IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

Comparative Analysis of Web Usage Mining

Adarsh Gupta[1], Mukul Atawnia[2], Rohan Wadhwa[3], Shreya Mahar[4], Vinita Rohilla[5]

Dept of Computer Science and Engineering, Maharaja Surajmal Institute of Technology, New Delhi, India[1,2,3,4]

Assistant Professor, Dept of Computer Science, Maharaja Surajmal Institute of Technology, New Delhi, India[5]

Abstract: The World-Wide Web provides every internet citizen with access to a lot of information, but it is becoming increasingly difficult to identify those pieces of information that are relevant to the user. Research in areas of web mining tries to address this problem of web data by applying techniques from data mining on it. The web mining tools are generally used for scanning the HTML documents and other medias, the results are provided for the web engines. It can provide content to web engines to provide productive results in order of relevance. In this paper, we will briefly introduce concepts of web mining and then we will give overview of FP-growth and Apriori algorithm which are important usage mining algorithm and provide their comparison.

Keywords: Web Mining, FP Growth, Apriori.

I. INTRODUCTION

Main aim of web mining is to discover information or knowledge which is useful from the hyperlink structure, content of page and usage data. Web mining is an application of data mining which has become an important area of research due to vast amount of World Wide Web services in recent years. The main objective of the emerging field of web mining is finding and extracting relevant information that is hidden in Web-related data, particularly in text documents published on the Web[1][2].

Although many data mining techniques are used in web mining, it can’t be considered as purely an application of traditional data mining style. Web data can be semi-structured or unstructured in nature. Large number of new mining tasks and algorithms were invented in the past decades. Based on main kinds of data and analysis used in the mining process, Web mining tasks can be categorized into three subtypes: Web content mining, structure mining and Web usage mining.

II. WEB MINING SUB CATEGORIES

Web mining can be broadly classified into three distinct sub categories.

A. Content Mining

It is a process of extracting convenient information from the content of a web files. Content data is set of important facts and figures rooted in a webpage[3]. It consists of text content, audio, images, video etc. Applications of text mining is most important and researched area in content mining. Extraction of useful knowledge from Web page contents is known as web content mining. For example, classification and clustering of web pages can be automatically done according to their topics. These tasks are similar to those in traditional data mining techniques. However, patterns can be discovered in Web pages for extracting useful data such as details about products, postings of forums, etc, for a lot purposes[4]. Furthermore, customer reviews and forum postings can be mined to discover consumer sentiments.

B. Web Structure Mining

Web structure mining actively concentrates on hyperlink structure of the web apps and websites. The different web documents are connected in some way or the other. The existence of links amid web pages makes the web site design methodical and user friendly. The web structure mining operation serves in locating and modeling the link structure of the web site. The sitemaps of the web site is generally helpful for this motive. Finding linking structure of web docs in the Web site is a biggest challenge in process of Web Structure Mining. This structured data is discoverable by study of available web structure.

C. Web Usage Mining

Web Usage Mining is Data Mining application to analyze patterns and discover new sequence in user transactions, click streams and other user interactions log within a website[3]. This web-mining category extracts usage pattern by exercising data mining techniques to server logs or tracking histories, along with employing some content discovery for the web site structure. Important steps in Web Usage Mining are: (1) Preprocessing, (2) Knowledge discovery (3) pattern analysis. Web Usage mining is materializing web activities into an important procedure for making more user

DOI10.17148/IJARCCE.2017.6461

approachable, custom made and business-first intelligent web services. Pre-processing ,data mining and demonstrating techniques when applied to the web sources, have already provided many effective applications for information systems, personally tailored pages for individual user characteristics and preferences , smart analytic tools designing and content management systems. The way the interaction between Users and Web is exponentially increasing, the need for smarter web usage analytic tools are also continuing to grow. As the complexity of interaction between users and web applications increase with these applications, the need for analysis of usage patterns also grows continuously.

III. APPLICATIONS OF WEB MINING

The main objective of web usage mining is to study user navigation and their usage of web resources. There are many applications of web usage mining in different areas, and these applications are:

- Personalization of web content

The web usage mining techniques can be applied to personalize websites, on the basis of user profile and behavior. Personalization is important to build acceptable marketing strategies, creating a deeper relationship, and to automate the promotion of products for potential customers.

- System improvement

The results given by web usage mining techniques is useful for improving the performance of web based applications and web servers. Policies and strategies can be produced for web caching, network transmission, load balancing and data distribution by understanding the behavior of web traffic.

- Security

Web usage mining can provide patterns that are useful in detecting intrusion, fraud, attempted break-ins etc.

- Site design support

Usability is one of the most important designing issues in implementation of websites. Designers are given information about user behaviors by web usage mining results that help in decisions about any redesign of the content and structure of the website.

- Enhance e-learning environment

Web Usage mining tools can be used to track the activities happening within the course’s website, and then extract patterns and behaviors that need to be updated, improved or adapted to the course contents.

IV. FP GROWTH ALGORITHM

FP-tree is a type of extended prefix-tree structure storing quantitative and crucial information about frequently occurring patterns. For ensuring that compactness and informative nature of tree structure, only frequent-length-1 no of items have nodes in the tree, and the arranged of nodes is in such a way that more often occurring nodes have a better chance of node sharing than the less frequently occurring ones. Experiments have shown that such a tree designing procedure generate compact tree, and sometimes orders of magnitude is smaller than the order of original database. Subsequent pattern mining runs will only have to work with generated FP-tree rather than the whole database.

The operation of FP Growth algorithm can be divided into following four modules.
Preprocessing module
FP Growth Module for generating FP-Tree
Association Rule Generating module
Result set generator[3][5].

Preprocessing Module
This module helps to convert the log file which is in ASCII format into a database like format, which can be processed by the FP-Growth algorithm[3][5].

FP Growth Module for generating FP-Tree
This is two steps process.
Generation of FP Tree
FP Growth algorithm for generating association rules

Algorithm[2][3]
Input: A database Db, FP-tree and ‘n’ minimum support threshold.
Output: Frequent pattern set.
Method: call FP(tree, null).
Procedure FP(FP-tree, n) {

DOI10.17148/IJARCCE.2017.6461

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

```
1)  if single prefix path is contained in FP-tree Mining single prefix-path in FP-tree {
2)  let FP-trees single prefix-path part  be P-path ;
3)  let the multipath path be Q-path with by a null root replacing top branching node ;
4)  for each combination of the nodes (denoted as B) in the P-path do
5)  generate pattern set B ∪ n with support >=minimum support of nodes in B;
6 )let frequent pattern set P-path be the set of generated patterns;
}
7)  else let Q-path be FP-Tree;
8)  for each item i present in Q-path do { /*Mining multipath FP-tree*/
9)  generate pattern B = i ∪ n with support = i .support;
10) construct B’s conditional pattern-base and then B’s conditional
FP-tree FP-tree B;
11) if FP-Tree B ≠ Ø then
12)call FP(FP-tree B , B);
13)let set of generated patterns Q-path;
}
14)return(frequent pattern set(P-path) ∪ frequent pattern set(Q-path) ∪ (frequent pattern set(P-path) × frequent pattern
    set(Q-path)))
```

Advantages

1. Compact data structure is used.
2. Repeated database scan is eliminated.
3. Less time consuming than Apriori algorithm.
4. By generating compressed version of dataset total no of candidate itemset are reduced

Disadvantages

1. Recursive calls requires more time .
2. Good only for common access paths
3. More memory is utilised
4. Building an FP-Tree is expensive

V. APRIORI ALGORITHM

Apriori algorithm is used for finding frequent itemsets and association rule learning over databases. In initial database passes it confines large data sets and this result is used as the base for finding large patterns in other sets during sequent passes It starts by identifying each frequently occurring data item in the dataset and if occurring large enough time in database they are extended to larger patterns. The frequent itemsets that are determined are used to determine association rules present in database highlighting general trends . Frequent itemsets are itemsets having support level above the minimum The base of this algorithm is large itemset property which states: “Any subset of a large itemset is large and any subset of frequent itemset must be frequent”. Apriori was traditionally known as AIS algorithm.

Apriori Algorithm

```
L1=frequent_1-itemsets(Db); for(n=2; Ln-1≠Φ; n++)
{
Cn=gen_apriori(Ln-1, min_support);
for each transaction tr@Db
{
Ctr=subset(Cn,tr);
for each candidate can@Ctr can.count++ ;
}
Ln={can@Cn |can.count≥min_support }
}
Ans=UnLn ;
Function gen_apriori(Ln-1:freq(n-1)-itemsets)
for each itemset I1 @Ln-1
{
for each itemset I2 @ Ln-1
{
```

```
if(I1 [1]= I2 [1])∧ (I1 [2]= I2 [2]) ∧ …∧ (I1 [n-2]= I2 [n-2]) ∧ (I1) [n-1]< I2 [n-1]) then
{
can=I1 I2;
if infreq_subset(can, Ln-1) then delete cna;
else add can to Cn ;
}
}
}
return Cn;
Procedure infreq_subset(can: candidate k-itemset; Ln-1:freq(n-1)-itemsets)
for each(n-1)-subset sub of can
{
if sub @Ln-1 then return true;
}
return false; where Db=database
min_support=minimum support defined by user
```

Apriori algorithm suffers from many inefficiencies and trade-offs, which have given rise to other algorithms overcoming these shortcomings.

Advantages
Simple and easy algorithm.
Implementation is easy.

Disadvantages
Multiple scans over the database are done to generate the candidate set.

The max length of frequent of item set is equal to the number of passes of the database.

Table 1: Features of Algorithm

Algorithm	Features
FP-Growth	● Compact data structures are used. ● Extended FP Tree structure implements efficient and scalable method for mining complete set of frequent pattern by pattern fragment growth. ● Repeated database scan are removed. ● Faster than Apriori algorithm. ● Log files which are generally in ASCII format are converted into a database supported format by preprocessing modules.
Apriori	● Frequent itemsets are mined. ● Breadth-first search and a Hash tree structure are used to count candidate item sets efficiently. ● Boolean association rule frequent itemsets are mined .

Table 2: Comparison of Algorithms

Parameters	Apriori	FP-Growth
Technique	Use apriori property and join prune property	Conditional frequent pattern tree and conditional pattern base from database are constructs satisfying minimum support
Memory Utilization	Large number of candidates are generated so require large memory space is used	As no candidate generation require so less memory is required
Number of scans	Multiple scans used to re-generate candidate sets	Database is scanned twice only.
Time	As each candidate is constructed every time execution time is increased.	Execution time is less as compared to Apriori algorithm

DOI10.17148/IJARCCE.2017.6461

VI. CONCLUSION

Web mining is one of the important field in data science. Web usage patterns are discovered. Need to analyze web data will increase with increasing web usage. Apriori algorithm and FP Growth satisfies various needs of web service providers and users, business analysts, etc as it provides the useful and relevant data to them. It improves techniques of Web Usage Mining by first working on the usage log of individual users and then matching with data of other users activity patterns. Hence the result set in improved .This combined knowledge can be used to design more effective business strategies which can boom revenue. Apriori algorithms main disadvantage is it using expensive candidate set generation process and for FP-growth algorithm is the lacks a good method for candidate generation . In future these algorithms have scope in web content mining. In this paper authors have made an attempt to discuss FP-Growth and Apriori algorithms, and advantages and disadvantages.

REFERENCES

[1] N. Mago, "Web Mining: Intelligent way of mining Web based data", Apeejay Journal of Computer Science and Applications, vol. 3, pp. 11-18,2015
[2] G. Bharamagoudar, "Literature Survey on Web Mining", IOSR Journal of Computer Engineering, vol. 5, no. 4, pp. 31-36, 2012.
[3] B. Skaria, D. Eldhose T John and P. Shajan, "Literature Review on Web Mining", BIJDM, vol. 6, no. 1, pp. 04-06, 2016. DOI: 10.9756/bijdm.8127.
[4] K.Dharmarajan-Scholar and D Dorairangaswamy, "CURRENT LITERATURE REVIEW - WEB MINING", Elysium Journal of Engineering Research and Management, vol. 1, no.1, pp 38-42, 2014.
[5] Santhosh Kumar B & K.V.Rukmani, “Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms”, International Journal.of Advanced Networking and Applications Volume:01, Issue:06, pp: 400- 404, 2010.
[6] Kargupta, Hillol et al. Next Generation Of Data Mining. 1st ed. Boca Raton: Chapman & Hall/CRC, 2009.
[7] Han, Jiawei and Micheline Kamber. Data Mining. 1st ed. San Francisco: Morgan Kaufmann Publishers, 2001.
[8] W. Chu and T. Lin, Foundations and advances in data mining, 1st ed. Berlin: Springer, 2005, pp. 275-307.
[9] J Vellingiri, and S.Chenthur Pandian,“A Survey on Web Usage Mining,” in Global Journals Inc. (USA), Vol 1, Issue 4 Version 1.0, March, 2011
[10] Aanum Shaikh, “Web Usage Mining Using Apriori and FP Growth Alogrithm”, Aanum Shaikh/ (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6, Issue 1, pp: 354-357, 2015
[11] [11]Clifton, Christopher, “Encyclopædia Britannica: Definition of Data Mining”, Retrieved 2016-10-10.
[12] T. Hastie, R. Tibshirani and J. Friedman, The elements of statistical learning, 1st ed. New York: Springer, 2001.
[13] "Web Structure Mining", Web-datamining.net, 2016. [Online]. Available: http://www.web-datamining.net/structure/.
[14] [14]G. Bharamagoudar, "Literature Survey on Web Mining", IOSR Journal of Computer Engineering, vol. 5, no. 4, pp. 31-36, 2012. DOI:10.9790/0661-0543136

DOI10.17148/IJARCCE.2017.6461