# Guest Editors' Introduction: Special Section on Intelligent Data Preparation

Chengqi Zhang, *Senior Member*, *IEEE Computer Society*, Qiang Yang, *Senior Member*, *IEEE*, and Bing Liu

## 1 Introduction

Data preparation (or data preprocessing) is the first step of data processing applications, including machine learning, data mining, data warehousing, information retrieval, and pattern recognition. Industrial practice indicates that more than 80 percent of the data mining work concentrates on data preparation. Indeed, once the data is well prepared, the mined results are more accurate and reliable. This means that data preparation is also a critical step for other data processing applications.

Among others, data preparation mainly involves data cleaning, data transformation, and data selection [1]. The principal aim of data preparation is to provide quality data for other steps of data processing applications, which are often accomplished in tandem with data mining and knowledge discovery. Indeed, the boundary can sometimes be blurry, but the general principle of data preparation can still be singled out from other data processing techniques. However, there has been relatively little focused work done for data preparation.

Recognizing the unique importance of data preparation, we have organized a series of workshops and special issues in top-quality journals to facilitate the development of data preparation techniques. Our first activity was the organization of the First International Workshop on Data Cleaning and Preprocessing, in conjunction with the IEEE International Conference on Data Mining (ICDM) in 2002, which was known as the DCAP workshop (i.e., data cleaning and preparation). High-quality papers from the workshop was also selected and published in a special issue on data preparation for data mining in the *Applied Artificial Intelligence* journal [1]. Following that activity, we then organized a special issue on information enhancement for data mining in *IEEE Intelligent Systems* magazine [2]. The current special section grew out of this continuing momentum. For illustration, Fig. 1, which shows the number of submissions to these journals and workshops, shows the growing interest in data preparation research.

The success of this special section can be attributed to the great support from many reviewers and the special support of Drs. Philip Yu and Xindong Wu, who are the former and current Editors-in-Chief of the *IEEE Transactions on Knowledge and Data Engineering*, respectively.

The papers selected in this special section emphasize both the practical techniques and new theoretical methodologies for data preparation, especially for data mining and machine learning applications. We have striven to include the papers in this special section that can benefit all areas of data processing.

## 2 The Contributions

Following our call for papers in 2004, we received 90 submissions. Following a rigorous period of peer review, we accepted 11 papers for this special section. These papers can be categorized into three main areas: discretization, feature selection, and Web intelligence.

### 2.1 Discretization

Discretization, as a necessary preprocessing step for many data mining and machine learning tasks, is the process of converting continuous attributes of a data set into discrete ones so that they can be treated as nominal features by many machine learning algorithms. Xiaoyan Liu and Huaiqing Wang designed a new criterion function, known as CF, that can assess the quality of discretization schemes. Based on CF, a heuristic method is proposed to find an approximately optimal discretization method. This function is based on a new measure of class heterogeneity in intervals from the view of class probability, which has been shown to outperform the class homogeneity based measures in accuracy.

Attribute correlation information has been difficult to handle by traditional discretization algorithms. To overcome this shortcoming, Sameep Mehta, Srinivasan Parthasarathy, and Hui Yang proposed a PCA-based unsupervised algorithm for discretizating continuous attributes in multivariate data sets. The algorithm leveraged the underlying correlation structure in the data set to obtain discrete intervals and ensured that the inherent correlations are preserved. They


- *C. Zhang is with the Faculty of Information Technology, University of Technology, Sydney, Broadway NSW 2007, Australia. E-mail: chengqi@it.uts.edu.au.*
- *Q. Yang is with the Department of Computer Science, Hong Kong University of Science and Technology, Room 3562 (Lift 25/26), Clearwater Bay, Kowloon, Hong Kong. E-mail: qyang@cs.ust.hk.*
- *B. Liu is with the Department of Computer Science, University of Illinois at Chicago, 851 S. Morgan (M/C 152), Chicago, IL 60607-7053. E-mail: liub@cs.uic.edu.*


*For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org.*

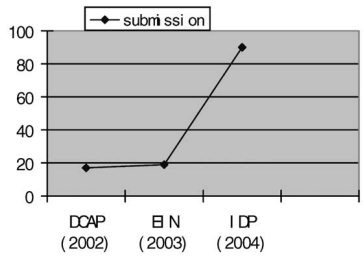



Fig. 1. The trend of submissions in our activities on data preparation.

considered the correlation among continuous attributes as well as the interactions between continuous and categorical attributes. The approach extends easily to data sets containing missing values. This algorithm can also discover hidden patterns in data.

### 2.2 Feature Selection

Feature selection is an important part of data preparation. It enables both cost-effective predictors to be built and a better understanding of the underlying process that generates the data. Hyunjin Yoon, Kiyoung Yang, and Cyrus Shahabi presented a family of novel methods for feature selection of multivariate time series data based on common principal component analysis. Their methods can overcome many shortcomings (such as the lack of correlation information among features) of the traditional algorithms.

To improve the quality of data, Guangzhi Qu, Salim Hariri, and Mazin Yousif designed an efficient method that is able to consider both the relevance of the features and the pair-wise feature correlation in order to improve the prediction and accuracy of data mining algorithms. They introduced a new feature-correlation metric and feature-subset-merit measure to quantify the relevance of and the correlation among features with respect to a desired data mining task (e.g., the detection of abnormal behaviors in a network service due to network attacks). The approach takes into account both the dependency among the features and their dependency with respect to a given data mining task. This was shown to lead to a higher accuracy.

To reduce the dimensionality, Jieping Ye, Qi Li, Hui Xiong, Haesun Park, Ravi Janardan, and Vipin Kumar proposed an LDA-based incremental dimensionality reduction algorithm called IDR/QR, which applies QR decomposition. Unlike other LDA-based algorithms, this algorithm does not require the whole data matrix to reside in the main memory of a computer. More importantly, with the arrival of new data items, the IDR/QR algorithm can limit the computational cost by applying efficient QR-based updating.

To automatically assign documents to a set of categories, Elias F. Combarro, Elena Montañés, Irene Díaz, José Ranilla, and Ricardo Mones proposed the selection of relevant features by means of a family of linear filtering measures that are simpler than the classic measures. This method is useful for information retrieval.

### 2.3 Web Intelligence

As the Web has emerged as a large, distributed data repository, it is easy nowadays to access a large number of data sources. To make the Web data usable, Sandip Debnath, Prasenjit Mitra, Nirmal Pal, and C. Lee Giles noticed the fact that there are certain similarities among the Web pages' complicated structures since dynamically generated Web documents have some form of underlying templates. Based on some observations, the authors proposed three algorithms: ContentExtractor, FeatureExtractor, and K-FeatureExtractor. With the above algorithms, "primary content block" could be identified and, in this way, it results in smaller indexed databases and improves the recall and precision of the search engine.

To integrate and reuse Deep Web data, James Caverlee and Ling Liu designed a new unsupervised algorithm with four steps: collecting, clustering, identifying, and partitioning to extract the deep Web page structure. In this algorithm, the structure information and content information are both used in constructing a similarity measure for reasoning.

Bin Gao, Tie-Yan Liu, Guang Feng, Tao Qin, Qian-Sheng Cheng, and Wei-Ying Ma observed that the hierarchical classifiers outperform corresponding flat ones in both efficiency and accuracy. However, hierarchical taxonomies are often not explicitly given. The authors proposed a novel algorithm to automatically mine a hierarchical structure from the flat taxonomy of a data corpus with three steps. The first step is to compute matrices to represent the relations among categories, documents, and terms. The second step is to cocluster the three substances at different scales based on the above matrices. The third step is to construct a hierarchical taxonomy from the category clusters. Their algorithms can discover very reasonable taxonomy hierarchy and help improve the classification accuracy as compared to previous methods.

To provide an integrated access to multiple, distributed, heterogeneous databases and other information sources, Zhiyong Peng, Qing Li, Ling Feng, Xuhui Li, and Junqiang Liu advocated using an object deputy model to realize materialized views for data warehouse. Inconsistency and conflicts from preparing data in data warehouses can be resolved by using the object deputy model.

Finally, Saeed Hashemi presents a wrapper-based method for data cleaning in which an atypical sequential removing (ASR) method is used for removing outliers. It offers a linear running time with little accuracy loss. This enables the wrapper methodology to be applicable for data cleansing, which, in the past, was prohibitive in time cost.

## 3 Conclusion

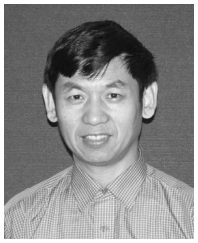
While previous works have been focused on data processing applications, there remains a large gap between the available data and the machinery available to process the data [2]. The large number of submissions for this special section encouraged us to continue to develop diverse forums for advancing the study of data preparation. We believe this research direction will receive more support from academic scholars and industrial practitioners.

Chengqi Zhang
Qiang Yang
Bing Liu
*Guest Editors*

## REFERENCES

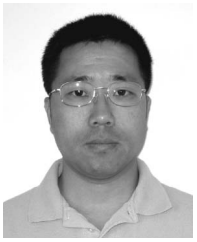
[1] S. Zhang, Q. Yang, and C. Zhang, "Data Preparation for Data Mining," *Applied Artificial Intelligence,* vol. 17, nos. 5-6, pp. 375-382, 2003.

[2] S. Zhang, C. Zhang, and Q. Yang, "Information Enhancement for Data Mining," *IEEE Intelligent Systems,* vol. 19, no. 2, pp. 12-13, 2004.



**Chengqi Zhang** has been a research professor on the Faculty of Information Technology at the University of Technology, Sydney (UTS) since 2001. He received the PhD degree from the University of Queensland, Brisbane, in computer science and a doctor of science (higher doctorate) degree from Deakin University, Australia. His research interests include data mining and multiagent systems. He has published approximately 100 refereed papers and three monographs, including 12 top-journal papers in IEEE/ACM transactions, artificial intelligence, and information systems. He is the leader of the Data Mining Program at the Capital Market Cooperative Research Centre and a member of UTS Research Management Committee. He is a senior member of the IEEE Computer Society, a member of the steering committee of PRICAI (Pacific Rim International Conference on Artificial Intelligence) and PAKDD (Pacific-Asia Conference on Knowledge Discovery and Data Mining), serves as an associate editor for three international journals, including the *IEEE Transactions on Knowledge and Data Engineering*, served as general chair, pc chair, or organizing chair for four international conferences and a member of the program committees for many international or national conferences. His personal Web page is at http://www-staff.it.uts.edu.au/~chengqi/.

**Qiang Yang** received the bachelor's degree from Peking University in 1982 and the PhD degree from the University of Maryland, College Park, in 1989. He was a faculty member at the University of Waterloo and Simon Fraser University in Canada between 1989 and 2001. At Simon Fraser University, he held an NSERC Industrial Chair from 1995 to 1999. He is currently a faculty member at the Hong Kong University of Science and Technology. Dr. Yang's research interest is in artificial intelligence planning, machine learning, and case-based reasoning, as well as data mining. He has published two books and more than 100 research articles in conferences and journals. He was the conference chair of the 2001 International Conference on Case-Based Reasoning, a program cochair of the 2000 Canadian AI Conference, and a tutorial cochair of the AAAI 2005 Conference. He has been a guest editor for the *IEEE Transactions on Knowledge and Data Engineering*, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *IEEE Intelligent Systems*, *Computational Intelligence Journal*, and *Applied Intelligence Journal*. He is also on the editorial board of the *IEEE Transactions on Knowledge and Data Engineering*, the *Knowledge and Information Systems Journal*, and *Web Intelligence Journal*. He is a senior member of the IEEE.



**Bing Liu** received the PhD degree in artificial intelligence from the University of Edinburgh. He is an associate professor in the Department of Computer Science, University of Illinois at Chicago (UIC). Before joining UIC in April 2002, he was with the National University of Singapore. His current research interests include data mining, Web and text mining, and machine learning. Since 1996, he has been active in data mining research and has published many papers in leading conferences and journals related to data mining, Web mining, and Artificial Intelligence (e.g., KDD, WWW, AAAI, IJCAI, ICML, and the *IEEE Transactions on Knowledge and Data Engineering*). He served (or serves) on the technical program committees of many data mining and Web related international conferences, including KDD, WWW, ICML, VLDB, and AAAI. He is currently an associate editor of *SIGKDD Explorations* and was an associate editor of the *IEEE Transactions on Knowledge and Data Engineering*. He also serves on the editorial boards of two other international journals related to data analysis and Web technology. Further information about him can be found at http://www.cs.uic.edu/~liub.