to achieve the above goals the following tasks needed to be completed

1. Gather literature on issues relating to Association Rule Mining especially in relation to XML data.

   We have studied various papers on different classes of ARM algorithms (BFS and counting occurrences, BFS and intersecting sets, DFS and counting occurrences, DFS and intersecting sets). It is observed that certain techniques are more suitable to relational data only and cannot be used to mine XML data because of various structural incompatibilities.

2. Compare the different approaches available and choose one suitable for implementation:

   The Apriori algorithm is by far the most popular implementation. In fact only few papers have even considered other classes of algorithms for mining XML data. We have chosen an implementation of the Apriori Algorithm which we wish to test and analyze. By doing so, we hope to gain an insight into the possibilities of

different approaches to mining XML data as well as improvements in current implementations

3. Find a suitable platform to study the current implementations of ARM algorithms and use the platform for any other implementations.

   Since XQuery is considered the standard for querying XML data, we studied an implementation of Apriori in XQuery. The usage of XQuery necessitates the need for a native XML database. In this light, we have chosen the Berkeley DB XML platform to perform our analysis. We have studied the working of BDB XML and have started testing the implementation of the Apriori algorithm on the same.

TASKS TO BE COMPLETED:

1. Implementation of a BFS intersection set algorithm for mining XML data.

   After looking through various possibilities, we have decided to implement a partition algorithm or the ARMOR Algorithm.

2. Performance evaluations:

   We intend to evaluate the performance of our implementation and compare it with the apriori system.