# Frequent Itemset Mining and Association Rules

F

**Susan Imberman**
*City University of New York, USA*

**Abdullah Uz Tansel**
*Bilkent University, Turkey*

## INTRODUCTION

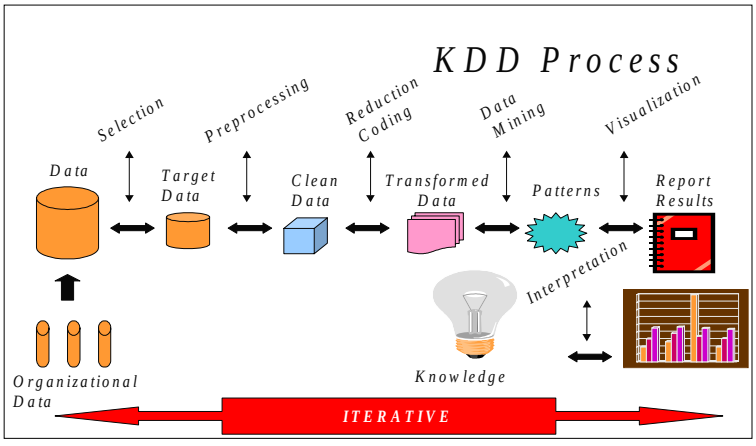
With the advent of mass storage devices, databases have become larger and larger. Point-of-sale data, patient medical data, scientific data, and credit card transactions are just a few sources of the ever-increasing amounts of data. These large datasets provide a rich source of useful information. Knowledge Discovery in Databases (KDD) is a paradigm for the analysis of these large datasets. KDD uses various methods from such diverse fields as machine learning, artificial intelligence, pattern recognition, database management and design, statistics, expert systems, and data visualization.

KDD has been defined as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad, Piaetsky-Shapiro, & Smyth, 1996). The KDD process is diagramed in Figure 1.

First, organizational data is collated into a database. This is sometimes kept in a data warehouse, which acts as a centralized source of data. Data is then selected from the data warehouse to form the target data. Selection is dependent on the domain, the end-user's needs, and the data mining task at hand. The preprocessing step cleans the data. This involves removing noise, handling missing data items, and taking care of outliers. Reduction coding takes the data and makes it usable for data analysis, either by reducing the number of records in the dataset or the number of variables. The transformed data is fed into the data mining step for analysis, to discover knowledge in the form of interesting and unexpected patterns that are presented to the user via some method of visualization. One must not assume that this is a linear process. It is highly iterative with feedback from each step into previous steps. Many different analytical methods are used in the data mining step. These include decision trees, clustering, statistical tests, neural networks, nearest neighbor algorithms, and association rules. Association rules indicate the co-occurrence of items in market basket data or in other domains. It is the only technique that is endemic to the field of data mining.

Organizations, large or small, need intelligence to survive in the competitive marketplace. Association rule discovery along with other data mining techniques are tools for obtaining this business intelligence. Therefore, association rule discovery techniques are available in toolkits that are components of knowledge management systems. Since knowledge management is a continuous process, we expect that knowledge management techniques will, alternately, be integrated into the KDD process. The focus for the rest of this article will be on the methods used in the discovery of association rules.

*Figure 1. The KDD process*




Copyright © 2006, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

## BACKGROUND

Association rule algorithms were developed to analyze market basket data. A single market basket contains store items that a customer purchases at a particular time. Hence, most of the terminology associated with association rules stems from this domain. The act of purchasing items in a particular market basket is called a transaction. Market basket data is visualized as Boolean, with the value 1 indicating the presence of a particular item in the market basket, notwithstanding the number of instances of an item; a value of 0 indicates its absence. A set of items is said to satisfy a transaction if each item's value is equal to 1. Itemsets refer to groupings of these items based on their occurrence in the dataset. More formally, given a set $I = \{ i_1, i_2, i_3, \ldots i_n \}$ of items, any subset of $I$ is called an itemset. A *k-itemset* contains $k$ items. Let X and Y be subsets of $I$ such that $X \cap Y = \phi$. An association rule is a probabilistic implication $X \Rightarrow Y$. This means if X occurs, Y also occurs. For example, suppose a store sells, among other items, shampoo (1), body lotion (2), hair spray (3), and beer (4), where the numbers are item numbers. The association rule *shampoo, hair spray* $\Rightarrow$ *beer* can be interpreted as, "those who purchase shampoo and hair spray will also tend to purchase beer."

There are two metrics used to find association rules. Given an association rule $X \Rightarrow Y$ as defined above, the support of the rule is the number of transactions that satisfy $X \cup Y$ divided by the total number of transactions. Support is an indication of a rule's statistical significance. Interesting association rules have support above a minimum user-defined threshold called *minsup*. Given the database represented in Figure 2, the support of the association rule *shampoo, hair spray* $\Rightarrow$ *beer* is equal to the number of transactions where shampoo, hairspray, and beer are equal to 1. This is equal to the shaded region

*Figure 2. Support of shampoo, hair spray* $\Rightarrow$ *beer 4/12 or 33%*

| 1. Shampoo | 2. Hair Spray | 3. Body Lotion | 4. Beer |
|---|---|---|---|
| 1 | 0 | 1 | 1 |
| 1 | 1 | 1 | 0 |
| 1 | 1 | 1 | 1 |
| 1 | 0 | 1 | 1 |
| 0 | 0 | 0 | 1 |
| 1 | 0 | 1 | 1 |
| 1 | 1 | 1 | 0 |
| 1 | 1 | 1 | 1 |
| 0 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 1 |

*Figure 3. Support of shampoo and hair spray*

| 1. Shampoo | 2. Hair Spray | 3. Body Lotion | 4. Beer |
|---|---|---|---|
| 1 | 0 | 1 | 1 |
| 1 | 1 | 1 | 0 |
| 1 | 1 | 1 | 1 |
| 1 | 0 | 1 | 1 |
| 0 | 0 | 0 | 1 |
| 1 | 0 | 1 | 1 |
| 1 | 1 | 1 | 0 |
| 1 | 1 | 1 | 1 |
| 0 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 1 |

and consists of a support of 4 out of 12 transactions, or 33%. Frequently occurring itemsets, called frequent itemsets, indicate groups of items customers tend to purchase in association with each other. These are itemsets that have support above the user-defined threshold, *minsup*.

Given an association rule $X \Rightarrow Y$ as defined above, the confidence of a rule is the number of transactions that satisfy $X \cup Y$ divided by the number of transactions that satisfy X. In Figure 3, the shaded portion indicates the support of Shampoo and Hair Spray. The confidence is then the support of the itemset Shampoo, Hairspray and Beer, divided by the support of Shampoo and Hairspray which equals 4/6 = 66%. It is common practice to define a second threshold based on a user-defined minimum confidence called *minconf*. A rule that has support above *minsup* and confidence above *minconf* is an interesting association rule (Agrawal, Imielinski, Swami, 1993; Agrawal & Srikant, 1994; Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996).

## FINDING ASSOCIATION RULES

Finding association rules above *minconf*, given a frequent itemset, is easily done and linear in complexity. Finding frequent itemsets is exponential in complexity and more difficult, thus necessitating efficient algorithms. A brute force approach would be to list all possible subsets of the set of items $I$ and calculate the support of each. Once an itemset is labeled frequent, partitions of the set's items are used to find rules above *minconf*. Continuing our example, assume *minsup* = 65%. Figure 4 lists all the subsets of the set of the items in Figures 2 and 3. The shaded areas indicate the frequent itemsets with support equal to or above 65%. The set of all itemsets forms a lattice, as seen in Figure 5.

5 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/frequent-itemset-mining-association-rules/16951

## Related Content

The Green Bay Chamber of Commerce: Foundation's Foundation
Philip Mattek (2010). *Knowledge Management Strategies for Business Development (pp. 84-114).*
www.irma-international.org/chapter/green-bay-chamber-commerce/38464

Promoting Organizational Knowledge Sharing
Jack S. Cookand Laura Cook (2004). *Innovations of Knowledge Management (pp. 300-321).*
www.irma-international.org/chapter/promoting-organizational-knowledge-sharing/23809

Dissemination in Portals
Steven Woods, Stephen R. Poteet, Anne Kaoand Lesley Quach (2008). *Knowledge Management: Concepts, Methodologies, Tools, and Applications (pp. 1521-1536).*
www.irma-international.org/chapter/dissemination-portals/25197

Organizing for Knowledge Management: The Cancer Information Service as an Exemplar
J. David Johnson (2008). *Knowledge Management: Concepts, Methodologies, Tools, and Applications (pp. 2261-2275).*
www.irma-international.org/chapter/organizing-knowledge-management/25257

The Impact of Personal and Positional Powers on Knowledge Management Systems
Vincent Scovetta (2017). *International Journal of Knowledge Management (pp. 18-34).*
www.irma-international.org/article/the-impact-of-personal-and-positional-powers-on-knowledge-management-systems/185762