RESEARCH REVIEW JOURNALS
Volume-04
Issue-06
June-2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com[UGC Listed Journal]

# A Study on Graph Mining Mathematical Model for Big Data Analysis for Big Networks

[1]Santanu Sikdar & [2]Dr. Anil Kumar

*[1]Research Scholar, Department of Computer Science & Engineering, Sri Satya Sai University of Technology and Medical Sciences, Sehore, Madhya Pradesh (India)*
*[2]Department of Computer Science & Engineering, Sri Satya Sai University of Technology and Medical Sciences, Sehore, Madhya Pradesh (India)*

**ARTICLE DETAILS**

**Article History**
*Published Online:* 12 June 2019

**Keywords**
*Graph Mining, Data Architecture.*

**ABSTRACT**

*Data mining is comprised of many data analysis techniques. Its basic objective is to discover the hidden and useful data pattern from very large set of data. Graph mining, which has gained much attention in the last few decades, is one of the novel approaches for mining the dataset represented by graph structure. Graph mining finds its applications in various problem domains, including: bioinformatics, chemical reactions, Program flow structures, computer networks, social networks etc. Different data mining approaches are used for mining the graphbased data and performing useful analysis on these mined data. In literature various graph mining approaches have been proposed. Each of these approaches is based on either classification; clustering or decision trees data mining techniques. In this study, we present a comprehensive review of various graph mining techniques. These different graph mining techniques have been critically evalnated in this study. This evalnation is based on different parameters. In our future work, we will provide our own classification based graph mining technique which will efficiently and accurately perform mining on the graph structured data.*

## 1. Introduction

Data mining in seeking for better performance and innovation. One innovation includes mining from structured data, which is a new challenge. Since a structure is represented by proper relations and a graph can easily represent such relations, knowledge discovery from graph-structured data poses a general problem for mining from structured data. Some examples amenable to graph mining are finding typical web browsing patterns, identifying typical substructures of chemical compounds, finding typical subsequences of DNA and discovering diagnostic rules from patient history records. Graph mining techniques have been categorized into following groups. (1) Graph clustering; is the task of grouping the vertices of the graph into clusters taking into consideration the edge structure of the graph in such a way that there should be many edges within each cluster and relatively few between the clusters? Graph clustering in the sense of grouping the 978-1-4673-2430-4112/$31.00 ©2012 IEEE 88 vertices of a given input graph into clusters graph clustering is based on unsupervised learning technique in which the classes are not known in prior to clustering. The graph clusters are formed based on some similarities in the underlying graph structured data graph. (2) Graph Classification; in graph classification the main task is to classify separate, individual graphs in a graph database into two or more categories/classes. Classification is based on supervised/semi supervised learning technique in which the classes of the data are defined in prior. (3) Sub graph mining; sub graph is a graph whose vertices and edges are subsets of another graph. The frequent sub graph mining problem is to produce the set of sub graphs occurring in at least some given threshold of the given n input example graphs

## 2. Graph-Oriented Databases

Graph-arranged databases influence graph theory from mathematics. This kind of database uses concepts of nodes, edges, and properties to store information and relationships. Coordinated graphs are especially useful when thinking about complex relationships like schedules with multiple dependencies, or in a social network, where you have to store information about individuals and their connectedness. In Fig. 4.1 you see how individuals associate with each other. For example, if this were a graph of individuals associated in Face book, you can see how sending information to one person would emanate out to others.

Graph theory is the science of review mathematics models in terms of graphs to relate objects with each other. A social network picture is an example of a graph. Nodes are individuals, and the edge connects these individuals. The properties can characterize the edges, or relationships.

## 3. Big Graph Mining

Algorithms and Discoveries We can discover patterns and anomalies in enormous graphs with billions of nodes and edges with mining which are useful most productively. Enormous graphs are all over the place, extending from social networks and portable call networks to biological networks and the World Wide Web. Mining enormous graphs leads to many interesting applications including digital security, extortion discovery, Web search, suggestion, and some more. In this subject we describe PEGASUS, a major graph mining system based over MAPREDUCE, a cutting edge distributed data processing stage. We present GIM-V, a significant crude that PEGASUS uses for its algorithms to break down structures of huge graphs. We also present

HEIGEN, a huge scale Eigen solve which is also a part of PEGASUS. Both GIM-V and HEIGEN are exceptionally advanced, accomplishing linear scale up on the number of machines and edges, and giving 9.2x and 76x faster execution than their gullible counterparts, respectively. Using PEGASUS, we break down exceptionally huge, certifiable graphs with billions of nodes and edges. Our findings incorporate anomalous spikes in the associated component size distribution, the 7 degrees of separation in a Web graph, and anomalous adult advertisers in the who-follows-whom Twitter social network.

**4. Distributed Big Graph Mining**

There are several works on distributed huge graph mining which can be gathered into two:

1. One not based on MAP REDUCE/HADOOP
2. The other over it.

The works not based on MAPREDUCE/HADOOP incorporate Graph Lab, Pregel, and Trinity. Graph Lab provides a framework for parallel machine learning and data mining, in a shared memory setting. As of late, they give the distributed Graph Lab to shared nothing machines. Pregel is a system for enormous scale graph processing where vertices trade messages and change their states in memory. Trinity is a memory-based distributed database and calculation stage. When all is said in done, those systems don't coordinate the MAP REDUCE/HADOOP'S high degree of adaptation to internal failure capabilities including 3-way replication and speculative execution. On the MAP REDUCE/HADOOPO side, Apache Mahout, a scalable machine learning library on HADOOP, provides an alternate set of operations contrasted with PEGASUS. Forecasting the future, the regularly developing sizes of graphs and diverse application needs will open numerous new opportunities for interesting researches in enormous grph mining. We list five of the significant research directions.

**5. Visualization and Understanding of Graphs**

A graph forms a confounded object with numerous interactions between nodes. Visualization of graphs helps us between understand the structure and the interactions in graphs. The test is to adequately summarize the graphs to those users can easily understand the graphs in a screen with constrained resolution.

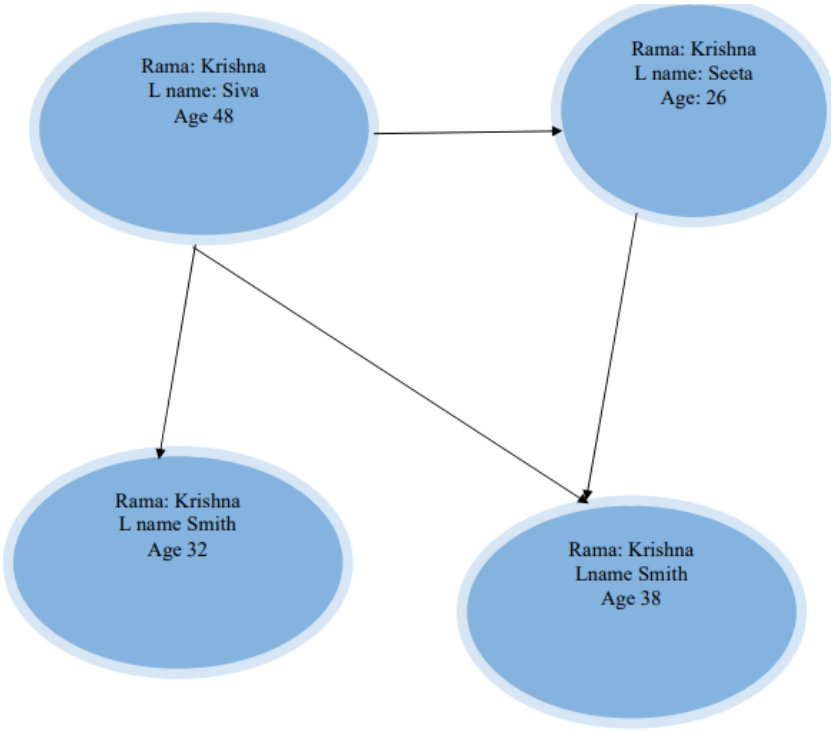


**Fig This binary relatin gives the Galois lattice**

**6. Big Data Architecture**

Enormous data is a wide descriptive term for non-transactional data tht is user produced and machine created. Data producing developed from transactional data to first collaboration data and after that sensor data. Web log was the first step in this advancement. These machines created logs of internet action caused the first development of data. Social media pushed data generation higher with human interactions. Robotized observations and wearable technologies make the following phase of huge data. Data volumes have been the primary focus of biggest data discussions. Architecture for huge data regularly focuses on storing huge volumes of data. Dollars per TB (Terabyte) becomes the measurement for architecture discussions.

We contend this is not the correct focus. Enormous data is tied in with determining value. Hence, analytics should be the goal behind investments in storing enormous volumes of data. The measurement should be dollars per analytic performed. There are three utilitarian aspects to huge data-data catch, data R&D, and data item. These three aspects must be put in a framework for creation the data architecture. We discuss every one of these aspects top to bottom in this chapter. The goal of huge data is data-driven decision

 

making. Decisions should not be made with data silos. At the point when setting is added to data items they become important. At the point when more contexts are included more insight is possible. Getting insight from data is tied in with reasoning with all data and not just huge data. We show examples of this and contend huge data architecture must give mechanisms to reason al data.

Huge data analytic requires all forms of various technologies including graph analytics, statistical analytics, way analysis, machine learning, neural networks, and statistical analysis be coordinated in an incorporated analytics environment. Enormous data architecture is architecture that professional vides the framework for reasoning with all forms of data. We end this chapter with such architecture. This chapter makes three points as follows: Big data analytics is analytics on all data and not just enormous data alone Data multifaceted nature not volume, is the primary worry of huge data analytics Measure of goodness of a major data analytic architecture is dollars per analytics and not dollars per TB.

**7. Big Data Processing Algorithms**

Information has been becoming bigger enough to understand the need to stretch out customary algorithms to scale. Since the data can't fit in memory and is distributed across machines, the algorithms should also follow the distributed storage. This chapter introduces some of the algorithms to take a shot at such distributed storage and to scale with massive data. The algorithms, called Big Data Processing Algorithms, comprise arbitrary walks, distributed hash tables, streaming, mass synchronous processing (BSP), and Map Reduce paradigms. Every one of these algorithms is unique in its approach and fits certain problems. The goal of the algorithms is to decrease network communications in the distributed network, limit the data movements, cut down synchronous delays, and upgrade computational resources. Data to be processed where it resides, peer-to-peer-based network communications, computational and amassing components for synchronization are some of the techniques being used in these algorithms to accomplish the goals. Guide "Diminish has been received in Big Data problems broadly. This chapter demonstrates how Map Reduce enables analytics to process massive data easily. This chapter also given example applications and codebase to readers to start hands-on with the algorithms.

**8. Big Data Search and Mining**

Most enterprises are producing data at an extraordinary way. Then again, customary consumers are transforming into computerized consumers because of high embracing of social media and networks by individuals. Since transactions on these sites are enormous and increasing quickly, social networks have turned into the new focus for several business applications. Huge Data mining deals with catching ale measure of data that is perplexing sixth a wide variety of data types and provides noteworthy insights at the opportune time. The search and mining applications over Big Data resulted in the improvement of another sort of technologies, platforms, and frameworks. This chapter introduces the idea of search and data mining in the Big Data setting and technologies supporting Big Data. We also present some data mining techniques that deal with scalability and heterogeneity of enormous data. We further discuss clustering social networks using topology discovery and furthermore address the issue of assessing and overseeing content based sentiments from social network media. Further, this chapter accentuates some of the open source tools for Big Data mining.

**9. Applications of Big Data**

Throughout the last couple of decades, enormous business houses in various disparate areas have been aggregating data from various departments in various formats and have been struggling to associate the datasets and settle on any important business decisions. The key stumbling b lock has been the failure of the accessible systems to process enormous data when the data are part structured and part unstructured. As witnessed in the previous chapters, the innovation strides made ove the last couple of years have broken the stigma of processing huge datasets and have empowered minng and analysis of enormous data. Corporations in the data warehousing space have seen this pattern as the following huge chance to enable their clients to mine their historical data and help further their businesses in terms of including strategic and strategic value based on the insights picked up from their aggregated data over decades.

In this chapter, we will see run of the mill examples of how various businesses dissect their data and upgrade their business objectives. We will present some examples in the fields of money related services, retail, producing, telecommunications, social media, and human services.

**10. Mining Graph Data**

Data mining or Knowledge discovery in databases is a huge territory of study and is populated with numerous hypothetical and commonsense textbooks. In this thesis, we investigate one subject inside this field: mining dta that is represented as a graph. We endeavor to cover the full broadness of the theme, including graph controlling, visualization, and representation, mining techniques for graph data, and application of these keas to problems of current interest. The thesis is partitioned into three parts. Part I, Graphs, offers a prologue to basic graph terminology and techniques. In Part II, Mining Techniques, we investigate computational techniques for extricating patterns from graph data. These techniques give a diagram of the state of the art in successive substructure mining, connect analysis, graph kernels, and graph grammars. Pat III, Applications, describe application of mining techniques to hide graph-based application domains: compound graphs, bioinformatics data, Web graphs, and social networks. The thesis is focused toward alumni students, staff, and researchers from industry and the scholarly community who have some nature with basic computer science and data mining concepts.

The thesis is designed so that individuals with no back ground in breaking down graph data can figure out how to represent the dta as graphs, separate patterns or concepts from the data, and see how researchers apply the methodologies to genuine datasets. For those readers who might want to try different things with the techniques found in this thesis or test their very own ideas on graph data. This site

 

contains extra information on current techniques for mining graph data. Links sre also given to implementations of the techniques described in this thesis, as well as graph datasets that can be used for testing new or existing algorithms. With the appearance of and proceeded with prospect for enormous databases containing social and graphical information, the discovery of knowledge in such data is a significant test to the scientific and industrial communities. Handled applications for mining graph data from genuine domains has the possibility to make significant contributions of new knowledge. We trust that this book accelerates progress toward gathering this test.

### 11. Big Data Security and Privacy Issues in the Cloud

Numerous organizations request proficient solutions to store and examine immense measure of information. Cloud computing as an empowering agent provides scalable resources and significant economic benefits as diminished operational costs. This worldview raises an expansive scope of security and protection issues that must be mulled over. Multi-tenure, loss of control, and trust are key challenges in cloud driving environments. This subject reviews the existing technologies and a wide array of both prior and state-or-the-art projects on cloud security and protection. We arrange the existing research as indicated by the cloud reference architecture orchestration, resource control, and physical resource, and cloud service management layers, notwithstanding checking on the ongoing developments for improving the Apache Hadoop security as a standout amongst the most sent enormous data infrastructures. We also outline the outskirts research on security preserving data-intensive applications in cloud computing such as protection threat modeling and protection upgrading solutions.

### 12. Conclusion

While a number of anonymization techniques have been designed, it remains an open problem on how to use the anonymized data. In our experiments, we randomly generate the associations between column values of a bucket. This may lose data utility. Another direction to design data mining tasks using the anonymized data computed by various anonymization techniques.

### References


1. Pratiyush, G & Manu, S 2014, 'Data Mining in Education A Review On The Knowledge Discovery Perspective', International Journal of Data Mining and Knowledge Management Process, vol. 4, no. 5, pp. 47-60
2. Shweta, K 2012, 'Using Data Mining Techniques For Diagnosis And Prognosis Of• Cancer Disease', International Journal of Computer Science Engineering and Information Technology, vol. 2, no. 2, pp. 55-66
3. Abhijit, R &Kulkarni, RV 2011, 'Datamining Techniques: a Source for Consumer Behavior Analysis', International Journal of Database Management Systems, vol. 3, no. 3, pp. 45-56
4. Sankar, R 2011, 'Customer Data Clustering Using Data Mining Technique', International Journal of Database Management Systems, vol. 3, no. 4, pp. 1-11
5. Neethu, B &Priyanka, LT 2012, 'Customer Classification And Prediction Based On Data Mining Technique', International Journal of Emerging Technology and Advanced Engineering, vol. 2, no. 12, pp. 315-318
6. Sarah, NK &Alaa, ME, 2011, 'Implementation of Data Mining Techniques for Meteorological Data Analysis', International Journal of Information and Communication Technology Research, vol. 1, no. 3, pp. 96-100
7. Divya, T &Sonali, A 2013, 'A survey on Data Mining approaches for Healthcare', International Journal of Bio-Science and Bio-Technology, vol. 5, no. 5, pp. 241 -266
8. David, M 2012, 'Computational Historiography: Data Mining in a Century of Classics Journals', ACM Journal on Computing and Cultural Heritage, vol. 5, no. 1, pp. 33-319
9. Lakshmi, SP & Mohamed, SAR 2014, 'Educational Data Mining Applications', Operations Research and Applications: An International Journal, vol. 1, no. 1, pp. 23-29
10. Hemlata, S, Shalini, S &Seema, G 2011, 'A Brief Overview on Data Mining Survey', International Journal of Computer Technology and Electronics Engineering, vol. 1, no. 3, pp. 114-122
11. Dileep, DB &Kulkarni, RV 2013, 'A Review: Application of Data Mining Tools in CRM for Selected Banks', International Journal of Computer Science and Information Technologies, vol. 4, no. 2, pp. 199-201
12. Hameetha, SB 2013, 'Data Mining Tools and Trends An Overview', International Journal of Emerging Research in Management & Technology, vol. 2, no. 2, pp. 6-12


 