Control and Cybernetics

vol. **38** (2009) No. 1

Guest Editors' Introduction

Database systems have been tremendously successful over the past three decades and have become the most important development in the field of software engineering. While database studies achieved a considerable success in the marketplace and data management became almost synonymous with database systems, the recent development of new computing and storage devices, communication systems, text and multimedia technology, Web technology, information services, etc., requires the development of new methodologies, environments, and tools to cope with new problems and challenges, in particular — to facilitate the storage, access, manipulation, and modification of new types of data.

This special issue consists of fourteen papers on research that lays foundations for the future of data management and processing technologies, involving data mining, data warehousing, XML, and Web technologies. The papers included in this issue are enhanced versions of eleven of the best papers from the "Data Processing Technologies" conference (TPD'2007) held on 24-26 September, 2007, in Poznań, and three best papers from the 3rd ADBIS Workshop on Data Mining and Knowledge Discovery (ADMKD'2007) held on 2 October, 2007, in Varna (Bulgaria). These papers were selected from 48 papers accepted by the main refereed paper track of the TPD'2007 conference, out of a total pool of 70 submissions, and from 12 papers accepted by the ADMKD program committee out of 20 submissions.

The first paper, entitled "Analysis of monotonicity properties of some rule interestingness measures" addresses the problem of properties of some quantitative, objective interestingness measures for "if...then..." association rules. The authors analyze the problem whether popular interestingness measures, i.e., interest function, gain measure, and dependency factor, satisfy the M property of monotonic dependency on the number of objects satisfying or not the premise or the conclusion of an association rule. The authors show also relevant relationships among objective interestingness measures for "if...then..." association rules.

The second paper, entitled "An evaluation of quality in context based sequential pattern mining", addresses the problem of the context based sequential pattern discovery, namely, the discovery of sequential patterns with additional sets of context attributes. The paper presents and compares two approaches to mine context based sequential patterns, and, additionally, presents a new measure for comparing sets of context sequential patterns.

The third paper, entitled “Integration of candidate hash trees in concurrent processing of frequent itemset queries using Apriori”, presents a new method for the concurrent execution of the set of frequent itemset queries using the Apriori algorithm. The new method processes a batch of frequent itemset queries utilizing a common candidate hash tree for all the concurrently processed queries. The experimental results show that the proposed method is much more efficient in comparison with the previously proposed solution to this problem because it scales better with respect to the number of queries, and consumes less memory.

The fourth paper, entitled “Evaluation of node position based on email communication”, addresses the problem of analysis of a node position within the graph representing the social network. The paper presents a new method and a new measure of evaluation of the node position, which takes into account both the topology of the social network and the strength of connections between network nodes. The experimental results shown in the paper illustrate the main properties of the proposed measure.

The fifth paper, entitled, “Mining online auction social networks for reputation and recommendation”, presents a new method for mining the reputation of sellers in online auctions. The proposed method is based on a novel reputation measure for sellers in online auctions. The measure considers the linkage between sellers and buyers and mines the topology of the linkage network to derive useful knowledge about sellers. An experimental evaluation of the proposed method demonstrates the feasibility and usefulness of the proposed approach.

The sixth paper, entitled “How to improve efficiency of analysis of sequential data?”, proposes a new indexing method for optimizing retrieval of a large collection of sequences of sets based on sequence containment. The paper presents the logical and physical structure of the new index scheme, the algorithm of index construction and maintenance, and the algorithm of set subsequence query processing.

The seventh paper, entitled “Predicting access to materialized methods by means of hidden Markov model”, presents a new technique of predicting access to the results of materialized methods and objects, for the purpose of selecting the most appropriate moment of recomputing the materialized results of methods. The proposed technique is based on the Hidden Markov Model. The technique was implemented and experimentally compared to other techniques such as immediate recomputation, deferred recomputation, random recomputation, and PMAP.

The eighth paper, entitled “Time complexity of page filling algorithms in materialized aggregate list (MAL) and MAL/TRIGG materialization cost”, presents and evaluates the performance of three versions of a page filling algorithm in the materialized aggregate list (MAL). The author calculates their time complexity and estimates the best parameters of the MAL configuration.

The ninth paper, entitled “Schema mapping and query reformulation in peer-to-peer XML data integration system", addresses the problem of XML data integration in P2P environment in order to answer queries formulated against arbitrary chosen peers. The paper presents a new method of specifying local data schema, schema constraints, schema mappings and query reformulation in a uniform and precise way. Additionally, the paper addresses the problem of query propagation and merging modes, when missing data is to be discovered in P2P integration processes.

The tenth paper, entitled "An efficient SQL-based querying method to RDF-schemata", presents a semantic-preserving method of translating a SPARQL query (defined on RDF data structures) into a SQL query (defined on relational tables). Transforming queries instead of transforming data may provide better readability of queries and may lead to a significant increase in data extraction efficiency.

The eleventh paper, entitled "SXCCP+: simple XML concurrency control protocol for XML database systems", proposes a new locking protocol for concurrency control access in XML database systems. The protocol is based on primitive and indivisible operations, which may be treated as basic components of any set of operations of any XML access interface. Therefore, the proposed protocol is general and is not dependent on any particular XML interface.

The twelfth paper, entitled "Tracing cluster transitions for different cluster types", presents a framework, called MONIC+, for cluster type specific transition modeling and detection. The framework encompasses a typification of clusters and specific cluster type indicators by exploiting cluster topology and cluster statistics for the transition detection process. The experiments on real as well as synthetic datasets demonstrate the applicability of the framework.

The thirteenth paper, entitled "Active learning using pessimistic expectation estimators", proposes a new algorithm for cost-sensitive active learning using a conditional expectation estimator. The new estimator focuses on acquisitions that are likely to improve the profit. The evaluation of the proposed algorithm on four benchmark datasets is also given in the paper and demonstrates the superiority of the proposed method in comparison with other approaches.

The last paper, entitled "A linear Support Vector Machine solver for a large number of training examples", presents a new linear Support Vector Machine (SVM) algorithm and a solver specialized to solve SVM optimization problem with a large number of training examples. The contribution of the paper is twofold. Firstly, the Author demonstrates through experiments and practical comparison that the analytical center cutting plane scheme from Nesterov, used by his algorithm, is more efficient for harder parameter settings than the Kelley's cutting plane scheme used by other solvers. Secondly, to reduce the volume of data read from the file the Author proposes two optimization mechanisms. First, it checks whether the number of examples read in the current file scan is sufficient to gain new information about the supporting hyperplane location, which allows to finish earlier the file scans at the early stages of the algorithm run. Then, at the later stages of the algorithm, it disregards some input data that are clearly far from the hyperplane. The experiments on real as well as synthetic datasets demonstrate the applicability and efficiency of the proposed algorithm.

We sincerely hope that the readers will benefit from this special issue. The topics covered in these papers are timely and important. We would like to express our gratitude to the authors and the reviewers of this special issue. Additionally, we would like to thank all the program committee members, organizers, and sponsors who contributed to the success of the "Data Processing Technologies" conference (TPD'2007) and the 3rd ADBIS Workshop on Data Mining and Knowledge Discovery (ADMKD'2007).

Guest Editors,

Marcin Gorawski
Silesian University of Technology, Institute of Computing Science,
M.Gorawski@polsl.pl
Tadeusz Morzy
Poznań University of Technology, Institute of Computing Science,
Tadeusz.Morzy@put.poznan.pl
Robert Wrembel
Poznań University of Technology, Institute of Computing Science,
Robert.Wrembel@cs.put.poznan.pl