BULETINUL Universității Petrol – Gaze din Ploiești	Vol. LX No. 1/2008	49 - 56	Seria Matematică - Informatică - Fizică

Association Rule Mining as a Data Mining Technique

Irina Tudor

Universitatea Petrol-Gaze din Ploieşti, Bd. Bucureşti 39, Ploieşti, Catedra de Informatică
e-mail: irinat @upg-ploiesti.ro

Abstract

Association rule mining represents a data mining technique and its goal is to find interesting association or correlation relationships among a large set of data items. With massive amounts of data continuously being collected and stored in databases, many companies are becoming interested in mining association rules from their databases to increase their profits. For example, the discovery of interesting association relationships among huge amounts of business transaction records can help catalog design, cross-marketing, loss leader analysis, and other business decision making processes. A typical example of association rule mining is market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets".

Key words: *data mining, association rule, market basket analysis, knowledge discovery*

About Association Rule Mining

The abundance of data generates the appearance of a new field named data mining. Data collected in large databases become raw material for these knowledge discovery techniques and mining tools for "gold" were necessary. The current expert system technologies, which typically rely on users or domain experts to manually input knowledge into knowledge bases. This procedure contains errors, and it is extremely time-consuming and costly. Data mining tools which perform data analysis may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research [1, 2, 4].

Data mining represents the automatic process to discover patterns and relations between data stored in large databases called warehouses, the final product of this process being the knowledge, meaning the significant information provided by the unknown elements [7].

An explicit representation of this process is shown in the figure below.

In literature [3, 5] there were identified two major classes of data mining algorithms: supervised (classification) represented by the following: Bayesian, Neural Network, Decision Tree, Genetic Algorithms, Fuzzy Set, K-Nearest Neighbor and unsupervised algorithms such as Association Rules, Clustering.

One of the most popular data mining techniques is association rule mining. The patterns discovered with this data mining technique can be represented in the form of *association rules* [5, 4]. *Rule support* and *confidence* are two measures of rule interestingness. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts.

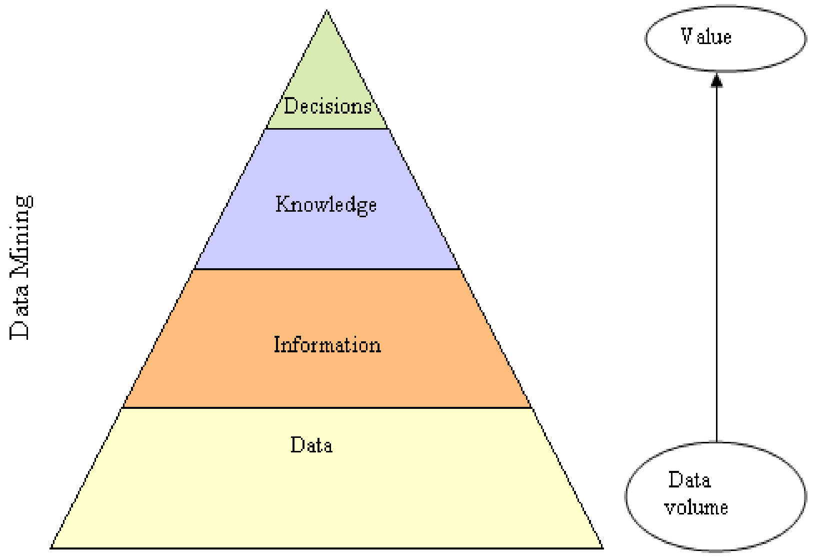

Fig. 1. Data mining process

Definition. Let I={I_1, I_2, ..., I_m} be a set of items. Let D, the task relevant data, be a set of database transactions where each transaction T is a set of items such that T $\subseteq$I. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if A $\subseteq$T. An association rule is an implication of the form A $\Rightarrow$ B, where A $\subset$I, B $\subset$I and A $\cap$B = Φ. The rule A$\Rightarrow$B holds in the transaction set D with support s, where s is the percentage of transactions in D that contain A $\cup$ B. The rule A $\Rightarrow$B has confidence c. in the transaction set D if c is the percentage of transactions in D containing A which also contain B.

That is, support(A $\Rightarrow$B) = Prob{A $\cup$B} and confidence(A $\Rightarrow$B) = Prob{B/A}.

Rules that satisfy both a minimum support threshold (minsup) and a minimum confidence threshold (minconf) are called *strong*. A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset [4].

Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties [6].

Example: Market Basket Analysis

Association rule mining searches for interesting relationships among items in a given data set. This paper presents a typical example of association rule technique and highlights the importance of knowledge discovery process in large databases.

Considering the example of a store that sells DVDs, Videos, CDs, Books and Games, the store owner might want to discover which of these items customers are likely to buy together. With the information above, the store could strive for more optimum placement of DVDs and Games as the sale of one of them may improve the chances of the sale of the other frequently associated item. The mailing campaigns may be fine tuned to reflect the fact that offering discount coupons on Videos may even negatively impact the sales of DVDs offered in the same campaign. A better decision could be not to offer both DVDs and Videos in a campaign. In this case, it is appropriate to use association rule mining to generate the optimum combination of products to increase sales.

Apriori algorithm data mining discovers items that are *frequently associated* together. The Apriori data mining analysis of the 9 transactions above is known as Market Based Analysis, as it is designed to discover which items in a series of transactions are frequently associated together.

In this example I considered a database, D, consisting of 9 transactions (table 1).

- Suppose minimum support count required is 2 (i.e. min_sup = 2/10 = 20 %);
- Let minimum confidence required is 60%;
- We have to first find out the frequent itemset using Apriori algorithm;
- Association rules will be generated using the two parameters minimum support and minimum confidence.

Assume that the dataset of 9 transactions below is selected randomly from a *universe* of 100000 transactions:

Table 1. An example of database with transactions

Transactions	**Items**
Customer1	BOOKS, CD, VIDEO
Customer2	CD, GAMES
Customer3	CD, DVD
Customer4	BOOKS, CD, GAMES
Customer5	BOOKS, DVD
Customer6	CD, DVD
Customer7	BOOKS, DVD
Customer8	BOOKS, CD, DVD, VIDEO
Customer9	BOOKS, CD, DVD

The Apriori algorithm [6] would analyze all the transactions in a dataset for each items support count. Any item that has a support count less than the *minimum support count* required is removed from the pool of candidate items.

First step of this algorithm is generating 1-itemset Frequent Pattern. The result is presented in the table 2.

Table 2. Itemsets and their support count

Itemsets	**Support Count**
{BOOKS}	6
{CD}	7
{VIDEO}	2
{GAMES}	2
{DVD}	6

At the beginning of association rule generation process each of the items is a member of a set of first candidate itemsets. The support count of each candidate item in the itemset is calculated (table 3) and items with a support count less than the minimum required support count are removed as candidates. The remaining candidate items in the itemset are joined to create second candidate itemsets each comprise of two items or members.

Table 3. Itemsets and their support count after condition verified

Itemsets	Support Count
{BOOKS}	6
{CD}	7
{VIDEO}	2
{GAMES}	2
{DVD}	6

Step 2 is represented by the generation of 2-itemset Frequent Pattern. After a combination process the matrix L1 provides the items for C2 matrix. The result is presented in the table 4.

Table 4. Itemsets and their support count after condition verified

Itemsets
{BOOKS, CD}
{BOOKS, VIDEO}
{BOOKS,GAMES}
{BOOKS,DVD}
{CD,VIDEO}
{CD,GAMES}
{CD,DVD}
{VIDEO,GAMES}
{VIDEO,DVD}
{GAMES,DVD}

Each item counts in the database to generate the next matrix, L2 as follow:

Table 5. Itemsets and their support count after L2xL2

Itemsets	Support Count
{BOOKS,CD}	4
{BOOKS,VIDEO}	2
{BOOKS,GAMES}	1
{BOOKS,DVD}	4
{CD,VIDEO}	2
{CD,GAMES}	2
{CD,DVD}	4
{VIDEO,GAMES}	0
{VIDEO,DVD}	1
{GAMES,DVD}	0

The next step is to discover the set of frequent 2-itemsets, L2 and the algorithm uses the L1 Join L1 procedure to generate a candidate set of 2-itemsets, C2. The transactions in D are scanned and the support count for each candidate itemset in C2 is accumulated (table 6). The set of frequent 2-itemsets, L2, is then determined, consisting of those candidate 2-itemsets in C2 having minimum support.

The following step consists in calculating the support count of each two member itemset from the database of transactions and 2 member itemsets that occur with a support count greater than or equal to the minimum support count are used to generate third candidate itemsets. The steps 1 and 2 are repeated to generate fourth and fifth candidate itemsets, the criteria used to stop this process being the value of support count of all the itemsets.

Table 6. Itemsets and their support count after condition verified

Itemset	**Support count**
{BOOKS, CD}	4
{BOOKS,VIDEO}	2
{BOOKS,DVD}	4
{CD,VIDEO}	2
{CD,GAMES}	2
{CD,DVD}	4

The next step consists of generation of 3-itemset frequent pattern as is shown in the table 7.

The generation of the set of candidate 3-itemsets, C3, involves use of the Apriori Property [6]. In order to find C3, a L2 Join L2 procedure is used. Join step is complete and prune step will be used to reduce the size of C3. Prune step helps to avoid heavy computation due to large Ck.

Table 7. 3-itemset Frequent Pattern

Itemsets
{BOOKS, CD,VIDEO}
{BOOKS, CD, DVD}
{BOOKS, CD, GAMES}
{CD,VIDEO,GAMES}
{CD,GAMES,DVD}
{CD, DVD, VIDEO}

Based on the Apriori property [6] that all subsets of a frequent itemset must also be frequent, we can determine that four latter candidates cannot possibly be frequent. Considering this example, lets take {BOOKS, CD, VIDEO}.The 2-item subsets of it are {BOOKS, CD}, {BOOKS, VIDEO} and {CD, VIDEO}. Since all 2-item subsets of {BOOKS, CD, VIDEO} are members of L2, we will keep {BOOKS, CD, VIDEO} in C3.

Therefore, C3= {{BOOKS, CD, VIDEO}, {BOOKS, CD, DVD}} after checking for all members of result of Join operation for Pruning. Now, the transactions in D are scanned in order to determine L3, consisting of those candidates 3-itemsets in C3 having minimum support.

Table 8. 3-itemset Frequent Pattern and their support count

Itemsets	**Support Count**
{BOOKS, CD,VIDEO}	2
{BOOKS, CD, DVD}	2

Step 4 is represented by the generation of 4-itemset frequent pattern.

Table 9. 4-itemset Frequent Pattern

Itemset
{BOOKS, CD, VIDEO, DVD}

The algorithm uses L3 *Join*L3 procedure to generate a candidate set of 4-itemsets, C4. Although the join results in {{BOOKS, CD, VIDEO, DVD}, this itemset is pruned since its subset {CD, VIDEO, DVD} is not frequent in L2.

Thus, C4= φ, and algorithm finishes, having found all of the frequent items. This completes the Apriori Algorithm.

All the candidate itemsets generated with a support count greater than the minimum support count form a set of frequent itemsets. These frequent itemsets will be used to generate strong association rules (where strong association rules satisfy both minimum support and minimum confidence).

The final step is to provide the association rules from frequent itemsets. The procedure consists in [6]:

For each frequent itemset “l”, generate all nonempty subsets of l;

For every nonempty subset s of l, output the rule “s ⇒(l-s)” if support_count(l) / support_count(s) >= min_conf where min_conf is minimum confidence threshold.

For the example given in this paper, L = {{BOOKS}, {CD}, {VIDEO}, {GAMES}, {DVD}, {BOOKS, CD}, {BOOKS, VIDEO}, {BOOKS, DVD}, {CD, VIDEO}, {CD, GAMES}, {CD, DVD}, {BOOKS, CD, VIDEO}, {BOOKS, CD, DVD}}.

For example L = {BOOKS, CD, VIDEO}. Its all nonempty subsets are {BOOKS, VIDEO}, {BOOKS, CD}, {CD, VIDEO}, {BOOKS}, {VIDEO}, {CD}.

Let minimum confidence threshold is 60%. The resulting association rules are shown below, each listed with its confidence.

R1: BOOKS and VIDEO ⇒ CD

Confidence = sc{BOOKS,CD,VIDEO}/sc{BOOKS,VIDEO} = 2/2 = 100% and R1 is selected.

R2: VIDEO and CD ⇒ BOOKS

Confidence = sc{BOOKS,VIDEO,DVD}/sc{VIDEO,CD} = 2/2 = 100% and R2 is selected.

R3: BOOKS and CD ⇒ VIDEO

Confidence = sc{BOOKS,VIDEO,DVD}/sc{BOOKS,CD} = 2/4 = 50% and R3 is rejected.

R4: BOOKS ⇒ VIDEO and CD

Confidence = sc{BOOKS,VIDEO,DVD}/sc{BOOKS} = 2/6 = 33% and R4 is rejected.

R5: VIDEO ⇒ BOOKS and CD

Confidence = sc{BOOKS,VIDEO,DVD}/{VIDEO} = 2/2 = 100% and R5 is selected.

R6: CD ⇒ BOOKS and VIDEO

Confidence = sc{BOOKS,VIDEO,DVD}/ {CD} = 2/7 = 28% and R6 is rejected.

In this way, we have found three strong association rules.

Conclusions

Association rule mining has a wide range of applicability such market basket analysis, medical diagnosis/ research, Website navigation analysis, homeland security, education, financial and business domain and so on. In this paper we present an example of data mining technique represented by association rule mining known as market basket analysis. In the market basket analysis example the existing database was analyzed to identify potentially interesting patterns. The objective is not only to characterize the existing database. What one really wants to do is, first, to make inferences to future likely co-occurrences of items in a basket, and, second and ideally, to make causal statements about the patterns of purchases: if someone can be persuaded to buy item I1 then they are also likely to buy item I2. The simple marginal and conditional probabilities are insufficient to tell us about causal relationships more sophisticated techniques are required. Association rule mining is easy to use and implement and can improve the profit of companies. The computational cost of association rule mining represents a disadvantage and future work will focus on reducing it.

References

1. Agrawal, R. Srikant, R. - Fast Algorithms for Mining Association Rules, *Proc. of the 20th Int'l Conference on Very Large Databases,* Santiago, Chile, 1994
2. Fayyad, U. M., Piatetsky - Shapiro, G., Smyth, P., Uthurusamy, R. - *Advances in Knowledge Discovery and Data Mining*, AAAI Press Series in Computer Science. A Bradford Book, The MIT Press, Cambridge Massachusetts, London Englan, 1996
3. Fayyad, W., Piatetsky - Shapiro, G., Smyth, P. - From data mining to knowledge discovery: An overview, In: *Advances in Knowledge Discovery and Data Mining*, W. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.), AAAI/MIT Press, Cambridge/USA, pp. 1 – 3, 1996

4. Han, J., Fu, Y. – Discovery of Multiple-Level Association Rules from Large Databases, *Proc. of 1995 Int'l Conf. on Very Large Data Bases (VLDB'95)*, Zürich, Switzerland, September 1995, pp. 420-431, 1995
5. Srikant, R., Agrawal, R. - Mining Generalized Association Rules, *Future Generation Computer Systems*, 13(2-3), 1997
6. Wasilewska, A. - *APRIORI Algorithm*, Lecture Notes, http://www.cs.sunysb.edu/~cse634/lecture_notes/07apriori.pdf, accessed 10.01.2007
7. * * * - *Data Mining*, CINECA site, http://open.cineca.it/datamining/, accessed 15.01.2008

Regulile de asociere - o tehnică de data mining

Rezumat

Regulile de asociere reprezintă o tehnică de data mining iar scopul acesteia este de a găsi relații de asociere sau corelații între datele dintr-un set mare de date. Odată cu colectarea şi depozitarea continuă a datelor, multe companii au devenit interesate de descoperirea regulilor de asociere în bazele lor de date pentru a le creşte profitul. Spre exemplu, descoperirea relațiilor interesante de asociere între volumele mari de date reprezentate de înregistrările tranzacțiilor poate ajuta la proiectarea catalogului de prezentare, încrucişarea vânzărilor şi a altor procese de luare a deciziilor în domeniul afacerilor. Un exemplu tipic de reguli de asociere îl reprezintă analiza coşului de piață (market basket analysis). Acest proces analizează clienții privind anumite obiceiuri ale acestora prin descoperirea asocierilor între diferite produse pe care clienții le plasează împreună în coşul de cumpărături.