Global Journals Inc.

GLOBAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY
SOFTWARE & DATA ENGINEERING
Volume 12 Issue 10 Version 1.0 May 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Understanding Rule Behavior through Apriori Algorithm over Social Network Data

By S.S.Phulari, P.U.Bhalchandra, Dr.S.D.Khamitkar & S.N.Lokhande

S.R.T.M.University, Nanded, MS, India

Abstract - APRIORI algorithm is a popular data mining technique used for extracting hidden patterns from data. This paper highlights practical demonstration of this algorithm for association rule mining over a survey data set of students related to social network usage. We concluded with discussions on the number of research observations including new rules generated during the process.

GJCST-C Classification: *E.m*

Understanding Rule Behavior through Apriori Algorithm over Social Network Data

S.S.Phulari [α], P.U.Bhalchandra [α], Dr.S.D.Khamitkar [α] & S.N.Lokhande [α]

Abstract - APRIORI algorithm is a popular data mining technique used for extracting hidden patterns from data. This paper highlights practical demonstration of this algorithm for association rule mining over a survey data set of students related to social network usage. We concluded with discussions on the number of research observations including new rules generated during the process.

I. Introduction

Data mining is a technique that helps to extract important data from a large database. It is the process of sorting through large amounts of data and picking out relevant information through the use of certain sophisticated algorithms. As more data is gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform this data into information. Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. The myth possible with data mining includes automated prediction of trends and behaviors and automated discovery of previously unknown patterns. The most commonly used techniques in data mining are:

1. Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
2. Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
3. Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
4. Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. Sometimes called the k-nearest neighbor technique.
5. Rule induction: The extraction of useful if-then rules from data based on statistical significance.
6. Apriori is a classic algorithm used in data mining for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no transactions (Winepi and Minepi), or having no timestamps (DNA sequencing).

II. Analysis of Apriori Algorithm

Apriori was proposed by Agrawal and Srikant in 1994. The algorithm finds the frequent set L in the database D. It makes use of the downward closure property. The algorithm is a bottom search, moving upward level; it prunes many of the sets which are unlikely to be frequent sets, thus saving any extra efforts. Apriori algorithm is an algorithm of association rule mining. It is an important data mining model studied extensively by the database and data mining community. It Assume all data are categorical. It is initially used for Market Basket Analysis to find how items purchased by customers are related.

The problem of finding association rules can be stated as : Given a database of sales transactions, it is desirable to discover the important associations among different items such the presence of some items in a transaction will imply the presence of other items in the same transaction. As example of an association rule is:

Contains (T, "baby food") → Contains (T, "diapers") [Support= 4%, Confidence=40%]

The interpretation of such rule is as follows:

- 40% of transactions that contains baby food also contains diapers;
- 4% of all transactions contain both of these items.

The calculations of the Support(S) and Confidence(C) are very simple:

- CONF (A → B) = SUPP(AUB)
- SUPP(A)
- S (A) = (Number of transactions containing item A) /(Total number of transactions in the database)

*Author α : School of Computational Sciences, S.R.T.M.University , Nanded, MS, India. E-mail : Santoshphulari@gmail.com ,
E-mail : srtmun.parag@gmail.com, E-mail : s.khamitkar@gmail.com*

- S (A → B) =(Number of transactions containing items A and B) / (Total number of transactions in the database)

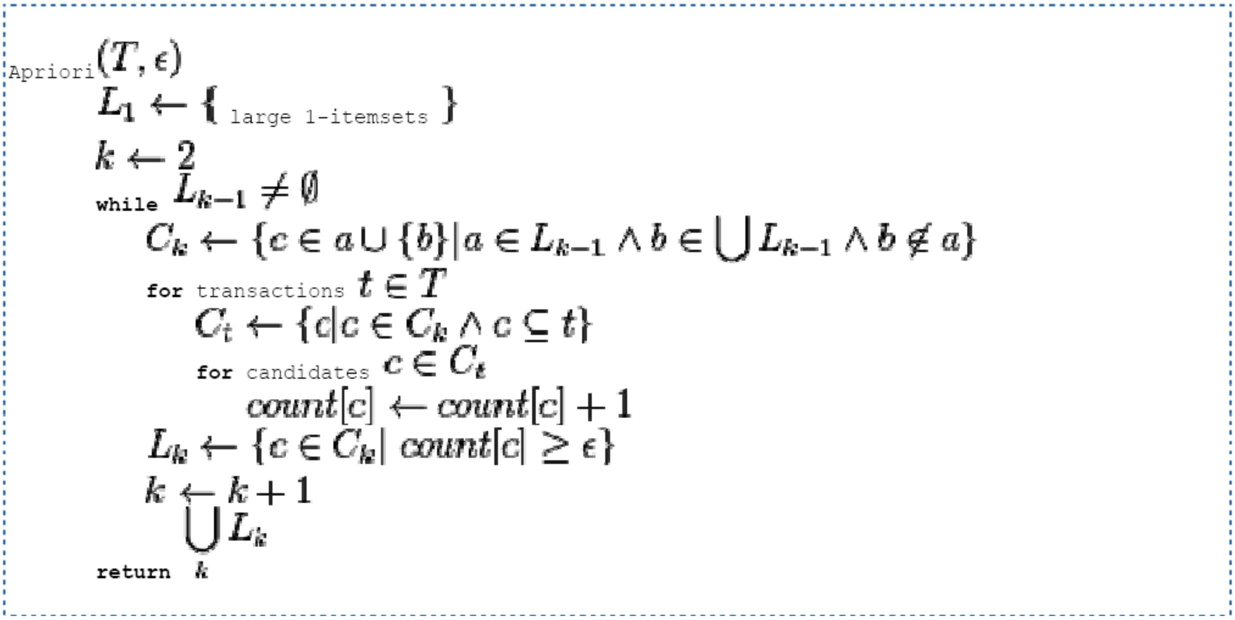
The above association rule is called single-dimension because it involves a single attribute or predicate (Contains). The main problem is to find all association rules that satisfy minimum support and minimum confidence thresholds, which are provided by user and/or domain experts. A rule is frequent if its support is greater than the minimum support threshold and strong if its confidence is more than the minimum confidence threshold.

Discovering all association rules is considered as two phase process where we find all frequent item sets having minimum support. The search space to enumeration all frequent item sets is on the magnitude of 2 * n. In second step, we generate strong rules. Any association that satisfies the threshold will be used to generate an association rule. The first phase in discovering all association rules is considered to be the most important one because it is time consuming due to the huge search space (the power set of the set of all items) and the second phase can be accomplished in a straightforward manner.

III. Algorithm for Apriori

The pseudo code for the algorithm is given below. For a transaction database T, and a support threshold of ϵ. Usual set theoretic notation is employed; though note that T is a multi set. C_k is the candidate set for level k. Generate () algorithm is assumed to generate the candidate sets from the large item sets of the preceding level, heeding the downward closure lemma.

$count[c]$ accesses a field of the data structure that represents candidate set c, which is initially assumed to be zero. Many details are omitted below, usually the most important part of the implementation is the data structure used for storing the candidate sets, and counting their frequencies.



```
Apriori$(T,\epsilon)$
    $L_1 \leftarrow \{$ large 1-itemsets $\}$
    $k \leftarrow 2$
    while $L_{k-1} \neq \emptyset$
        $C_k \leftarrow \{c \in a \cup \{b\} | a \in L_{k-1} \wedge b \in \bigcup L_{k-1} \wedge b \notin a\}$
        for transactions $t \in T$
            $C_t \leftarrow \{c | c \in C_k \wedge c \subseteq t\}$
            for candidates $c \in C_t$
                $count[c] \leftarrow count[c] + 1$
        $L_k \leftarrow \{c \in C_k | count[c] \geq \epsilon\}$
        $k \leftarrow k + 1$
    return $\bigcup_k L_k$
```

IV. Steps in Finding the Association Rules Using Apriori

A large supermarket tracks sales data by stock-keeping unit (SKU) for each item, and thus is able to know what items are typically purchased together. Apriori is a moderately efficient way to build a list of frequent purchased item pairs from this data. Let the database of transactions consist of the sets {1,2,3,4}, {1,2}, {2,3,4}, {2,3}, {1,2,4}, {3,4}, and {2,4}. Each number corresponds to a product such as "butter" or "bread". The first step of Apriori is to count up the frequencies, called the supports, of each member item separately: Table 1 explains the working of Apriori algorithm. We can define a minimum support level to qualify as "frequent," which depends on the context. For this case, let min support = 3. Therefore, all are frequent. The next step is to generate a list of all 2-pairs of the frequent items. Had any of the above items not been frequent, they wouldn't have been included as a possible member of possible 2-item pairs. In this way, priori *prunes* the tree of all possible sets. In next step we again select only these items (now 2-pairs are items) which are frequent and generate a list of all 3-triples of the frequent items (by connecting frequent pairs with frequent single items). In the example, there are no frequent 3-triples. Most common 3-triples are {1,2,4} and {2,3,4}, but their support is equal to 2 which is smaller than our min support. Table 2 explains these items.

Item	Support
1	3
2	6
3	4
4	5

Table 1:

Item	Support
{1,2}	3
{2,3}	3
{2,4}	4
{3,4}	3

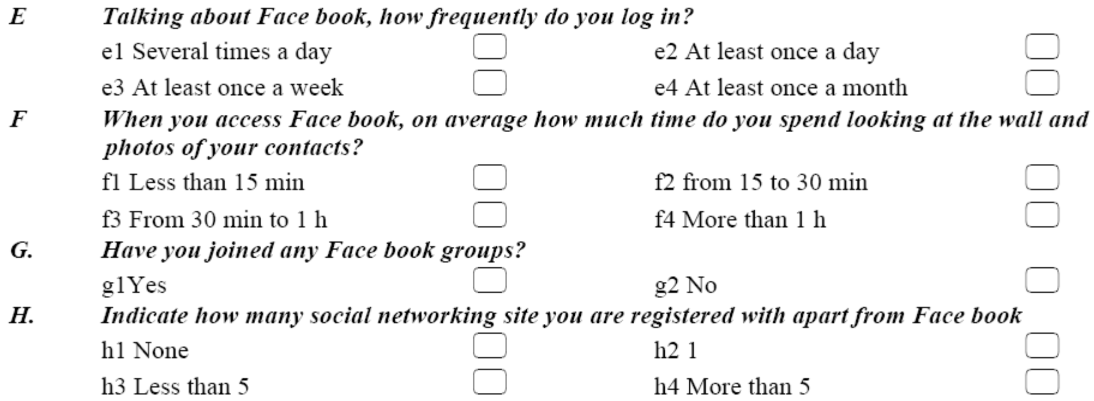
Table 2 :

V. Implementing Apriori Algorithm in Weka

WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. WEKA is open source software issued under the GNU General Public License.

VI. Data Set Features

A closed questionnaire of 56 questions, labeled A,B,C...... BB was prepared and circulated among 56 students. Maximum questions were having four options to answer. Theses answers were caught as a1, a2, a3, a4 (a means answer) . These questionnaire were circulated randomly to avoid mass copying of the answers and then collected after one hour. Out of 56 , only 43 questionnaire were correct in all respects. Remaining 13 needed interactions with the corresponding students as few questions on them were not answered by them. Since 13 students refused to re answer these, we have rejected them out. Microsoft Excel was used to tabulate the data in the questionnaire and 43 rows were created . A CSV(Comma Separated Values) sheet was made from it which has been fed as input to the WEKA Algorithm.



E ***Talking about Face book, how frequently do you log in?***
e1 Several times a day □ e2 At least once a day □
e3 At least once a week □ e4 At least once a month □

F ***When you access Face book, on average how much time do you spend looking at the wall and photos of your contacts?***
f1 Less than 15 min □ f2 from 15 to 30 min □
f3 From 30 min to 1 h □ f4 More than 1 h □

G. ***Have you joined any Face book groups?***
g1Yes □ g2 No □

H. ***Indicate how many social networking site you are registered with apart from Face book***
h1 None □ h2 1 □
h3 Less than 5 □ h4 More than 5 □

VII. Rules Generated

1. G=g1 34 ==> K=k1 34 conf:(1) [Those who join face book groups also have knowledge of Security Settings, Accuracy -34 %]
2. Ae=ae1 33 ==> Af=af1 33 conf:(1) [Those who use internet for preparing projects also use internet for preparing seminars , Accuracy -33 %]
3. D=d1 Ac=ac1 33 ==> Af=af1 33 conf:(1) [Those who have active account on facebook , and download lecture notes , also use internet for preparing seminars, Accuracy -33%]
4. G=g1 Af=af1 33 ==> K=k1 33 conf:(1) [Those who joined groups in facebook , and download seminar from internet , also have knowledge of security settings of facebook Accuracy -33%]
5. K=k1 Ae=ae1 32 ==> Af=af1 32 conf:(1) [Those who download lecture notes , also use internet for preparing projects and seminars, Accuracy -32%]
6. D=d1 G=g1 31 ==> K=k1 31 conf:(1) [Those who have active account on facebook and joins groups on facebook , can have knowledge of security settings, Accuracy-31%]

VIII. Conclusions

In this paper, we have studied association rule mining over survey dataset. Our study shows that mining multiple-level association rules from databases has wide applications and efficient algorithms can be developed for discovery of interesting and strong such rules in the database The larger the set of frequent item sets the more the number of rules presented to the user, many of which are redundant.

References Références Referencias

1. Arun K. Pujari, " Data Mining Techniques", 14th impression, 2008
2. R. Agrawal, T. Imielinski, A. Swami,"Mining Association Rules Between Sets of Items in Large Databases", Proc. SIGMOD Conference, 1993.
3. Rakesh Agrawal and Ramakrishnan Srikant, "Fast algorithms for mining association rules in large

databases", In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, Proc. of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.

4. Klementtinen M., et al "Finding interesting rules from large sets of discovered association rules." Proceedings of the CIKM 1994.
5. http://en.wikipedia.org/wiki/Apriori_algorithm