

PART IV

Data Mining and Web-Knowledge Management

Introduction to PART IV: Data Mining and Web-Knowledge Management

Sergio A. Alvarez¹ and Carolina Ruiz² and Virginia Dignum³

¹ Boston College, Department of Computer Science, Chestnut Hill, MA 02467 USA

² Worcester Polytechnic Institute, Department of Computer Science, Worcester, MA
01609 USA

³ University of Utrecht, Institute of Information and Computing Sciences, 3508 TB
Utrecht, The Netherlands

Data Mining

The field of data mining comprises the design and analysis of efficient methods for the extraction of patterns from large data sets. Interest in this field has grown enormously over the past decade. The many scientific, industrial, and commercial applications of data mining include genomic analysis, intrusion detection, and personalization for electronic commerce. The first International Workshop on Rule-Based Data Mining, RBDM2001, held in conjunction with the 14th International Conference on Applications of Prolog (INAP2001), focused on data mining techniques, tools, and applications that employ a rule-based formalism as a primary component. Twelve papers from eight different countries were submitted to RBDM2001. Of these, six were accepted for inclusion in the conference proceedings. Three of these papers appear in the present collection and are briefly described below.

Association rule mining continues to be a topic of great interest within the field of data mining. Association rules provide a powerful representational formalism and are comparatively easy for humans to understand. The Apriori algorithm has been widely used for association rule mining in various contexts. Despite many successes, obstacles remain, both in terms of the efficiency of the mining process as well as in the generality of the rules allowed. The paper by Groth and Robertson (RBDM2001) addresses the problem of mining association rules in situations in which some itemsets are highly frequent. The paper introduces a method to lower the number of candidate itemsets whose frequencies are counted against the dataset during the generation of association rules. This is done by introducing a set of inference rules that represent sufficient conditions under which a candidate itemset can be proven to satisfy the minimum support requirement. They propose an exploratory analysis procedure based on those inference rules that effectively mines a subset of association rules from datasets in which the number of frequent itemsets is overwhelmingly high. The paper by Okoniewski, Gancarz, and Gawrysiak (RBDM2001) presents an extension of the quantitative association rules introduced by Aumann and Lindell in their KDD-1999 paper. Quantitative association rules capture deviations of subpopulations in numeric data. The original definition of these rules and the procedure

to mine them are limited to rules whose antecedents consist of just one item (or attribute–value pair). The paper extends quantitative associations to multiple dimensions by introducing an algorithm that constructs rules that can contain several items in their antecedents. These antecedents represent hyper–cubes in the high dimensional space that contain dataset instances which are significantly different from the rest with respect to a fixed data attribute of interest. The paper discusses the benefits of this approach to several application domains.

The computational complexity of a data mining task depends significantly on the set of attributes used to describe data instances. Attribute selection and reduction techniques are therefore of great interest in data mining. The paper by Okubo, Kudoh, and Haraguchi (RBDM2001) deals with attribute preprocessing for classification. They describe a technique for attribute coarsening based on an existing Is–A hierarchy. The objective is to find suitable abstractions for a given attribute, that is, partitions of the domain of the attribute that may be used instead of the original attribute without significantly affecting performance. The desirability of a given abstraction may be measured by the degree to which it affects the probability distribution of the class attribute. The focus of the paper is on using the Is–A dictionary to efficiently carry out a search in the lattice of abstractions. The authors propose lower and upper bounds for the search based on the dictionary.

Web-Knowledge Management

Knowledge has widely been acknowledged as one of the determining factors for corporate competitiveness and advantage. In the past years we have witnessed an explosion of approaches to knowledge management (KM). Practitioners and business managers alike agree that issues of technology, process, people, and content must be addressed to achieve success. The SOL Workshop Series offers a communication forum and meeting ground for practitioners and researchers engaged in developing and deploying advanced solutions for Organizational Learning, Knowledge Management, Case-based Reasoning and topics relating to the integration of these fields. The World Wide Web provides an exceptional means for the distribution, storage and sharing of knowledge. However, finding information becomes excessively more difficult as the Web grows. Moreover, knowledge in the web is not only embedded in sites but valuable knowledge can also be derived from the way individuals navigate and relate sites. The paper by Jung sees bookmarks has evidence for user preferences (SOL2001). The hierarchical trees used to characterize and store bookmarks through public web directory services provide valuable information for others. The paper proposes a method for normalization and management of topic hierarchies based on Bayesian networks. The method is applied to implement collaborative web browsing that efficiently and adaptively supports web browsing.