Volume 5, Issue 3, March 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Web Usage Mining in Online Social Network

Tamanna Garg
M.Tech., Software Engineering
Department of Computer Science & Engineering
U.I.E.T, Kurukshetra University
Kurukshetra, Haryana, India

Dr. Sanjeev Dhawan
Faculty of Computer Science & Engineering
Department of Computer Science & Engineering
U.I.E.T, Kurukshetra University
Kurukshetra, Haryana, India

Abstract— The web content in present scenario is mainly comprised of* Social media systems such as blogs, photo and link sharing sites and on-line forums. . Web Usage Mining is the application of data mining techniques in the field of social networks to discover exciting usage patterns from SNS data and to serve the needs of SNS applications in a better manner. The major use of web usage mining techniques has been confined to web logs. But we use the same techniques to discover the hidden relationships between different nodes within a network or across networks. The most commonly used Apriori algorithm had a major disadvantage of performing multiple database scans for candidate set generation. The FP Tree structure solved this problem by restricting the database scans to two times. But the FP growth algorithm was very complicated and time consuming since it recursively created trees at every step during frequent item-set generation. The FP-Split algorithm further improved candidate set generation by doing a single scan of the date for candidate generation. The Apriori growth algorithm when used for mining the FP Tree performed better in frequent item-set generation. So we now propose a hybrid technique for web usage mining using FP Split Tree and Apriori Growth algorithm in the field of social network mining to perform analysis of user behavior on these sites. This can be advantageous while creating communities within a network, defining trust relationships, user behavior analysis and for friend suggestions.

Keywords—Apriori Growth, FP Split, SNS, frequent patterns.

I. INTRODUCTION

Social media systems such as blogs, photo and link sharing sites, wikis and on-line forums approximately produce up to one third of Web content in present scenario. Web Usage Mining is the application of data mining techniques to discover exciting usage patterns from SNS data, in order to understand and serve the needs of Web-based SNS application clients and the application designers in a better manner. Usage data helps to capture the identity or origin of the SNS users along with their browsing behavior on the SNS Web site. Web usage mining itself can be classified further depending on the kind of usage data considered. This research work concentrates on web usage mining and in particular focuses on discovering the web usage patterns of SNS websites from the client interaction logs.

Social Network Mining Process [3]

The following steps are included in web usage mining:

a) **Data Collection**: It is the very first step of Web usage mining. Data collection includes collecting data from various data sources which can be a single source or more than one source. In the latter case, it may be required to fuse multiple data files into a single file and then perform mining. In order to ensure authenticity and integrity of data, scientific and advanced techniques should be used for data collection.

b) **Preprocessing**: Data needs to be treated before mining as data has to be procured from different sources (user communities, blogs, likes, etc.) . Mining directly from raw data would lead to deceptive or incomplete results. Hence the following preprocessing methods should be adopted to remove noise and inconsistency:

   i. **Data aggregation**: The main purpose of this step is to collect relevant data from database. Usually user requests to the web server contain many useless entries.

   ii. **Knowledge discovery**: Pattern discovery includes discovering knowledge by applying methods such as clustering, classification, association rule etc. to the SNS data. Various methods that can be applied to web data in order to find useful patterns are:

   iii. **Statistical analysis**: This is the most commonly used method in discovering knowledge about web users. Presently, there are many traffic analysis tools which generate a report depicting statistics such as mean length of path accessed, mean time of page viewing, most frequently accessed pages etc.

   iv. **Association rules**: This method can be used to find group of pages which are frequently accessed together with support exceeding a threshold. It is not necessary that the pages are connected directly. For example, Apriori algorithm can be used to find relation between users who access a faculty page of a college and those who access a syllabus download page. It is not only applicable to marketing as well as business domains but also is very useful for web designers and administrators in deciding the layout of a website and its contents based on the correlations generated.

v. **Classification**: It is mapping a page to a set of predefined groups or classes. This is applicable in web domain as one may be interested in class or category of users having stated characteristics. Various algorithms that can be used for classification include decision tree, naïve Bayesian, k-nearest neighbor algorithm etc. For example, rules of the form 40% of the users who placed an order for iphone online are youngsters between the age group of 20-40 years with majority living in United States of America.

vi. **Clustering**: It is grouping together a set of items having similar features. Two types of clusters can be found in a web usage mining: user clusters and page clusters. User clusters will discover users having same browsing patterns whereas page clusters will discover pages possessing similar content.

vii. **Sequential patterns**: It is same as association rule with the difference of time ordering. It finds patterns such that one or set of pages are accessed after the other set but in a time sequence. Its application is the prediction of future visitors so as to target advertising on a group of users.

c) Pattern analysis: It includes filtering uninteresting patterns and to present the interesting ones to the user in a human understandable format such as graphs, table, pie charts, reports, rules generated from discovered patterns .

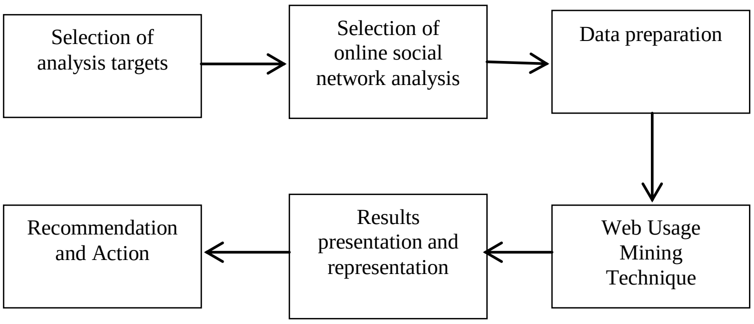
Selection of analysis targets
Selection of online social network analysis
Data preparation
Web Usage Mining Technique
Results presentation and representation
Recommendation and Action

Figure-1: Complete Process of Web Usage Mining

TECHNIQUES FOR FREQUENT PATTERNS MINING

APRIORI Algorithm

- Apriori principle: Apriori uses breadth-first search and a tree structure to count candidate item sets in an efficient manner. It generates candidate item sets of length k from item sets of length k-1. Then it prunes the candidates which have an infrequent sub pattern.

Apriori Principle

- If an itemset is frequent, then all of its subsets must also be frequent
- Apriori principle holds due to the following property of the support measure:
- Support of an itemset never exceeds the support of its subsets
- This is known as the anti-monotone property of support

Limitations:

Apriori algorithm, in spite of being simple, suffers from certain limitations.

A. It is costly to handle a huge number of candidate sets. The number of candidates to be generated increases exponentially with increasing n-itemset. This is the inherent cost of candidate generation, no matter what implementation technique has been applied.

B. 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 is an alternative way to find frequent itemsets without using candidate generations, thus improving performance.
- The core of this method is the usage of a special data structure named frequent-pattern tree (FP-tree), which retains the itemset association information.
- The FP Growth algorithm operates in the following four modules:

❑ Preprocessing module
❑ FP Tree an FP Growth Module
❑ Association Rule Generation
❑ Results

➢ 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.

Pass 2:

Nodes correspond to items and have a counter

1. FP-Growth reads 1 transaction at a time and maps it to a path
2. Fixed order is used, so paths can overlap when transactions share items (when they have the same prefix).

- In this case, counters are incremented
- Pointers are maintained between nodes containing the same item, creating singly linked lists (dotted lines)
- The more paths that overlap, the higher the compression. FP-tree may fit in memory.
- Frequent itemsets extracted from the FP-Tree.

Inspite of being time efficient, this algorithm suffers from following limitations.

- It scans the entire database twice for FP Tree construction.
- The FP Tree construction algorithm can take a lot of time.

FP Split algorithm

The FP-split algorithm comes in three phases.

- Phase l: Scanning the database to create equivalence class of item.
- Phase 2: Calculating support to filter out non-frequent items.
- Phase 3: Constructing the FP-split tree using equivalence classes of frequent items. In the construction phase of FP-split algorithm with no more database scan, it mainly deals with the intersection and difference of itemsets for fast operations in the memory, and then efficiently deals with the linkage maintenance. Therefore, a lot of time can be saved for successful constructing the FP-split tree.

The FP-split algorithm excels the FP-tree algorithm in three ways:

- Only once scanning of the database.
- No filtering and sorting of each item of the transaction.
- No repeatedly searching the header table for maintaining links, while inserting a new node into tree.

APRIORI Growth algorithm

The Apriori-Growth mainly includes two steps.

- First, the data set is scanned one time to find out the frequent 1 itemsets, and then the data set is scanned again to build an FP-tree.
- At last, the built FP-tree is mined by Apriori-Growth instead of FP-Growth.

Benefits of Apriori Growth algorithm

- ✓ Apriori-Growth works much faster than Apriori. It uses a different method FP tree Calculate to calculate the support of candidate itemsets.
- ✓ Second, Apriori-Growth works almost as fast as FP-Growth. But it consumes less memory than FP Growth because it doesn’t need to generate conditional pattern bases and build sub-conditional pattern tree recursively.

II. LITERATURE REVIEW

Initial research of 1994, Fast Algorithms for Mining Association Rules by Rakesh Agrawal *et al.*[1] considered the problem of discovering association rules between the items in a large database of sales transactions. The author presented two new algorithms for solving this problem that were fundamentally different from the known algorithms. The research also showed how only the best features of the two proposed algorithms could be combined to create a hybrid algorithm which they named as Apriori Hybrid. Experiments showed that Apriori Hybrid scaled linearly with the no. of transactions and it also had excellent scale-up properties w.r.t. the no. of items in the database and the transaction size. According to Hsein Ting[2], the author studied the issues around using web mining techniques for analysis of on-line social networks. The author introduced and reviewed techniques and concepts of web mining and social networks analysis along with a discussion about how web mining techniques could be used for on-line social networks analysis. In addition this paper initiated a process to use web mining for on-line social networks analysis, which could be treated as a general process in this research area. Discussions of the challenges and future research were also included. Anupam Joshi *et al.*[3], described recent work on building systems that analyzed these emerging social media systems to recognize spam blogs, find opinions on topics, identified communities of interest, derived trust relationships, and detected influential bloggers. The authors described some initial results from ongoing work that was focused on extracting, and exploiting this structural information. We note that there is a lack of adequate data sets to fully test the new approaches.

Another research, Discovery of Interesting Association Rules Based on Web Usage Mining by Huiping Peng[6] that was published in 2010 suggested that in web usage mining, mining of association rules is an important topic. The purpose of this paper is to discover how to dig association rules effectively from the Web Log files after been obviously preprocessed. Firstly, using the FP-growth algorithm for processing the web log files and obtaining a set of frequent access patterns.Then using the combination of site topology and browse interestingness for web mining, hence discovering a whole new pattern to provide valuable data for the site’s construction. As stated by Maja Dimitri *et al.*[7] their research Association rules for improving website effectiveness: Case Analysis published in 2013, association rule mining of the web usage log files could be used to extract patterns of a website users’ behavior. This knowledge can then be used for enhanced web marketing strategies and improved web browsing experience. They used lift and confidence as the association rule interestingness measures. Drawback: They have left studies employing more sophisticated pruning techniques, which might exploit in more detail the interconnectedness of the web pages, as well as applying other association rule interestingness measures.

Ruben *et al.* [10], created a 'BRINCA' project to support a set of analyses of the social networks from this particular INCT(created by Brazilian Department of Science and Technology). These analyses were created by mining curricular and publications bases, and identifying different types of scientific relationships and areas. The authors were able to observe, for instance, how the interaction was amongst researchers from related areas, which researchers were more collaborative and which ones were isolated from the network. These analyzes were used by the INCT coordination to understand and acted to improve scientific collaboration. Chin Feng Lee *et al.* [11] proposed a fast algorithm called frequent pattern split, simply FP-split, for improving the process of the FP-tree construction. The proposed FP-split algorithm contained two main steps. The first step was to scan transaction database only once for generating equivalence classes of frequent items. The second step was to sort these equivalence classes of frequent items in descending order so as to construct the FP-split tree. Through detailed experimental evaluations under various system conditions, this method showed excellent performance in terms of execution efficiency and scalability. The FP-split algorithm was superior to FP-tree construction algorithm. There were three reasons to support that the proposed method out performed FP-tree construction algorithm in terms of tree construction. The first one was that this method scanned the database only once. The second one was that filtering out and sorting the items in each transaction record was no longer employed in this method. The third one was that the header table and links were not repeatedly searched, while adding a new node in the FP-split tree. Drawback: It needs to be optimized for counting the support of the candidates and expanded for mining more larger database.

This section summarizes the invaluable researches done by different researchers in the field of frequent pattern mining and association mining. The algorithms in these works mainly use or modify the apriori and frequent pattern growth algorithm for the purpose. The FP Split tree [32] explains a novel technique which improves the performance of FP Tree which has long been considered the best prevailing technique for frequent pattern mining. The different modifications of apriori suggested by different authors highlight the shortfalls in these basic algorithms.

III. PROPOSED METHODOLOGY

This Research work concentrates on web usage mining and in particular focuses on discovering the web usage patterns of SNS websites from the server database. The comparison of memory usage and time usage will be compared using FP split tree and Apriori growth algorithm. The previous works in web usage mining in Apriori algorithm and FP Tree mining algorithm had their disadvantages. Here we create a hybrid of FP-split tree and Apriori growth mining algorithm to take advantage of positives of both schemes. The Apriori algorithm performs repeated scans of the database while generating candidates while the FP tree mining algorithm is a time consuming, complicated algorithm since it recursively creates trees at every step during frequent item-set generation. So we create a hybrid by combining FP Split tree for candidate generation and Apriori growth for mining.

IV. CONCLUSION

Web content mining on social networks means to categorize or classify documents on an on-line social networking website, especially articles on blogs or text forums. This research discusses differenttechniques for content mining on social networks. The two main methods in this context- Apriori and FP Tree method are the traditional methods. The most commonly used Apriori algorithm had a major disadvantage of performing multiple database scans for candidate set generation. The FP Tree structure solved this problem by restricting the database scans to two times. But the FP growth algorithm was very complicated and time consuming since it recursively created trees at every step during frequent item-set generation. The FP-Split algorithm further improved candidate set generation by doing a single scan of the date for candidate generation. The Apriori growth algorithm when used for mining the FP Tree performed better in frequent item-set generation. So we now propose a hybrid technique for web usage mining using FP Split Tree and Apriori Growth algorithm in the field of social network mining to perform analysis of user behavior on these sites

REFERENCES

[1] RakeshAgrawal, RamakrishnanSrikant, "*Fast Algorithms for Mining Association Rules*", Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994

[2] I-Hsien Ting, "*Web Mining Techniques for On-line Social Networks Analysis*", IEEE 2008. Pp. 696-700.

[3] Anupam Joshi, Tim Finin, Akshay Java, Anubhav Kale, and PranamKolari, "*Web (2.0) Mining: Analyzing Social Media*", IEEE 2008.

[4] K. R. Suneetha, Dr. R. Krishnamoorthi, "*Identifying User Behavior by Analyzing Web Server Access Log File*", IJCSNS International Journal of Computer Science and Network Security. VOL.9 No.4, April 2009

[5] Mehdi Heydari, Raed Ali Helal, Khairil Imran Ghauth, "*A Graph-Based Web Usage Mining Method Considering Client Side Data*", 2009 International Conference on Electrical Engineering and Informatics, 2009.

[6] HuipingPeng, "*Discovery of Interesting Association Rules Based on Web Usage Mining*", 2010 - International Conference on Multimedia Communications.

[7] Maja Dimitrijevic, TanjaKrunic, "*Association rules for improving website effectiveness: case analysis*", Online Journal of Applied Knowledge Management Volume 1, Issue 2, 2013

[8] Kirti S. Patil, Sandip S. Patil, "*Sequential Pattern Mining Using Apriori Algorithm & Frequent Pattern Tree Algorithm*", IOSR Journal of Engineering (IOSRJEN) Vol. 3, Issue 1 (Jan. 2013), Pp. 26-30

[9] Shipra Khare, Prof Vivek Jain, Prof Manoj Ramaiya, "*Implementation of Web Usage Mining with Customized Web Log Using FP Growth Algorithms*", International Journal of Engineering & Managerial Innovations (IJEMI) Volume I (II), September 2013.

[10] Ruben P. Albuquerque, Jonice Oliveira, Fabrício F. Faria, Rafael Monclar and Jano M. de Souza, "*Studying Group Dynamics through Social Networks Analysis in a Medical Community*"**,** Journal of Social Networking, 2014, Vol. 3, Pp. 134-141.

[11] Chin Fewng Lee and Tsung-HsienShen, "*An FP-split method for fast association rules mining*", IEEE 2005. Pp. 459-464.