

## Editorial

Wee Keong Ng · Masaru Kitsuregawa · Jianzhong Li

Published online: 10 May 2007  
© Springer-Verlag London Limited 2007

We are pleased to present this special issue of the Knowledge and Information Systems Journal consisting of six selected papers from the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) held in Singapore in April 2006.

The Pacific-Asia Conference on Knowledge Discovery and Data Mining is a leading international conference in the area of data mining and knowledge discovery. It is an annual conference that provides an international forum for researchers and industry practitioners to share their new ideas, original research results and practical development experiences from all aspects of KDD data mining, including data warehousing, machine learning, databases, statistics, knowledge acquisition, automatic scientific discovery, data visualization, causal induction, and knowledge-based systems.

PAKDD 2006 received 501 paper submissions from 38 countries and regions in Asia, Australia, North America and Europe, of which 67 (13.4%) papers were accepted as regular papers and 33 (6.6%) papers as short papers. The six papers presented in this issue are selected from among the best papers accepted in PAKDD 2006 in terms of their technical contributions and paper presentation.

The paper on “Learning to Extract and Summarize Hot Item Features from Multiple Auction Web Sites” by Tak-Lam Wong and Wai Lam presents a graph-based approach to model dependencies among entities using Conditional Random Fields. This allows one to automatically semantically tag text segments in Web pages and express their relationships and dependencies. In this way, tags corresponding to product features can be identified from

---

W. K. Ng (✉)  
Nanyang Technological University,  
Singapore, Singapore  
e-mail: AWKNG@ntu.edu.sg

M. Kitsuregawa  
University of Tokyo, Tokyo, Japan

J. Li  
Harbin Institute of Technology, Harbin, China

auction sites and tasks such as hot item feature mining can be performed through an inference process using the automatically generated graphical structure.

The paper on “Privacy-Preserving SVM Classification by Jaideep Vaidya, Hwanjo Yu, Xiaoqian Jiang addresses a security issue in data mining: preserving data privacy while allowing useful knowledge to be discovered. The paper presents a timely discussion of privacy-preservation in SVM-based classification methods, after various existing data mining techniques in association analysis,  $k$ -means clustering, decision tree induction, naïve Bayes decision analysis, and Bayesian networks have been empowered with privacy-preserving feature.

The paper on “Mining Interesting Imperfectly Sporadic Rules” by Yun Sing Koh, Nathan Rountree, and Richard O’Keefe is on the topic of infrequent pattern mining. In particular, sporadic rules are association rules with low support but high confidence. Such patterns are useful for detecting anomalies but present difficulty in their discovery. The authors proposed a method to discover imperfectly sporadic rules—rules whose support of the antecedent falls below a maximum support but whose items may have quite individual support counts.

The paper on “Fast and Effective Clustering of XML Data Utilizing their Structural Information” by Richi Nayak utilizes intrinsic structural information in heterogeneous XML documents to perform clustering. A global criterion function is defined to determine the goodness of clustering that does not require the computation of pair-wise similarity between documents. The resultant clustering is accurate, fast, scalable, and robust.

The paper on “A Systematic Study on Parameter Correlations in Large Scale Duplicate Document Detection” by Shaozhi Ye, Ji-Rong Wen, and Wei-Ying Ma is a systematic study of the correlations among various parameters in large-scale duplicate document detection, such as similarity threshold, precision/recall, sampling ratio, and document size. The work provides significant insights, such as the finding that precision varies greatly on document with different sizes even with the same sampling ratio.

The paper on “Parallel Randomized Sampling for Support Vector Machine and Support Vector Regression” by Yumao Lu and Vwani Roychowdhury proposed a parallel randomized SVM and Support Vector Regression algorithm that has an average convergence rate that is fast and that works for general linear non-separable SVM and non-linear support vector regression problems. The authors have demonstrated the performance of the algorithm on real and synthetic applications.

We want to thank all authors for their contributions and support. We also express our heartfelt appreciation to the Editors-in-Chief at KAIS—Nick and Xindong—for making this issue possible. We hope you will find this issue useful and enlightening.