Fundamenta Informaticae 78 (2007) i-iii
IOS Press

Special Issue on Knowledge Discovery

1. Introduction

This Special Issue contains modified and extended versions of nine papers presented at the Fifteenth International Symposium on Methodologies for Intelligent Systems (ISMIS′05) which took place in the Inn Hotel at Saratoga Springs, New York on May 25-28, 2005. The goal of *ISMIS* symposia [1], [2], [3], [4], [5], [6] is to provide a platform for a useful exchange between theoreticians and practitioners, and to foster the cross-fertilization of ideas in the following areas: Active Media, Human-Computer Interaction, Autonomic and Evolutionary Computation, Intelligent Agent Technology, Intelligent Information Retrieval, Intelligent Information Systems, Knowledge Representation and Integration, Knowledge Discovery and Data Mining, Logic for Artificial Intelligence, Soft Computing, Web Intelligence, Web Services. All papers presented in this special issue focus on problems related either directly or indirectly to knowledge discovery and data mining.

ISMIS symposium has been organized to meet on annual basis (1986-1990). Since 1991, *ISMIS* meets only twice every three years (1991, 1993, 1994, 1996, 1997, 1999, 2000, 2002, 2003, 2005).

The symposium was sponsored by DOD/US Army grant No. W911NF-05-1-0103.

Editors of this special issue would like to dedicate it to Professor Zdzislaw Pawlak who passed away in Warsaw on April 7, 2006.

2. Papers in this Special Issue

This Special Issue of *Fundamenta Informaticae* brings together 9 contributions, all in the area related to knowledge discovery.

The paper *SPICE: A New Framework for Data Mining based on Probability Logic and Formal Concept Analysis*, by Liying Jiang and Jitender Deogun, integrates formal concept analysis (FCA) and probability logic to provide a more flexible and robust data analysis model. The proposed model is called SPICE - **S**ymbolic integration of **P**robability **I**nference and **C**oncept **E**xtraction. SPICE integrates the good features of both FCA and probability logic by employing probability logic to reason and make assertions on the concepts imported by the FCA. Within SPICE, authors reformulate some important notions in KDD such as *concepts*, *patterns*, and develop *maximal potentially useful patterns*. Based on these, they formalize association rule mining.

The paper *Hierarchical Hidden Markov Models for User/Process Profile Learning*, by Ugo Galassi, Marco Botta, and Attilio Giordana, presents a method for automatically synthesizing complex profiles

from traces, that builds on Hierarchical Hidden Markov Model and Profile Hidden Markov Model. The main contribution of the paper consists in organizing and generalizing in a unique framework different methods developed in the past. The outcome is a new architecture of the Hierarchical Profile Hidden Markov Model (HPHMM), which is powerful enough to model many real world problems, and has an affordable computational complexity. In principle, a HPHMM can have an arbitrarily large number of levels in the hierarchy. Nevertheless, up to now, no need emerged of going beyond two levels.

In the paper *Privacy Aware Data Management and Chase* by Seunghyun Im, the author shows that hiding confidential values of attributes *AV* from an information system cannot provide sufficient security against Chase when knowledge base is present. Author presents two algorithms that reduce the risk of confidential *AV* leakage with small amount of additional *AV* loss. The first algorithm is a bottom up approach that uses Chase closure to find the maximum set of *AV*s that can stay in an information system. The second algorithm is a top down approach that hides a set of *AV*s that can eliminate the largest number of rules involved in hidden *AV* reconstruction. Experimental results, provided by author, show that a substantial number of confidential *AV*s can be revealed by Chase.

In the paper *Towards Efficient Searching on the Secondary Structure of Protein Sequences*, by Minkoo Seo, Sanghyun Park, and Jung-Im Won, authors propose CSI (Clustered Segment Indexing), an efficient indexing scheme for approximate searching on the secondary structure of protein sequences. The proposed indexing scheme exploits the concept of *clustering* and *lookahead* to overcome a number of limitations listed in the paper. A pre-determined number of neighboring segments are grouped into a cluster which is then represented by three attributes. The first one denotes the type string of the cluster obtained by concatenating the *Type* attributes of the underlying segments, the second one denotes the length of the cluster obtained by summing up the *Len* attributes of the underlying segments, and the last one denotes the lookahead of the cluster obtained by concatenating the *Type* attributes of the segments *following* the cluster. Algorithms for exact match, range match, and wildcard match queries are also proposed and evaluated.

The paper *Unifying Framework for Rule Semantics: Application to Gene Expression Data*, by Marie Agier, Jean-Marc Petit, and Einoshin Suzuki presents three semantics devoted to gene expression data. The first one generates rules between genes according to their expression levels, i.e. under- or over-expressed genes. The second semantics analyzes the variations of gene expression levels and finally, the third semantics studies the evolution of gene expression levels between two consecutive samples, being understood that an order has to exist among them. Authors proposition has been implemented in a friendly graphical user interface to make it useful by biologists.

In the paper *Visualization of Differences between Rules' Syntactic and Semantic Similarities using Multidimensional Scaling*, by Shusaku Tsumoto and Shoji Hirano, authors propose a new visualization approach to show the similarity relations between rules based on multidimensional scaling, which assigns a two-dimensional cartesian coordinate to each data point from the information about similarities between this data and other data. Authors evaluated their method on three medical data sets. Experimental results show that knowledge useful for domain experts can be found.

In the paper *A Contribution to the Use of Decision Diagrams for Loading and Mining Transaction Databases*, by Ansaf Salleb and Christel Vrain, authors study the interest of *Binary Decision Diagrams (BDDs)* as a data structure for representing and loading transaction datasets. They introduce a coefficient, called the *Sparseness Coefficient* and experimentally show that it can be an interesting measure for evaluating the density of a database. Next, they investigate the feasibility of some common Data Mining steps, as computing the support of an itemset and even mining frequent itemsets. In their framework, a dataset

is viewed as a vectorial function, thus allowing, when possible, to load only this vectorial function into memory by means of a BDD.

In the paper *Privacy Preserving Database Generation for Database Application Testing*, by Xintao Wu, Yongge Wang, Songtao Guo, and Yuliang Zheng, authors investigate how to use the *general location model* to model an existing production database and how to use the model learned to generate a synthetic database. They examine in detail how to extract statistics and rules to estimate parameters of the general location model and how to resolve the potential disclosure of confidential information in data generation using model learned. This problem is related to, but not identical to, the widely recognized problem of privacy preserving data mining. For the scenario presented in this paper, the disclosure analysis is conducted at model level instead of tuple level.

In the paper *Sound Isolation by Harmonic Peak Partition For Music Instrument Recognition*, by Xin Zhang and Zbigniew Raś, authors propose a novel music information retrieval system with MPEG-7-based descriptors and they construct several classifiers which retrieve the time-frequency timbre information and isolate sound sources in polyphonic musical objects. Authors apply a clustering technique to separate the energy of FFT points in the power spectrum by taking each harmonic peak as a cluster centroid. Therefore, sound separation can be achieved by clustering the energy around each harmonic peak.

References

[1] *Foundations of Intelligent Systems*, Proceedings of ISMIS′05, M.-S. Hacid, N. Murray, Z.W. Raś, S. Tsumoto (editors), LNAI, No. 3488, Springer, 2005.

[2] *Foundations of Intelligent Systems*, Proceedings of ISMIS′03, N. Zhong, Z.W. Raś, S. Tsumoto, E. Suzuki (editors), LNAI, No. 2871, Springer, 2003.

[3] *Foundations of Intelligent Systems*, Proceedings of ISMIS′02, M.-S. Hacid, Z.W. Raś, D. Zighed, Y. Kodratoff (editors), LNAI, No. 2366, Springer, 2002.

[4] *Foundations of Intelligent Systems*, Proceedings of ISMIS′00, Z.W. Raś, S. Ohsuga (editors), Lecture Notes in Artificial Intelligence, No. 1932, Springer, 2000.

[5] *Foundations of Intelligent Systems*, Proceedings of ISMIS′99, Z.W. Raś, A. Skowron (editors), LNAI, No. 1609, Springer-Verlag, 1999.

[6] *Foundations of Intelligent Systems*, Proceedings of ISMIS′97, Z.W. Raś, A. Skowron (editors), LNAI, No. 1325, Springer, 1997.

Zbigniew W. Raś
Department of Computer Science
University of North Carolina
Charlotte, N.C. 28223, USA

Agnieszka Dardzińska
Department of Mathematics
Bialystok Technical University
ul. Wiejska 45A, 15-351 Bialystok, Poland