Modern Association Rule Mining Methods

Er.Anand Rajavat[1] and Er.Pranjal singh solanki[2]

[1]Asst. Prof, S.V.I.T.S, Indore and [2]Dept.of C.S.E, S.V.I.T.S, Indore (M.P)

ABSTRACT

Mining academic social network is becoming increasingly necessary with the increasing amount of data. It is a favorite topic of research for many researchers. The data mining techniques are used for the mining of academic social networks. In this paper, we are presenting an efficient frequent item set mining technique for social academic network. The proposed framework first processes the research documents and then the enhanced frequent item set mining is applied to find the strength of relationship between the researchers. The proposed method will be fast in comparison to older algorithms. Also it will takes less main memory space for computation purpose.

Keywords

Association Rules, Academic social network, CF tree, frequent item set mining, Support count.

1.Introduction

Social networks are built upon social relations among people who share common interests, activities etc. Social networking sites represent members through profile pages which contain their personal information like homepage, fields of interest, hobbies, contact details etc. However, the information is hidden in heterogeneous and distributed web pages. Many social networking sites offer search and mining services, by allowing a person to register and identify ties based on profile. Such networks can be briefly classified as informational, professional, educational, academic, news groups, sports based and so on. These networks can be visualized using graphs where users represent nodes and their relation represents edges. Social network analysis will be made simpler with the graphical representation and graph metrics would be useful to measure the dynamics of the network and individual properties of nodes.

Academic social networks represent collaboration of researchers and their publications. Different academic social networks include academia.edu, arnetminer.org, academic.research.microsoft.com, researchgate.net etc. It is observed that collaborative effort by people across the globe makes research strong from different perspectives. This proposed approach is based upon services offered by http://arnetminer.org which concentrates on accurately extracting researcher profile information from the web by integrating data from different sources. Coauthor path and coauthor graph presents visually the relation of co-authorship between researchers.

Resource Description Framework (RDF) is a W3C standard for describing resources. The necessity of Machine understandable data from Machine-readable data has lead to the specification of RDF framework. It provides basis for analysing metadata and provides interoperability between applications that exchange information on the web.In this context people i.e. members of social network are referred as the resource. In RDF, the *resource* is identified by

DOI : 10.5121/ijcsity.2014.2401

an URI and is represented as ellipse. Resources have *properties* which in turn have *values*. RDF graph has nodes and arcs that connect nodes and the RDF triple comprises subject, predicate and object nodes. Concrete syntax needed to create and exchange metadata is developed by using Extensible Markup Language, XML. RDF also requires namespace facility to precisely associate property with the schema that defines the property.

Consider a simple example of the sentence: *The individual whose name is Ora Lassila, email <lassila@w3.org>, is the creator of http://www.w3.org/Home/Lassila.*

<*rdf*:RDF> <*rdf*:Description

about="http://www.w3.org/Home/Lassila"> <*s*:Creator

rdf:resource="http://www.w3.org/staffId/8 574"/>

</*rdf*:Description> <*rdf*:Description

about="http://www.w3.org/staffId/8574”> <*v*:Name>Ora Lassila</*v*:Name> <*v*:Email>lassila@w3.org</*v*:Email> </*rdf*:Description> </*rdf*:RDF>

“Friend of A Friend”(FOAF) [18] project is based around the use of machine readable web home pages for people, groups, companies and others. FOAF is a linked information system that is built using decentralized semantic web technology designed to allow integration of data across variety of applications, web services and software systems. FOAF vocabulary is defined as a dictionary of named properties and classes using W3C’s RDF technology. FOAF integrates three kinds of networks; a) *social networks* b) *representational networks* and c) *information networks*. FOAF provides authoritative documentation of contents, status and purpose of RDF/XML vocabulary and are published as linked documents in the web. Main FOAF terms are grouped as Core, Social Web and Linked Data utilities. The Core terms describe characteristics of people and social groups that are independent of time and technology. FOAF defines classes of foaf:Person, foaf:Document, foaf:Image etc. and properties such as foaf:name, foaf:mbox , foaf:homepage etc. A member of a social network is represented with basic RDF template as given below.

<foaf:Person rdf:about ="#danbri" xmlns:foaf="http://xmlns.com/foaf/0.1/">

<foaf:name>Dan Brickley</foaf:name> <foaf:homepage

rdf:resource="http://danbri.org/"/> <foaf:openid

rdf:resource="http://danbri.org/" /> <foaf:img rdf:resource="/images/me.jpg"/> </foaf:Person>

With the increase in size and number of databases, the need of extracting useful information out of huge amount of data in databases keeps on growing. This is where knowledge discovery in databases comes into play.KDD helps in extracting out precious information and pattern from huge database. This helps in delivering better services, reducing cost and taking accurate and informed future decisions.

In recent years the size of database keeps on growing. This has led to a growing interest in the development of tools capable in the automatic extraction of knowledge from data. The term data mining or knowledge discovery in database has been used for digging out important information from large volume of database. The implicit information within databases thereby extracted out by data mining helps in the unraveling of patterns that can be used in variety of applications.

2. Problem Statement

Let I = {I1, I2, In} be a set of all items. A size k item set α, which consists of k items from I, is said to be frequent if α occurs in a transaction database D no lower than θ |D| times, where θ is a user-specified minimum support threshold (called MST), and the |D| is the total number of transactions in D.

Proposed Technique:

The proposed algorithm is as follows:

Inputs:

- Transaction data bases D1
- MST – Minimum Support Threshold
- MCT – Minimum Confidence Threshold

Output:

- A Set of Association Rules

Procedure:

Step 1: Scan the transaction data base & find the support count of each item

Step 2: Eliminate the infrequent items

Step 3: Now arrange the frequent items with higher support count first. This order will be used in the construction of the compact tree (CF-Tree)

Step 4: Construct CF-Tree by reading 1 transaction at a time

Step 5: Extract a sub tree ending in an item(For example, if e is the last item in a transaction database than we have to find a sub tree ending in e)

Step 6:

- Check that the item of step 5 (i.e. e) is frequent or not
- If it is frequent then extract it as frequent item
- New item (e) is frequent so now find the other frequent items ending with e(i.e. be,de,ce,..)
- Continue this recursive procedure until no item found

3.Result Analysis

1. Input Data Set

The input data set is as follows:

1 3 4

2 3 5

1 2 3 5

2 5

1 2 3 5

The MST is 40%.

4. Previous Algorithm

The results of the algorithm are as follows:

Output:

1 supp: 3

2 supp: 4

3 supp: 4

5 supp: 4

1 2 supp: 2

1 3 supp: 3

1 5 supp: 2

2 3 supp: 3

2 5 supp: 4

3 5 supp: 3

1 2 3 supp: 2

1 2 5 supp: 2

1 3 5 supp: 2

2 3 5 supp: 3

1 2 3 5 supp: 2

========== IM - STATS ==========

Candidates count : 15

The algorithm stopped at size 5, because there is no candidate

Frequent itemsets count : 15

Maximum memory usage : 0.55 mb

Total time ~ 40 ms

==========

BUILD SUCCESSFUL (total time: 0 seconds)

5. Proposed Algorithm

The results of the algorithm are as follows

1 :3

2 1:2

3 1 2:2

5 1 2 3:2

5 1 2:2

3 1:3

5 1 3:2

5 1:2

2 :4

3 2:3

5 2 3:3

5 2:4

3 :4

5 3:3

5 :4

run:

========== New Algo - STATS ============

Number of frequent itemsets: 15

Total time ~: 25 ms

Max memory:0.5078125

=====================================

BUILD SUCCESSFUL (total time: 0 seconds)

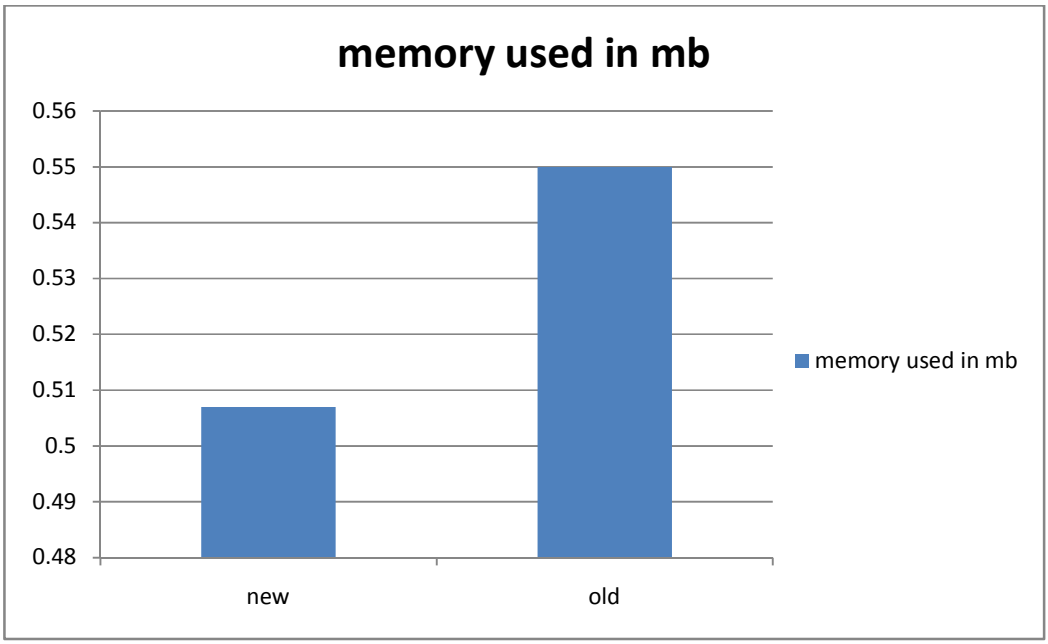

Figure: Memory Comparison

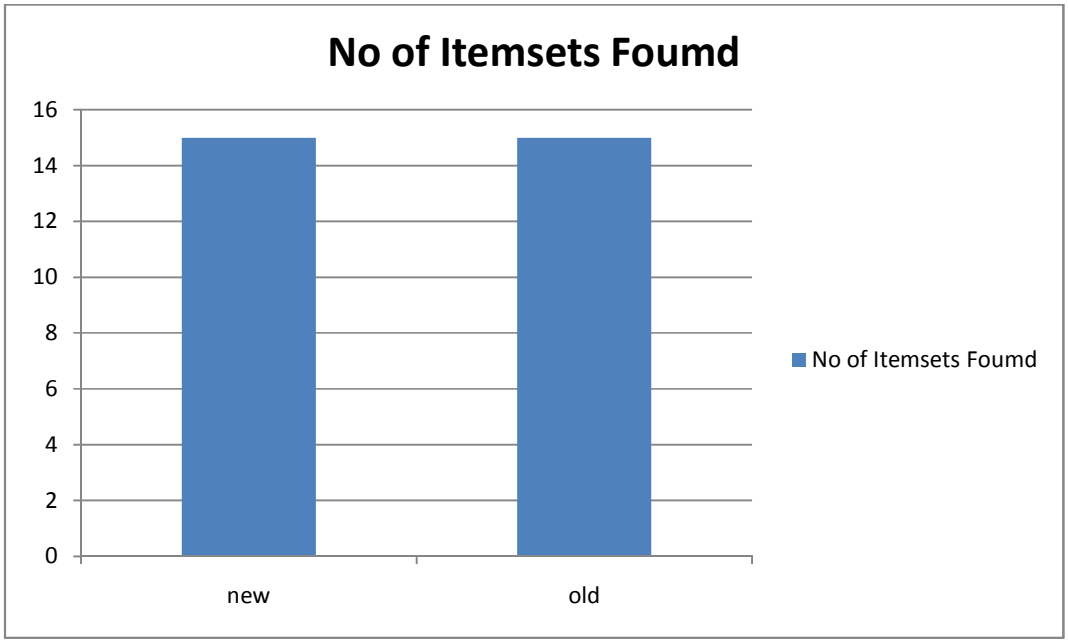

Figure: Result Comparison

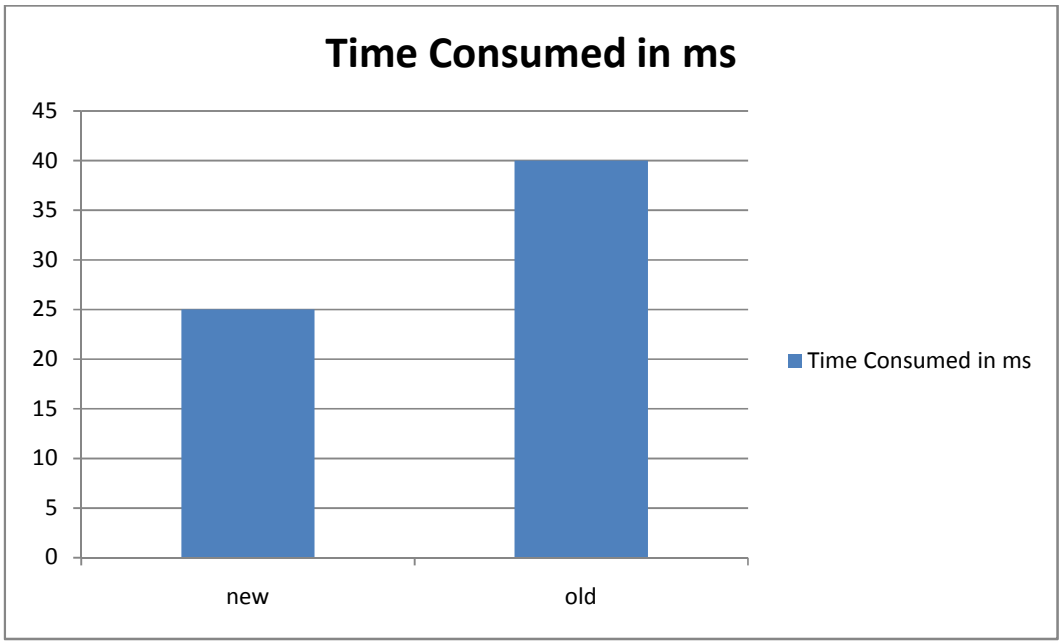

Figure: Time Consumption Comparison

6.Conclusion

In this paper, we presented a novel algorithm for mining frequent item sets. Frequent item set mining is crucial for association rule mining. We have compared our proposed algorithm with previous algorithm, it superseded previous algorithm both in terms of time taken and main memory used.

7. References

[1] A. Savasere, E. Omiecinski, and S. Navathe. "An efficient algorithm for mining association rules in large databases". In Proc. Int'l Conf. Very Large Data Bases (VLDB), Sept. 1995, pages 432–443.

[2] Aggrawal.R, Imielinski.t, Swami.A. "Mining Association Rules between Sets of Items in Large Databases". In Proc. Int'l Conf. of the 1993 ACM SIGMOD Conference Washington DC, USA.

[3] Agrawal.R and Srikant.R. "Fast algorithms for mining association rules". In Proc. Int'l Conf. Very Large Data Bases (VLDB), Sept. 1994, pages 487–499.

[4] Brin.S, Motwani. R, Ullman. J.D, and S. Tsur. "Dynamic itemset counting and implication rules for market basket analysis". In Proc. ACM-SIGMOD Int'l Conf. Management of Data (SIGMOD), May 1997, pages 255–264.

[5] C. Borgelt. "An Implementation of the FP- growth Algorithm". Proc. Workshop Open Software for Data Mining, 1–5.ACMPress, New York, NY, USA 2005.

[6] Han.J, Pei.J, and Yin. Y. "Mining frequent patterns without candidate generation". In Proc. ACM-SIGMOD Int'l Conf. Management of Data (SIGMOD), 2000

[7] Park. J. S, M.S. Chen, P.S. Yu. "An effective hash-based algorithm for mining association rules". In Proc. ACM-SIGMOD Int'l Conf. Management of Data (SIGMOD), San Jose, CA, May 1995, pages 175–186.

[8] Pei.J, Han.J, Lu.H, Nishio.S. Tang. S. and Yang. D. "H-mine: Hyper-structure mining of frequent patterns in large databases". In Proc. Int'l Conf. Data Mining (ICDM), November 2001.

[9] C.Borgelt. “Efficient Implementations of Apriori and Eclat”. In Proc. 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations, CEUR Workshop Proceedings 90, Aachen, Germany 2003.
[10] Toivonen.H. “Sampling large databases for association rules”. In Proc. Int’l Conf. Very Large Data Bases (VLDB), Sept. 1996, Bombay, India, pages 134–145.
[11] http://www.citeulike.org/tag/file-import-09-01-28
[12] http://www.ijarcsse.com/docs/papers/Volume_3/6_June2013/V3I6-0531.pdf
[13]http://f3.tiera.ru/2/Cs_Computer%20science/CsLn_Lecture%20notes/D/Data%20Warehousing%20and%20Knowledge%20Discovery,%207%20conf.,%20DaWaK%202005(LNCS3589,%20Springer,%202005)(ISBN%20354028558X)(550s).pdf
[14] http://www2.fiit.stuba.sk/iit-src/2005/zbornik.pdf