Editorial Notes
Classification and Assessment of Large Amounts of Data: Examples in the Healthcare Industry and Collaborative Digital Libraries

New ways of producing and using data provide an opportunity for researchers to seek a revised understanding and unprecedented solutions for analyzing the data. Exchanging and sharing large amounts and diverse uses of data demand an efficient and effective way of classifying the data. Researchers increasingly seek improved methods for classification and taxonomies for analyzing large amounts of datasets. For example, healthcare, financial industries, and national security areas have critical needs to analyze large amounts of data or what some might call, "big data".

Contemporary information forums, such as digital libraries and wikis, where vast amounts of freely contributed, collaboratively processed, and openly shared information reside, seek effective ways of collecting, processing, analyzing, and using high-quality data.

In this issue, two research articles are showcased that deal with improved ways of managing large amounts of data in the contemporary settings: (1) coded healthcare events and (2) freely-contributed information in collaborative digital libraries. Using the experimental research method, the first article, "Combining Bayesian Text Classification and Shrinkage to Automate Healthcare Coding: A Data Quality Analysis" by Eitel J. M. Lauria and Alan D. March, explores the feasibility of using an automated machine-learning classification method. The second article, "Automatic Assessment of Document Quality in Web Collaborative Digital Libraries" by Daniel H. Dalip, Marcos A. Goncalves, Marcos Cristo, and Pavel Calado, examines an automated machine-learning assessment method of the quality of the content information.

Both articles provide an application and extension of existing theories and methods, yielding new understanding and solutions. The authors pose and answer research questions examined through a data quality lens. Related to the issues of the classifications under study by the authors, Parsons and Wand [2008] stress the need for well-grounded guidelines for identification, evaluation, and selection of "classes" when modeling or designing information system artifacts. Their argument also applies to information and data. They also provide cognitive principles to guide classification for modeling in information systems.

We believe that the insights gleaned from the research results of the two articles can be adapted to provide the next step for both advancing theories and providing useful and effective solutions for researchers and practitioners.

ARTICLE 13

Lauria and March's study focuses on the classification of coded healthcare data. They employ the "shrinkage"-based automatic classification method, with a goal of attaining the quality of classification data at a reasonable level. The authors investigate the feasibility of automated classification of the International Classification of Diseases and Related Health Problems (ICD) coding system, using a statistical classifier trained with data of varying quality. They found that machine learning or automated mining yielded reasonable quality results in classifying coded classes and categories of healthcare data.

The authors' work significantly contributes to potentially overcoming the frequent challenges often experienced in analyzing and managing healthcare data, which

© 2011 ACM 1936-1955/2011/12-ART12 $10.00
DOI 10.1145/2063504.2063505 http://doi.acm.org/10.1145/2063504.2063505

typically involves a large number of class values, a limited amount of training data for experiments, and questionable data quality.

ARTICLE 14

Dalip et al. explore a number of quality indicators for assessing the content quality of articles in digital libraries such as Wikipedia. They use a machine-learning technique to provide an improved assessment method. Using the experimental study method, the authors report that the structure of an article is the most important indicator of article quality in collaborative digital libraries. In assessing quality, the authors propose to treat quality estimation as a regression problem. They estimate the quality of articles in digital libraries as a grade on a continuous quality scale.

This study is an important contribution to research and practice because assessment of the quality of text-based articles is inherently subjective and difficult to automate. If a salient factor such as the structure of the content can be further confirmed, this investigation can be a starting point for future adoption and enhancement.

Both articles have been revised once by the authors since the submission of the original manuscripts.

Researchers are encouraged to extend the results of these studies to evaluate additional datasets and also the data from contemporary and emerging uses. For practitioners, the results of these studies can be applied to develop a pragmatic method in utilizing the algorithm and the methods suggested by the authors in order to find a feasible mechanism for testing out their effective application and ease of implementation in their organizations.

We continue to welcome new and innovative research articles in a broader context of data and information quality. Particularly, we look forward to publishing articles that ask fresh and innovative questions related to quality information in organizations and society, as well as in information systems. Therefore, we seek research that applies theories from diverse disciplinary areas, and we welcome research work that employs both quantitative and qualitative research methods.

We look forward to engaging in discussions with prospective authors and to providing future articles with fresh ideas and useful insights.

REFERENCES

PARSONS, J. AND WAND, Y. 2008. Using cognitive principles to guide classification in information systems modeling. *MIS Quart. 32,* 4.

—Stuart E. Madnick
Yang W. Lee
Editors-in-Chief
http://jdiq.acm.org