

Guest Editors' Introduction

The papers included in this special issue are extended versions of eight out of the best papers from the "Data Processing Technologies" conference (TPD'2010) held on June 21-23, 2010, in Pozna. The papers cover sequential pattern mining, transactional processing of complex business processes in service-oriented architecture, compression of indexes, reputation and trust systems, data streams, XML clustering, extraction and integration of Web resources, performance evaluation of consistency protocols in distributed systems, and bipartite graph modeling. After they had been selected, they underwent a significant revision and review process, during which they were considerably improved and extended. The result of this process is presented in the following pages.

The first paper, entitled "Accuracy of generalized context patterns in the context based sequential patterns mining", addresses the problem of generalized context patterns discovery. Context patterns can be mined in a database of sequences, where every sequence and element of a sequence may have an additional description specified by sets of context attributes. The goal of the generalization method is to reduce the size of the context patterns set by introducing a more compact and descriptive kind of patterns. It is achieved by finding clusters of similar context patterns in the mined set and transforming them to a smaller set of generalized context patterns. The paper introduces a definition of generalized context patterns, an algorithm developed for the generalization, and presents performance results. Additionally, an accuracy measure is also proposed to evaluate the generalized patterns accuracy.

Recently, the interest in the service oriented architecture (SOA) as the platform for the development of loosely coupled distributed applications increased tremendously. The second paper, entitled "Transaction mechanisms in complex business processes", addresses the problem of transactional properties of applications developed in the SOA environment. Despite various formal specifications defining the coordination of business processes in the SOA environment, solutions specifying how to use system mechanisms of transaction processing for this environment are still lacking. The paper discusses various transaction concepts in the context of SOA-based applications: transaction models, transaction properties and transaction processing, and presents a Transaction Coordinator environment, developed and implemented as a response to the needs of distributed applications processing.

Bitmap indexes are used in multiple technologies, including data warehouses, databases and other data repositories, and are among the basic data structures applied to query optimization. The size of a bitmap index strongly depends on the domain of indexed attribute. For wide domains, the size of bitmap index is too large to be efficiently processed. For this reason, various techniques of bitmap index compression have been proposed. However, compressed indexes have to be decompressed before being used by a query optimizer. The third paper, “GPU-PLWAH: GPU-based implementation of the PLWAH algorithm for compressing bitmaps”, presents a new approach to compressing and decompressing bitmap indexes supported by the processing power of graphics processing units (GPUs). The paper presents the parallel GPU version of the PLWAH compression technique. The technique was experimentally compared to the traditional CPU version of PLWAH and previously developed GPU parallelized algorithm for the WAH compression technique. The experiments show that the proposed solution significantly reduces compression/decompression time of bitmap indexes.

Online auctions have become an important business and the number of auction users is growing rapidly. At the same time, it is also an interesting field of research, concerning the analysis of trust and reputation. Auction portals require an efficient mechanism for building trust between participants, whereas most of them provide simple participation counts for reputation rating. The fourth paper, “Asymptotic trust algorithm: extension for reputation systems in online auctions”, describes a new trust system called Asymptotic Trust Algorithm (ATA) for online auction portals. The system can prevent fraud attempts and is easy to understand for users. The method developed for ATA can be applied on top of the participation count systems, currently used in most online auction sites, as not requiring any additional information other than positive, negative or neutral feedback on auction transactions.

The fifth paper, entitled “Semi-supervised approach to handle sudden concept drift in Enron data”, addresses the problem of concept drift, i.e. of detecting changes of concept definitions in data streams and of adapting classifiers to these changes. Previous research on this topic usually assumed that examples in a data stream are labeled. However, it is not always the case. The paper presents an approach that detects a concept drift in unlabeled data and retrains a classifier using a limited number of labeled examples. The usefulness of this approach is evaluated in the experimental study with Enron data concerning classification of user’s emails to multiple folders. The experimental results presented in the paper demonstrate that the proposed approach can detect frequent sudden changes of concepts in data streams and that the approach provides the classification accuracy comparable to the periodic retraining of the classifier based on windowing, reducing, however, the number of examples to be labeled.

The sixth paper, “Random graph generator for bipartite networks modeling”, presents a preferential attachment model adjusted to generation of bipartite graphs. The classical model can produce unipartite graphs with node

degree distribution following power law relation. Since, however, multi-partite graph topologies are becoming increasingly popular in social networks, there is a need to extend the classical model to enable it to create power law degree distribution in bipartite graphs. The paper provides formal mathematical analysis of the new model and presents empirical results on node degree distribution in real-life bipartite networks.

Extraction and integration of web data is an inspiring and important research area. It aims at tools and methods for extracting pieces of information from third-party web sites. Existing solutions rely on specialized APIs or XPath querying tools and are therefore not easily accessible to non technical end users. The seventh paper, "Similarity-based Web clip matching", presents a new comprehensive, non-Xpath integration platform, allowing end users to extract web page fragments using a simple query-by-example approach and, then, to combine these fragments into custom, integrated web pages. The presented platform handles both static and dynamic web resources by using automated matching of page fragment definitions against updated or restructured original web pages.

The final paper in this issue, entitled "Performance evaluation of different types of session guarantees version vectors", presents the results of performance evaluation of consistency protocols that are used for ensuring session guarantees using ODSAP protocols for server synchronization. Session guarantees define consistency properties of replicas in a distributed system from the standpoint of a migrating client. Consistency protocols of session guarantees are used for efficient representation of sets of writes performed in the distributed system, which are required to achieve the consistency. The evaluation of protocols was performed by means of simulation experiments using different forms of write sets representation.

We sincerely hope that readers will benefit from this special issue. The topics covered here are timely and important. We would like to thank the authors and the referees of this special issue. Additionally, we would like to thank all program committee members, organizers and sponsors who contributed to the success of the "Data Processing Technologies" conference TPD'2010.

Marcin Gorawski
Tadeusz Morzy
Robert Wrembel
Aleksander Zgrzywa
Guest Editors

