

Information Enhancement for Data Mining

Shichao Zhang and Chengqi Zhang, *University of Technology, Sydney*

Qiang Yang, *Hong Kong University of Science and Technology*

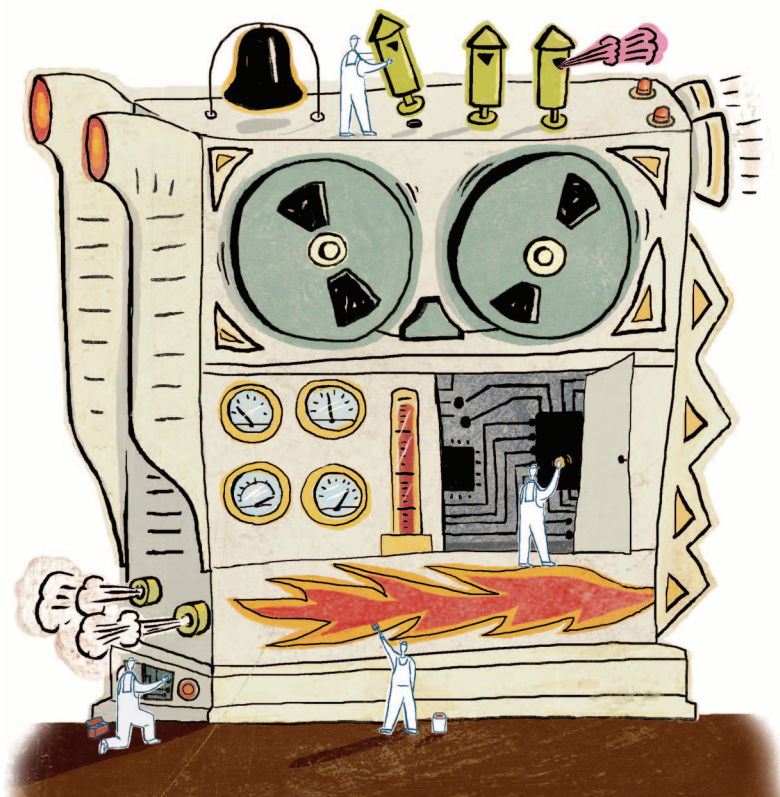
Many data analysis applications—such as data mining, information retrieval, machine learning, Web data management, data warehousing, and pattern recognition—need information enhancement. This involves taking data in their raw form, removing as much noise and redundancy as possible, and bringing out a core that's ready

for further processing. Indeed, information enhancement, which straddles data preprocessing and data mining, often presents itself as a less glamorous but more critical step than other steps in data mining applications; a minor information enhancement adjustment could bring higher effectiveness. Therefore, information enhancement is a crucial research topic. However, much work in relevant fields, such as data mining, is based only on quality data. That is, researchers have been assuming that the input to the data mining algorithms conforms to well-defined data distribution, containing no missing, inconsistent, or incorrect values.¹ This leaves a large gap between the available data and the machinery available to process the data.

The need for information enhancement

Information enhancement is important in three aspects:

- Real-world data is not pure. In fact, real-world data might be incomplete, noisy, or inconsistent, which can disguise useful patterns.
- High-performance data mining systems require quality data. Information enhancement, therefore, is critical in generating a data set that is cleaner and smaller than the original, which can significantly improve data mining's efficiency.
- Quality data yields more useful patterns for discovery. Often, the preprocess step in which we enhance the data is interleaved with the data mining step.



The Authors

We observe that data preprocessing and data mining aren't two separate steps in the data mining life cycle; instead, they form a process that crosses boundaries by straddling the traditional phases. Indeed, in real-world practice, data cleaning and data mining are intertwined.

In particular, as the Web rapidly becomes a channel for information flood, individuals and organizations take into account the Internet's low-cost information and knowledge when making decisions. So, researchers and practitioners must intensify efforts to develop appropriate techniques for efficiently using and managing data. Although data mining technology can support data analysis applications within these organizations, we must be able to enhance information from raw data to enable efficient and quality knowledge discovery. Thus, developing information enhancement technologies and methodologies is a challenging and critical task.

The articles in this special issue emphasize practical techniques and methodologies for information enhancement for data mining applications. We have striven to include in this issue articles that can benefit all areas of data analysis.

The contributions

We can categorize the articles into three main parts: data clustering, data cleaning, and Web intelligence.

Data clustering

Eugene Tuv and George Runger propose a new scoring method for variables in heterogeneous (mixed-type) data, using a nontraditional clustering approach called supervised-contrastive-independence clustering. Their method is computationally efficient and flexible in mapping categorical variables to numeric scores in mixed-type data.

Taghi M. Khoshgoftaar, Naeem Seliya, and Shi Zhong describe an interactive approach for software quality estimation that combines unsupervised learning and experts. This approach is effective in predicting both software modules' fault proneness and potential "noisy" (for example, mislabeled) modules.

Data cleaning

Mong Li Lee, Wynne Hsu, and Vijay Kothari formalize a solution to a new real-life data-quality problem that's just one of potentially many. Their approach uses context information to clean up spurious links in data by first identifying and retrieving the data containing potential spurious links, then performing a context similarity comparison to determine records with high overlaps. In the process, the degree of over-

lapping context indicates the likelihood of existing spurious links.

Data quality is of prime concern to any task involving data analysis. Choh Man Teng designs a process to correct potential errors by tenfold cross-validation comprising prediction and adjustment. In the prediction phase, suspect elements in data are identified together with a nominated replacement value. In the adjustment phase, the algorithm selectively incorporates the nominated changes into the analyzed data set.

In the CRM (customer relationship management) industry, analyzing customer churn is important for keeping customers and providing higher values. To highlight the importance of preprocessing data as a preparation for predicting customer behavior such as customer churn, Lian Yan, Richard H. Wolniewicz, and Robert Dodier present techniques, experiences, and lessons for customer behavior prediction in the telecom industry. The data preparation involves understanding customers' data and business practice, defining modeling targets, extracting data, reprocessing raw data, and compensating the available data for greater model accuracy.

Ying Yang and Xindong Wu analyze four parameters that can help measure the performance of an induction algorithm in feature elimination. They design a feature elimination method that considers not only the data and the target concept, but also the induction algorithm that will learn the target concept from the data.

Web intelligence

Doru Tanasa and Brigitte Trousse advocate an approach for preprocessing multiple Web server logs for Web usage mining. This method can increase data quality and reduce in a significant but relevant manner the size of the Web servers' log files.

The diversity of data and data mining tasks offer many challenging research issues for information enhancement:

- Constructing interactive and integrated information enhancement and data mining environments
- Establishing information enhancement theories
- Developing efficient and effective information enhancement methods and systems for mining multiple data sources, including internal and external data
- Exploring efficient information enhancement techniques for Web intelligence ■



Shichao Zhang is an assistant president at the Guangxi Teachers University. He is also a senior research fellow at the Faculty of Information Technology, University of Technology, Sydney. His recent research interests include data analysis and smart e-intelligence. He received his PhD in computer science from Deakin University. Contact him at the Univ. of Technology, Sydney, Broadway NSW 2007, Australia; zhangsc@it.uts.edu.au.



Chengqi Zhang is a research professor in the Faculty of Information Technology at the University of Technology, Sydney. His research interests include data mining and multiagent systems. He received his PhD in computer science from the University of Queensland, Brisbane and a Doctor of Science degree from Deakin University. He is a member of the IEEE, an associate editor on the editorial board for *Knowledge and Information Systems: An International Journal*, and a member of the editorial board of the *International Journal of Web Intelligence and Agent Systems*. Contact him at the Univ. of Technology, Sydney, Broadway NSW 2007, Australia; chengqi@it.uts.edu.au.



Qiang Yang is a faculty member in the Department of Computer Science, Hong Kong University of Science and Technology. His research interests include planning, case-based reasoning, and data mining. He received his PhD in computer science from the University of Maryland. He is a member of the IEEE. Contact him at the Dept. of Computer Science, Hong Kong Univ. of Science and Technology, Clearwater Bay, Kowloon, Hong Kong, China; qyang@cs.ust.hk.

Acknowledgments

Grants from the Australian ARC and UTS in Australia support Shichao Zhang's work. Grants from the Australian ARC, UTS in Australia, and the Australian CMCRC support Chengqi Zhang's work. Grants from the Hong Kong Research Grant Committee and the Hong Kong University of Science and Technology support Qiang Yang's work.

References

1. S. Zhang, Q. Yang, and C. Zhang, "Data Preparation for Data Mining," *Applied Artificial Intelligence*, special issue on data cleaning and preprocessing, vol. 17, nos. 5-6, 2003, pp. 375-382.