INTERNATIONAL JOURNAL OF TECHNOLOGY AND COMPUTING (IJTC)
ISSN-2455-099X,
Volume 2, Issue 4 April 2016.

SURVEY PAPER ON FREQUENT PATTERN MINING ON WEB SERVER LOGS

Manjot Kaur [a,*], Rupinder Kaur Gurm[b]

[a] *MTech Student*, Manjotnagpal189@gmail.com, *RIMT, Mandigobindgarh, Punjab, India*
[b] *Assistant Professor*, rupindergurm@gmail.com ,*RIMT, Mandigobindgarh,Punjab,India*

ABSTRACT

At earlier the frequent patterns are prepared from web server logs using Apriori Algorithm and FP tree Algorithm. These algorithms require time and cost, as in Apriori algorithm we have to scan the database again and again in order to create frequent patterns from web server log. By scanning the data several times it will take lot of time and cost .Same in FP tree Algorithm when the tree is created, then we have to scan the tree at least two times to create frequent patterns, this also takes time and become complicated .In FP Split Algorithm to make as much frequent pattern set from web server log, we have to scan the data only once. There is no recursive scan of database and no complicacy is there in making frequent pattern sets.

Keywords: Data Mining, Web Usage Mining, Apriori Algo, FP Algo, FP Split Algo, JDK.

I. INTRODUCTION

The Web is a huge, explosive, diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information, and also raises the complexity of how to deal with the information from the different perspectives of view, users, web service providers, business analysts. The users want to have the effective search tools to find relevant information easily and precisely. The Web service providers want to find the way to predict the users' behaviors and personalize information to reduce the traffic load and design the Web site suited for the different group of users.

Characteristics of Web data

a. Massive amount of data: The data on the Web is tremendously large in size. This makes it hard to implement traditional data mining techniques on the data.

b. Distributed: The data on the Web is distributed across computers spread over diverse locations. So before mining there is a need to gather the documents together.

c. Heterogeneous: Web data comprises of diverse types of data such as textual data as well as multimedia data.

d. Semi structured: Web documents have a structure which allows for querying but that is without a query language. Hence to process the documents easily they should be represented in a correct format.

e. Dynamic: The content and the structure of Web documents always changes. Hence keeping up with the change is essential so that information retrieval is not affected.

II. VARIOUS ALGORITHM USED

Apriori Algorithm

In computer science and data mining, Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Apriori uses breadth-first search and a tree structure to count candidate item sets efficiently. After that, it scans the transaction database to determine frequent item sets among the candidates.

Limitations of Apriori Algorithm

Apriori algorithm, in spite of being simple, has some limitation. They are,

a. It is costly to handle a huge number of candidate sets. It is tedious to repeatedly scan the database and check a large set of candidates by pattern matching, which is especially true for mining long patterns.

FP Growth Algorithm

The FP Growth algorithm operates in the following four modules.

a. Preprocessing module
b. FP Tree an FP Growth Module
c. Association Rule Generation
d. Results

The preprocessing modules convert the log file, which normally is in ASCII format, into a database like format, which can be processed by the FP Growth algorithm.

The 2nd module is performed in two steps.

a. FP Tree generation

b. Applying FP Growth to generate association rules

FP tree is a compact data structure that stores important, crucial and quantitative information about frequent patterns.

Second, an FP-tree-based pattern-fragment growth mining method is developed, which starts from a frequent length-1 pattern (as an initial suffix pattern), examines only its conditional-pattern base (a "sub-database" which consists of the set of frequent items co-occurring with the suffix pattern), constructs its (conditional) FP-tree, and performs mining recursively with such a tree. The pattern growth is achieved via concatenation of the suffix pattern with the new ones generated from a conditional FP-tree.

Advantages of FP Growth Algorithm

The major advantages of FP-Growth algorithm is,

a. Uses compact data structure

b. Eliminates repeated database scan

FP-growth is an order of magnitude faster than other association mining algorithms and is also faster than tree- Researching. The algorithm reduces the total number of candidate item sets by producing a compressed version of the database in terms of an FP-tree. The FP-tree stores relevant information and allows for the efficient discovery of frequent item sets.

FP-SPLIT Tree Algoritthm

Tree construction using Fp-Split Tree Algorithm.

Step-1. Scanning the database to create equivalence class of item. Let the equivalence class of item be EC,= {Tid I Tg are the identifier of transaction ti; 1 is an item of ti).

Step-2. Calculating support to filter out non-frequent items. The support of each item I- refers to the number of records contained in the equivalence class EC,. Let pCLl denote the support of the equivalence class EC,. After calculating the supports of items, we delete the items whose supports are below the predefined minimum support

Step-3. After generating frequent items, the equivalence class of item is then converted into nodes for the construction of FP-split tree. To facility tree traversal, a header table is built in advanced so that each item can point to its first occurrence in the FP-split tree.

Step-4. While constructing the FP-split tree, a root will be generated, which is a dummy node.

Step-5. There are the four rules for the constructing of FP tree, where p stands for a specific node in the FP tree.

III. OBJECTIVES

In this research, we will emphasize on Web usage mining. Reasons are very simple: With the explosion of E-commerce, the way companies are doing businesses has been changed. E-commerce, mainly characterized by electronic transactions through Internet, has provided us a cost-efficient and effective way of doing business. The growth of some E-businesses is astonishing, considering how E-commerce has made Amazon.com become the so-called "on-line Wal-Mart". Unfortunately, to most companies, web is nothing more than a place where transactions take place. They did not realize that as millions of visitors interact daily with Web sites around the world, massive amounts of data are being generated. And they also did not realize that this information could be very precious to the company in the fields of understanding customer behavior, improving customer services and relationship, launching target marketing campaigns, measuring the success of marketing efforts, and so on.

Uses of web usage mining:

- Enhance server performance
- Improve web site navigation
- Improve system design of web applications
- Target customers for electronic commerce
- Identify potential prime advertisement locations

IV. METHODOLOGY

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered. They are web server data, application server data and application level data. Web server data correspond to the user logs that are collected at Web

server. Some of the typical data collected at a Web server include IP addresses, page references, and access time of the users and is the main input to the present Research. This Research work concentrates on web usage mining and in particular focuses on discovering the web usage patterns of websites from the server log files. The mining is done using Frequent Pattern Growth algorithm.

V. LITERATURE REVIEW

For review literature, I studied research papers of previous years published by various researchers around the globe. By these papers ideas for our research build up. I collected data for review literature from the search engines as mentioned below.

The search engines used are:

www.google.com

Comprehensive search across websites

www.wikipedia.com

Excellent source for general description of theories, models, terms with references to other sources.

Under the guidance of my guide , I built up the idea for review literature.

VI. TECHNICAL TOOLS FOR IMPLEMENTATION

Java JDK 1.7

A Java Development Kit (JDK) is a program development environment for writing Java applets and applications. It consists of a runtime environment that "sits on top" of the operating system layer as well as the tools and programming that developers need to compile, debug, and run applets and applications written in the Java language.

REFERENCES

[1] Federico Michele Facca and Pier Luca Lanzi , "*Recent Developments in Web Usage Mining Research"*

[2] JaideepSrivastava, Robert Cooley, MukundDeshpande, Pang-Ning Tan, "*Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data"*

[3] Amit Kumar Mishra, Mahendra Kumar Mishra, VivekChaturvedi, Santosh Kumar Gupta, Jaiveer Singh, "*Web Usage Mining Using Self Organized Map",* International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013, ISSN: 2277 128X

[4] RajniPamnani, PramilaChawan, "*Web Usage Mining: A Research Area in Web Mining"*

[5] DimitriosPierrakos, GE OrgiosPaliouras, Christos Papatheodorou and Constantine D. Spyropoulos, "*Web Usage Mining as a Tool for Personalization : A Survey*"

[6] Anand Sharma, "*Web Usage Mining: Data Preprocessing, Pattern Discovery and Pattern Analysis on the RIT Web Data*".