

Chapter 33

Frequent Itemset Mining and Association Rules

Susan Imberman

College of Staten Island, City University of New York, USA

Abdullah Uz Tansel

Baruch College, City University of New York, USA

Category: Technologies for Knowledge Management

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,

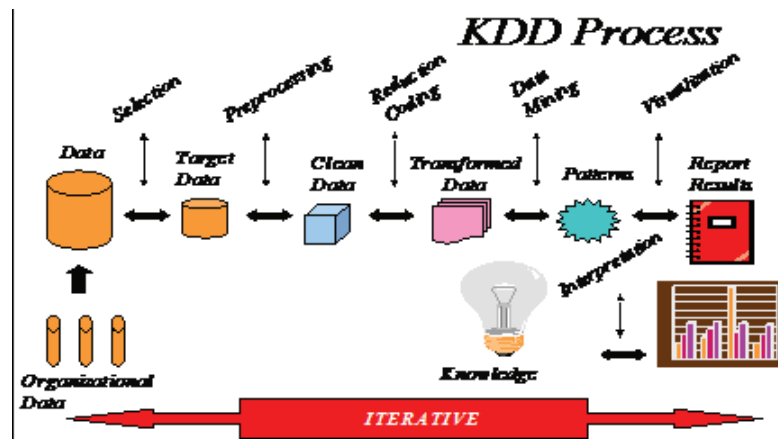
pattern recognition, database management and design, statistics, expert systems, and data visualization.

KDD has been defined as “the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.” (Fayyad, Piatetsky-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

DOI: 10.4018/978-1-59904-931-1.ch032

Figure 1. The KDD Process



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 market place. Association rule discovery along with other data mining techniques are tools for obtaining this business intelligence. Therefore, association rule discovery techniques are available in tool kits 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.

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, and a value of 0 indicating 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 $I = \{i_1, i_2, i_3, \dots, i_n\}$ 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 $X \cap Y = \emptyset$. 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.

9 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/48984

Related Content

Knowledge Forms and Enterprise Innovation Performance: An Evidence from the Dimensions of Stock and Flow

Qian Sunand Renyong Hou (2017). *International Journal of Knowledge Management* (pp. 55-70).

www.irma-international.org/article/knowledge-forms-and-enterprise-innovation-performance/193194

Challenges in Developing a Knowledge Management Strategy: A Case Study of the Air Force Materiel Command

Summer E. Bartczak, Jason M. Turnerand Ellen C. England (2008). *International Journal of Knowledge Management* (pp. 46-50).

www.irma-international.org/article/challenges-developing-knowledge-management-strategy/2720

Twenty-First Century Issues Impacting Turnover of IT Professionals: From Burnout and Turnover to Workplace Wellbeing

Valerie Fordand Susan Swayze (2018). *Innovative Applications of Knowledge Discovery and Information Resources Management* (pp. 29-62).

www.irma-international.org/chapter/twenty-first-century-issues-impacting-turnover-of-it-professionals/205397

Examining the Role of Technology Leadership on Knowledge Sharing Behaviour

Anugamini Priya Srivastavaand Yatish Joshi (2018). *International Journal of Knowledge Management* (pp. 13-29).

www.irma-international.org/article/examining-the-role-of-technology-leadership-on-knowledge-sharing-behaviour/213942

Can Soft Systems Methodology Identify Socio-Technical Barriers to Knowledge Sharing and Management?: A Case Study from the UK National Health Service

Alan C. Gilliesand Jeanette Galloway (2008). *International Journal of Knowledge Management* (pp. 90-111).

www.irma-international.org/article/can-soft-systems-methodology-identify/2740