Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 1, January 2014, pg.255 – 261

SURVEY ARTICLE

Surveillance on Bigdata to Mine Pattern

N. Monica[1], Dr. K. Ramesh Kumar[2], T. Nelson Gnanaraj[3]

[1]MTech Scholar, Department of Information and Technology
Hindustan University, Chennai, TamilNadu, India
monitech26@gmail.com
[2]Associate Professor, Department of Information and Technology
Hindustan University, Chennai, TamilNadu, India
rameshkumar.dr@gmail.com
[3]Mtech Scholar, Department of Information and Technology
Hindustan University, Chennai, TamilNadu, India
nelsonraj.27@gmail.com

Abstract

Big data is a collection of large amount of data with various types of data and usable to be processed at much higher frequency. One of the most popular knowledge discovery approaches is to find frequent items from a transaction data set and derive association rules. Pattern finding is one of the most computationally expensive steps in large data sets. Patterns often referred to association rules. Association rule plays an important role in the process of mining data for sequential pattern. Association rules are used to acquire interesting rules from large collections of data which expresses an association between items or sets of items. Apriori is a classic algorithm for learning association rules. It is designed to operate on databases containing transactions. Apriori algorithm attempts to find subsets which are common to at least a minimum number C of the item sets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time and groups of candidates are tested against the data. The algorithm terminates when no further Successful extensions are found. In this paper we enhance Apriori algorithm to solve its complexity over large data sets. We first collect variety of data and then integrate both structured and unstructured data using MapReduce to find out sequential pattern from the required data sets.

Keywords: ***MapReduce; Apriori; Association Rule; Pattern mining; Variety of data***

1. INTRODUCTION

Knowledge discovery is the computer assisted processes which are used to analyze large sets of data and to extract the meaning of those data. Knowledge discovery uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. It is the practice to discover hidden patterns and unexpected trends in the data using a combination of techniques from machine learning statistics and database technologies automatically from very large data storage. It involves in cleaning and integrating data from data sources like databases, flat files, pre-treatment of selecting and transferring target data, mining the required knowledge and finally evaluate and present the knowledge [1].

1.2 ASSOCIATION RULE

Association is a knowledge discovery function that discovers the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rules. Association rules are created by analyzing data for frequent if or then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of items if or then statements have been found to be true. An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. In knowledge discovery, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout.

1.3 FREQUENT PATTERN MINING

Frequent pattern mining is an essential step in the process of association rule mining and has been a focused theme in knowledge discovery research for over a decade. Frequent patterns are item sets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold. For example, a subsequence, such as buying first a PC, then a digital camera and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern. A substructure can refer to different structural forms, such as subgraphs, subtree or sublattices which may be combined with itemsets or subsequences. If a substructure occurs frequently in a graph database, it is called a (frequent) structural pattern. Finding frequent patterns plays an essential role in mining associations, correlations and many other interesting relationships among data. Moreover, it helps in data indexing, classification, clustering and other knowledge discovery task as well. Thus, frequent pattern mining has become an important knowledge discovery task and a focused theme in database community.

1.4 APRIORI

In knowledge discovery, Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions. It is proceeds by indentifying the frequent individual items in the database and extending them to larger item sets as long as those item sets appear sufficiently often in the database. In the given sets of itemsets, the algorithm attempts to find subsets which are common to at least a minimum number C of the itemsets. Apriori uses a bottom up approach, where frequent subsets are extended one at a time this step known as Candidate generation in which group of candidates are tested against the data. Apriori algorithm terminates when no further successful extensions are found.

1.4.1 PSEUDOCODE FOR APRIORI ALGORITHM

```
Join Step: $C_k$ is generated by joining $L_{k-1}$ with itself
Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

$C_k$: Candidate itemset of size k
$L_k$: frequent itemset of size k
L1= {frequent items};
for(k=1; $L_k$!=0; k++) do begin
    $C_{k+1}$= candidates generated from $L_k$;
    for each transaction T in database do
        increment count of all candidates in $C_{k+1}$ that are contained in T
    $L_{k+1}$= candidates in $C_{k+1}$ with min_support
end
return $U_kL_k$;
```

2. BIG DATA

Big data is a popular term which are describe the exponential growth, availability and use of information both structured and unstructured data. It enables organizations to store, manage and manipulate vast amounts of disparate data at the right speed at the right time.

Big data analytics is an important tool to improve efficiency and quality in an organization. Sampling and distributed system are two main strategies for dealing with big data. Sampling is used when the data set is too large which could obtain an approximate solution in a subset. A good sampling method will try to select the best instance to perform good using a small quantity of memory and time. The most popular distributed system used nowadays are based on the map-reduce framework.

2.1 MAPREDUCE

MapReduce is a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand alone computers.
The framework is divided as following:

- ❖ Map: A function that parcels out work to different nodes in the distributed cluster
- ❖ Reduce: A function that collects the work and resolves the results into a single value.
- ❖ MapReduce framework is fault-tolerant because each node in the cluster is expected to report back periodically with completed work and status updates. If a node remains silent for longer than the expected interval, a master node makes note and re-assigns the work to other nodes.

2.2 VARIETY OF DATA

Big data incorporates varieties of data. Data today comes in all types of form traditional databases to hierarchical data stores created by end users and OLAP system, to text documents, email, meter-collected data, video, audio and financial transactions.

2.2.1 STRUCTURED DATA:

Any data stored in a well defined non propriety system. Data is primarily text based. Typically conforms to ACID property.
Example: Database, Data Warehouses, Electronic Spreadsheets

2.2.2 SEMI STRUCTURED DATA:

Any data stored in a system that conforms to some rules and can be propriety. Data is primarily text based which does not have to conform to ACID property.
Example: Web Posts, Blogs, Wiki pages, Forums, Tweets, Instant Messages

2.2.3 UNSTRUCTURED:

Any data stored in a well defined propriety system. Binary data that conforms mostly to an agreed standard.
Example: PowerPoint, Word Documents, Email, PDF, Audio files, Video, Graphics and Multimedia

.

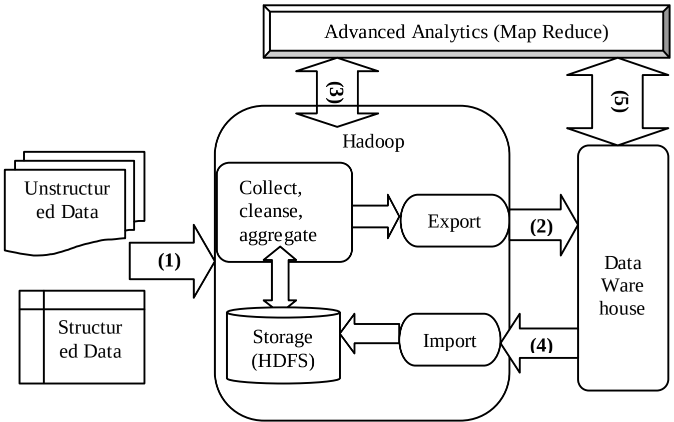


Figure 1: Integrate Structured and Unstructured Data

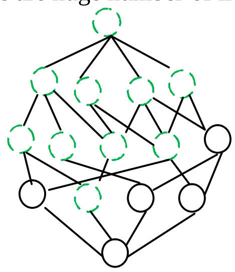
3. PROBLEM DESCRIPTION

Frequent Pattern Mining is an important knowledge discovery task and it has been a focus theme in knowledge discovery research. Frequent Pattern Mining over large database is fundamental to many knowledge discovery applications. One of the main issues in Frequent Pattern Mining is Sequential Pattern Mining retrieved the relationships among objects in sequential dataset[2]. AprioriAll is a typical algorithm to solve the problem in Sequential Pattern Mining but its complexity is so high and it is difficult to apply in large datasets. Recently, to overcome the technical difficulty, there are a lot of researches on new approaches as follow

- Custom built Apriori algorithm
- Modified Apriori algorithm
- Frequent Pattern-tree and its development
- Integrating Genetic algorithms
- Rough set Theory/ Dynamic Functions

3.1 PROBLEMS OF FREQUENT PATTERN MINING

In frequent Pattern Mining there are huge number of frequent itemsets which are hard to analyze and most of them are similar.



	1	2	3	4	5
I1	O		O		
I2		O		O	
I3	O	O	O	O	
I4		O	O		O
I5	O	O		O	

Figure 2: Problem in Frequent Pattern Mining

4. APPROACHES FOR PATTERN MINING

Following are the approaches of pattern mining

4.1 APRIORI-BASED METHOD

The Apriori property of sequences states that, if a sequences S is not frequent, then none of the super-sequences of S can be frequent[3] . E.g.<pq> is infrequent implies that its super-sequences like <prq> and <(rp)q> would be infrequent too. The GSP algorithm finds all the length-1 candidates(using one database scan) and orders them with respect to their support ignoring ones for which support < min_sup. Then for each level (i.e., sequences of length-k), the algorithm scans database to collect support count for each candidates sequence and generates candidate length-(k+1) sequences from length-k frequent sequences using Apriori. This is repeated until no frequent sequence or no candidate can be found.

Advantages

- Uses large itemset property.
- Easy to implement.
- Easily parallelized.

Disadvantages

- Large number of candidates sets are generated.
- Multiple database scans are required.
- The system I/O cost increases due to multiple scanning of transactional database.

4.2 VERTICAL FORMAT-BASED METHOD

This is a vertical format sequential pattern mining method. SPADE first maps the sequence database to a vertical id-list database format which is a large set of items <SID (Sequence ID),EID (Event ID)>. Sequential pattern mining is performed by growing the subsequences (patterns) one item at a time by Apriori candidate generation.

Advantages

- Reduces the time of scanning candidate sets.
- Hash structure is used for storage of candidate sets.

Disadvantages

- As the minimum confidence value is increased, the time is also increased.

4.3 PATTERN GROWTH BASED METHOD

Pattern growth based method uses frequent items to recursively project sequence database into a set of smaller projected databases and grows subsequence fragments in each projected database. This process partitions both the data and the set of frequent patterns to be tested and confines each test being conducted to the corresponding smaller projected database. It first scans the database, collects the support for each item and finds the set of frequent items.

Advantages

- Decreases the system overhead.
- Reduces the running time.

Disadvantage

- Multiple database scans required.

4.4 CONSTRAINT BASED METHODS

In sequential pattern mining algorithm users can specify only min_sup as a parameter. There are two major difficulties in sequential pattern mining:

Effectiveness: The mining may return a huge number of patterns, many of which could be uninteresting to users.

Efficiency: It often takes substantial computational time and space for mining the complete set of sequential patterns in a large sequence database.

To prevent these problems, users can use constraint based sequential pattern mining for focused mining of desired patterns. Constraints could be anti-monotone, monotone, succinct, convertible or inconvertible. Anti-monotonicity means “if an item-set does not satisfy the rule constraint, then none of its supersets satisfy”. Monotonicity means “if an item-set satisfies the rule constraint, then all of its supersets satisfy”. Succinctness means “All and only those patterns guaranteed to satisfy the rule can be enumerated”. Convertible constraints are those which are not any of anti-monotonic, monotonic, succinct but can be made anti-monotonic or monotonic constraints by changing order of elements in the set. Inconvertible constraints are the ones which are not convertible.

Advantage

- Decreases the system overhead.
- Easily parallezied

Disadvantage

- The system I/O cost increases due to multiple scanning of transactional database

5. APPLICATIONS OF PATTERN MINING

5.1 MINING TRANSACTIONAL DATA

It is possible to mine sequential patterns in sequences of transactions from a store. In this case, each sequence represents the transactions from a customer at the store. From this, a sequential pattern mining could find patterns common to several customers. For example, 30% of the customers who buy milk and bread will also buy jam. This could be used for taking marketing decisions or for product recommendation on a web store.

5.2 MINING WEB LOGS

Pattern mining can be done on web logs. In this case sequences of pattern could be drawn by analyzing sequences of webpages visited by users on a website. From this data, a sequential pattern mining algorithm could use sequential pattern mining to discover sequences of web pages that are often visited by users. The website could then use these patterns to generate suggestions to the user such as recommended links.

5.3 MINING MEDICAL RECORDS

Sequential pattern mining algorithm could be used to find patterns in medical records. For example, each sequence is the medical record of a person in a hospital. Patterns could be found such as that people who took the medicine A and the medicine B and then the medicine C, will have a heart attack.

5.4 MINING EDUCATIONAL DATA

Sequential pattern mining can be used to find patterns in educational data. For example, consider that each sequence of a sequence database is the course that a student took at University. It would be possible to discover patterns such as people who took course A and B will always take the course C.

5.5 MINING STOCK MARKET

It is possible to discover pattern in stock market by applying sequential pattern mining to a sequence of events on the stock market.

5.6 MINING SOFTWARE ENGINEERING

Sequential pattern mining could be applied in software engineering to find out patterns in source code.

6. FUTURE WORK

Ideas for future work in pattern mining include:

- Applying telescoping in tree projection pattern-growth algorithm to reduce the tree size, where more than one item can be compressed into one node or edge.
- Distributed mining of sequences can provide a way to handle scalability in very large sequence databases and long sequences. In the area of web usage mining, this can be applied to mine several web logs distributed on multiple servers.
- Extending the capability of existing approaches. For example, modifying the S-Matrix to include support counts of candidates sequences that are only characteristics of the underlying application, extending the PLWAP algorithm to handle general multi itemset event sequence mining
- Using the Fibonacci sequence to partition or sample the search space may be useful for effective mining of very long sequence.

7. CONCLUSION

In this survey, we present a brief overview of pattern mining and its algorithm which are used to find patterns by integrating structured and unstructured data together using Mapreduce framework. Sequential pattern mining methods have been used to analyze this data and identify patterns. Such patterns have been used to implement efficient systems that can recommend based on previously observed patterns, help in making predictions, to improve usability of system, detect events and in general help in making strategic product decisions. We have also seen applications of pattern mining in variety of domain. Apart from this, new sequential pattern mining methods may also be developed to handle special scenarios of colossal patterns, approximate sequential patterns and other kinds of sequential patterns specific to the applications. We have given a overall coverage on this topic. Hopefully, this short overview may provide a rough outline of the recent work and given people a general view of the field. In general, we feel that as a young research field in knowledge discovery, pattern mining has achieved tremendous progress and claimed a good set of applications. However, in-depth research is still needed on several critical issues soothe the field may have its long lasting and deep impact in knowledge discovery applications.

REFERENCES

[1] Albert Bifet, Mining Big Data in Real Time. Informatica 37, 15-20 (2013).

[2] Amit Ganatra, Amit Thakkar and Cletna Chand. Sequential Pattern mining: Survey and Current Research Challenges. International Journal of Soft Computing and Engineering, Volume 2, March 2012.

[3] Aniket Mahanti and Reda Alhajj. Visual Interface for Online Watching of Frequent Itemset Generation in Apriori and Eclat, Fourth International Conference on Machine Learning and Application (2005).

[4] Anuradha Sharma, Arvind Sehwal and Harleen Puri. An Empirical Proposal towards the Algorithmic Approach and Pattern in Web Mining for Assorted Applications. International Journal of Innovative Research in Computer and Communication Engineering, Volume 1. April 2013.

[5] Aramburu, Juan Manuel Perez, Maria Jose, Rafael Berlanga and Torben Bach Pedersen, Integrating Data warehouse with Web Data: A Survey, IEEE transaction on knowledge and Data Engineering, Volume 20, July 2008.

[6] Assured Research, Essay on Big Data. June 2012.

[7] Bart Goethals. Survey on Frequent Pattern Mining.

[8] Basel Kayyali, David Knotl, Peter Crroves and Steve Van Kuiken. The Big Data

[9] Bid Data, ANew World of Opportunities. NESSI White Paper, December 2012.

[10] Big Data Strategy- Issues Paper. March 2013.

[11] C.I. Ezeife, R. Mabroukeh and Nizar. A Taxonomy of Sequential Pattern Mining Algorithms. ACM Computing Surveys, Volume 43, November 2010.

[12] Challenges and Opportunities with Big Data, A Community white paper developed by leading researchers across the United States.

[13] D. Magdalene Delighta Angeline and I. Samuel Peter James. Association Rule Generation using Apriori Mend Algorithm for Student's Placement. Int. J. Emerging. Science. 2(1). March 2012.

[14] D.P Agrawal, R.K. Gupta and A. Tiwari. A Survey on Frequent Pattern Mining: Current Status and Challenging Issues. Information Technology Journal 9(7), 2010.

[15] David Loshin, Integrating Structured and Unstructured Data. TDWI Checklist Report.

[16] DonXin, Hong Cheng, Jiawei Han and Xifeng Yan. Frequent Pattern Mining : Current Status and Future direction. Knowledge discovery Knowledge Disc, 2007.

[17] Dr.L. DE Radt, Mining patterns in Structured Data. September 2009.

[18] Ekta Garg and Meenakshi Bansal. A Survey on Improved Apriori Algorithm. IJERT, Volume 2, July 2013.

[19] Huajun Chen, Jun Ma , Xiangyu Zhang, Xiaolongyu and Yang Liu. Map Reduce- Based Pattern Finding Algorithm Applied in Motif Detection for Prescription Compatibility Network, APPT 2009.

[20] Hui Xion, Jian Pei and Yan Huang. Mining Co-Location Patterns with Rare Events from Spatial Data Sets. Geoinformatica (2006).

[21] Imran R. Mansuri and Sunita Sarawagi. Integrating Unstructured data into relational Databases.

[22] Jaiwei Han and Micheline Kamber, "Knowledge discovery Concepts and Techniques", Second Edition , Morgan Kaufmann Publishers.

[23] Jiaivei Han and Jing Gao. Research Challenges for Knowledge discovery in Science and Engineering, Chapter 1.

[24] Jiawei and Manish Gupta. Approaches for Pattern Discovery using Sequential Knowledge discovery.

[25] Jiaweri Han, JiongYang, Lei Lui and Peter Bajcsy. Survey of Biodata Analysis from a Knowledge discovery Perspective, Chapter2.

[26] Joseph Mckendrick. Big Data, Big Challenges, Big Opportunities, IOUG September 2012.

[27] Olivera Marjanouic and Sarah Anstiss. Understanding Data Quality Issues in Dynamic Organizational Environments- A Literature Review. 23rd Australasian Conference on information System.Greelong, 3-5 December 2013.

[28] PS Sastry and Srivatsan Laxman . A Survey of Temporal Knowledge discovery.Sadhana Volume 31, April 2006.

[29] Qiankum Zhao, Sourav S. Bhowmick. Sequential Pattern Mining: A Survey, Tehnical Report,2003.

[30] R. Deepalakshmi, Dr.S.P. Shantharajah and S.Suriya. A Complete Survey on Association Rule Mining with Relevance to Different Domain, International Journal of Advanced Scientific and Technical Research, Volume 1, February 2012.

[31] R.V.Kulkarni and Tejaswini Abhijit Hilage. Review of Literature on Knowledge discovery. IJRRAS 10(1), January 2012.

[32] Research Paper on A Total data management approach to Big Data of Computing Research.

[33] Research Trends Special Issue on Big Data, September 2012.

[34] Sunita Sarawagi. Information Extraction Foundation and Trends in Databases. Volume 1,2007.

[35] The Challenges of Integrating Structured and Unstructured Data. 14th PNEC Conference.

[36] U.Chandrasekhar, Sandeep Kumar.K, Yakkala Uma Mahesh. A Survey of latest Algorithms for Frequent Itemset Mining in Data Stream. International Journal of Advanced Computer Research (ISSN (Print): 2249-7277 ISSN(Online): 2277-7970) Volume-3, March 2013.