IJRSET 2018 Volume 5, Special-Issue-1 Pages: 41-44

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

ANALYSIS OF WEB USAGE USING APRIORI AND FP-GROWTH DATA MINING ALGORITHMS

[1] R. Manimegalai
[1] Asst.Professor,
[1] Department of Computer Science,
[1] Rathinam College of Arts and Science,
[1] Coimbatore-21

ABSTRACT: Web mining is the application of data mining techniques to discover patterns from the World Wide Web. The goal of web mining is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a Web site. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. It is an important technology for understanding user's behaviors on the web. Web server data corresponds to user logs are the collection of web server. The data collected at the web server includes IP address, Page reference, Access time and Memory usage. The main purpose of this research is the comparison of web sites service and performance of time and memory usage. The comparison can be implemented using the Apriori and FPgrowth Algorithm.

Keywords: [Apriori, Data cleaning, FP Growth, FP-tree, Web Usage Mining].

1. INTRODUCTION

The web is a huge collection of data such as distributed content creation, linking. The Web (World Wide Web) consists of information organized into Web pages containing text and graphic images. It contains hypertext links, or highlighted keywords and images that lead to related information. A collection of linked Web pages that has a common theme or focus is called a Web site. The users wants to have the effective search tools to find relevant information easily and precisely. The Web service providers want to find the way to implement the users' behaviors and the information to reduce the traffic load and design the Web site suited for the different group of users. The user or consumer needs to have some tools for business analysts .Such expecting tools or techniques as to satisfy their demands and/or solve the problems encountered on the Web. Therefore, Web mining becomes a powerful area and is taken as the research topic for this paper. Web mining is to apply data mining techniques to extract and uncover knowledge from web documents and services. A huge, widely-distributed, highly heterogeneous, semi-structured, hypertext/hypermedia, interconnected information repository. Web is a huge collection of documents plus Hyper-link information, Access and usage information. The web is not a relation as textual information and linkage structure. The data usage in web is huge and growing rapidly and the Google's usage logs are bigger than the web crawl. Data generated per day is comparable to largest conventional data warehouses. Ability to react in real-time to

usage patterns as no human in the loop. This Research work focuses on web usage mining and discovering the web usage patterns of websites performance.

APRIORI ALGORITHM:

Data Mining, also known as Knowledge Discovery in Databases(KDD), to find anomalies, correlations, patterns, and trends to predict outcomes. Apriori algorithm is a classical algorithm in data mining. It is used for mining frequent itemsets and relevant association rules. It is devised to operate on a database containing a lot of transactions, for instance, items brought by customers in a store. It is very important for effective Market Basket Analysis and it helps the customers in purchasing their items with more ease which increases the sales of the markets. It has also been used in the field of healthcare for the detection of adverse drug reactions.

Step 1: Apply minimum support to find all the frequent sets with k items in a database.

Step 2: Use the self-join rule to find the frequent sets with k+1 items with the help of frequent k-itemsets. Repeat this process from k=1 to the point when we are unable to apply the self-join rule.

This approach of extending a frequent itemset one at a time is called the “bottom up” approach. The candidate generation could be extremely slow (pairs, triplets, etc.). The candidate generation could generate duplicates depending on the implementation. The counting method iterates through all of the transactions each time. Constant items make the algorithm a lot heavier. Huge memory consumption

1.2. FP GROWTH ALGORITHM

FP-Growth is an improvement of apriori designed to eliminate some of the heavy bottlenecks in apriori. The algorithm was planned with the benefits of MapReduce taken into account, so it works well with any distributed system focused on MapReduce. 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 it's current count, and each branch represents a different association. The biggest advantage found in FP-Growth is the fact that the algorithm only needs to read the file twice, as opposed to apriori who reads it once for every iteration.

Another huge advantage is that it removes the need to calculate the pairs to be counted, which is very processing heavy, because it uses the FP-Treee. This makes it O(n) which is much faster than apriori.

The FP-Growth algorithm stores in memory a compact version of the database. The pattern growth is achieved via concatenation of the suffix pattern with the new ones generated from a conditional FP-tree. Since the frequent item set in any transaction is always encoded in the corresponding path of the frequent-pattern trees, pattern growth ensures the completeness of the result [1].Apriori is very much a horizontal, breadth-first, algorithm. The trie structure of FP-Growth provides a vertical view of the data. However, FP-Growth also adds a header table for every individual item that has support above the threshold support level [2].

2. SYSTEM ANALYSIS

2.1 EXISTING SYSTEM

The Research work for system study is for analysis for understanding the existing system.Apriori Algorithm is one of the technique for webmining.The explosive growth of the World Wide Web has proven to be a double-edged sword. While an immense amount of material is now easily accessible on the Web, locating specific information remains a difficult task.

Finding Frequent Information-People either browse or use the search service for retrieving or searching some related data .The search tools have low precision, low recall, low speed available on the web.

To Understanding the use of browsers the site and find out which is the most frequent used link and pattern of using the features available in the site.

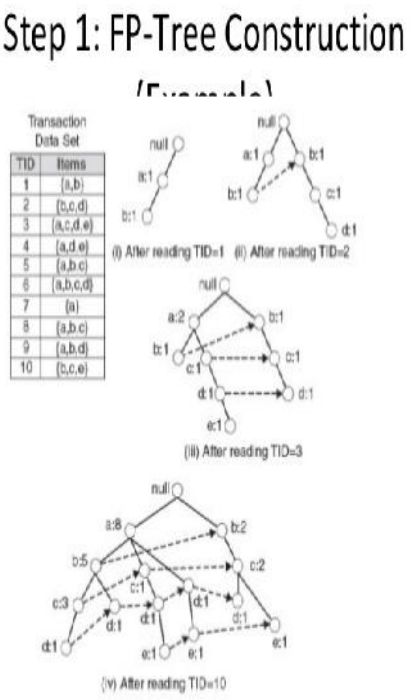
Creating new knowledge out of the information available on the web and this problem has to overcome for the collection of data and extract the discovered as well as useful knowledge of it.

2.2 LIMITATIONS OF APRIORI ALGORITHM

Apriori Algorithm is very slow and bottleneck in candidate generation. For example the transaction DB has 104 frequent 1-itemsets then generate 107 candidate 2 itemsets after employing the download closure. To compute sup more than min sup,the database scanned at everylevel,needs n+1 scans where n is the length of the longest pattern.

2.3 PROPOSED SYSTEM

The Proposed work for this paper uses a technique for mining web dataset.The Apriori and FP gorwth algorithms are used for frequent item sets for large database using association rules.In Apriori Algorithm candidate generation for large database requires significant amount of time.To overcome the drawback in Apriori, an efficient FP-tree based mining method, FP-growth, is used for this paper.FP-Growth is an improvement of apriori designed to eliminate some of the heavy bottlenecks in apriori. The algorithm was planned with the benefits of MapReduce taken into account, so it works well with any distributed system focused on MapReduce. 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 it's current count, and each branch represents a different association.



2.4 ADVANTAGES OF FP-GROWTH

FP-Growth allows frequent itemset discovery without candidate itemset generation. It is a two step approach:

Step 1: Build a compact data structure called the FP-tree

Step 2: Extracts frequent itemsets directly from the FP-tree

FP-Tree is constructed using 2 passes over the data-set:

Pass 1:

–Scan data and find support for each item.

–Discard infrequent items.

–Sort frequent items in decreasing order based on their support.

3.Comparison of Apriori vs FP-Growth

In Apriori Candidate generation is extremely slow. Runtime increases exponentially depending on the number of different items. In FP-Growth Runtime increases linearly, depending on the number of transactions and items. The comparisons between both the algorithms based on technique, memory utilization, number of scans and time consumed are given below:

Technique: In Apriori Generate singletons, pairs, triplets, etc.But in FPgrowth Insert sorted items by frequency into a pattern tree and also satisfies the minimal support.

Runtime: Candidate generation is extremely slow. Runtime increases exponentially depending on the number of different items for Apriori and Execution time more waste for candidate generation every time.In FP-Growth Runtime increases linearly, depending on the number of transactions and items

Memory Utilization: Due to Large number of candidate are generated for Apriori,so it requires large memory space.But in FP-Growth stores a compact version of the database and also it needs less memory.

Search method: Apriori use the search method as breadth first search method and FP Growth uses divide and conquer method.

No of Scans: Multiple scans for generating candidate sets for Apriori but in FP-growth scans the DB twice only.

CONCLUSION

Web usage mining is the application of data mining techniques to discover knowledge patterns from Web data. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. In this paper we have made a comparison for APriori and FP-Gowth in Webusage.Data Structure storing compressed, crucial information about frequent patterns, highly FP-tree.It is also an efficient mining methods of frequent patterns in large database. FP-Growth beats Apriori by far. It has less memory usage and less runtime. FP-Growth is more scalable because of its linear running time.

REFERENCES

[1]. B. Santhosh Kumar,K.V. Rukmani “Implementation of web usage mining using Apriori and FP growth algorithms”

[2]. J. Han, H. Cheng, D. Xin, and X. Yan, “Frequent pattern mining: Current status and future directions,”Data Mining Knowledge Discovery, vol. 15,no. 1, pp. 55–86, Aug. 2007

[3]. Mannila H, Toivonen H, Verkamo A I., "Efficient algorithms for discovering association rules." AAAI

[4]. Jeff Heaton "Comparing Dataset Characteristics that Favor theApriori, Eclat or FP-Growth Frequent Itemset Mining Algorithms

[5]. Agrawal R, Srikant R., "Fast Algorithms for Mining Association Rules", VLDB. Sep 12-15 1994, Chile, 487-99, pdf, ISBN 1-55860-153-8.

[6]. R. Srikant, “Fast algorithms for mining association rules and sequential patterns,” UNIVERSITY OF
WISCONSIN, 1996.

[7]. Agarwal, R., Aggarwal, C., and Prasad, V.V.V. 2001. A tree projection algorithm for generation of frequent itemsets. Journal of Parallel and Distributed Computing, 61:350–371,