Information Processing and Management 51 (2015) 141–143

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Editorial

Introduction to the special issue on "*interactive data analysis*"

The way in which scientific data analysis is carried out has dramatically changed in past years, and data analysis has been recently defined as the *fourth paradigm* in the investigation of nature, after empiricism (describing observed natural phenomena), theory (using models and generalizations) and simulation (simulating complex phenomena) (Bell, Hey, & Szalay, 2009; Hey, Tansley, & Tolle, 2009). Today data analysis is an own e-science that unifies theories, experiments and simulations. In the classical scientific hypothetico-deductive approach (Holzinger, 2011) the scientist asks at first a question, forms a hypothesis, carries out an experiment and collects the data to be analyzed. In eScience it is the reverse, as the data has already been collected and scientists are asking questions to the data. However, whether in astronomy or in life sciences, the increasing flood of data requires sophisticated methods for their handling (Hirsh, 2008). A typical example is the biomedical area, an extreme data-intensive science, where professionals are confronted with huge masses of complex, high-dimensional data sets from diverse sources; in this context *interactive data analysis* is a grand challenge (Holzinger, Dehmer, & Jurisica, 2014). The "*interactive*" part in this context is aimed at supporting professional end-users in learning to interactively analyze information properties, thus enabling them to visualize the relevant parts of their data, possibly via specific visualization techniques (Heer & Shneiderman, 2012; Turkay, Jeanquartier, Holzinger, & Hauser (2014)). In other words, the main aim is to enable an effective human control over powerful machine intelligence, and to integrate statistical methods with information visualization, to support both human insights and decision making processes (Mueller, Reihs, Zatloukal, & Holzinger, 2014). Although we have been in the information processing and management business since more than four decades, we are now facing a tremendous lack of integrated interactive systems, tools and methods. We are still lacking methods that support in the process of interactively finding relevant data – which is essential for sensemaking. It is not sufficient to just deliver to the end user more and more data. Most professionals, such as medical professionals, are in fact primarily *not* interested in data – they are indeed interested in *relevant* information. Hence, finding relevant, usable and useful information within the data to support their knowledge and decision making – this is the grand challenge. Consequently, making data both useful and usable is a major topic for information processing and management.

A big issue is the so called knowledge acquisition bottleneck. Many researchers from the data mining community expected that machine learning techniques would have automated the knowledge acquisition process, thus excluding experts from the process of building models – but the opposite is true, we need the knowledge of the domain expert, hence data mining techniques must put the human intelligence into the loop (Holzinger, 2013). In complex domains we need experts who understand the domain, the problem, and the data sets, hence the context (Berka, Rauch, & Tomecková, 2007). Context is extremely important in big data analytics, as by using context with big data, relationships from unstructured information and related structured data can be derived (Cao et al., 2014; Sokol & Chan, 2013).

Additionally, it is important to acknowledge that decision making is often a team effort, i.e. medical experts rarely work independently in modern hospital settings, but in interdisciplinary and/or multifunctional teams (Gorman et al., 2000), and key for success is their ability to work together to manage information both efficiently and effectively (Reddy & Spence, 2008).

A very closely related issue with interactivity is interpretability: much research in data mining focuses on well-defined metrics such as classifier accuracy, which does not always match the goals of particular data mining tasks. In biomedicine, or customer relationship management, or domains with appropriate regulatory characteristics (where you may need to explain the results of a decision), data mining is often irrelevant if it does not produce results that can be explained to others (Petz et al., 2014).

Some areas of data mining rely heavily on benchmark data sets (see e.g., Kreuzthaler, Bloice, Faulstich, Simonic, & Holzinger, 2011), which allow us to compare results across competing methods. However, benchmarks are a means to an end, not the end in itself. These benchmark problems are intended to be representative of the sorts of problems that our algorithms will see in practice, but what we will see in practice will change over time.

http://dx.doi.org/10.1016/j.ipm.2014.11.002

For sure we can emphasize that the need for interactive data analysis will be rapidly increasing in the future, particularly in integrative and interactive machine learning solutions (Holzinger & Jurisica, 2014).

In this special issue some of the discussed problems are addressed by four papers that have been peer reviewed and revised; the four papers are introduced here below.

Mario Mezzanzanica, Roberto Boselli, Mirko Cesarini and Fabio Mercorio from the Department of Statistics and Quantitative Methods of the Università degli Studi di Milano Bicocca, propose in their paper entitled "A model-based Evaluation of Data Quality Activities in KDD" the Multidimensional Robust Data Quality Analysis. This is a novel domain-independent technique aimed to improve data quality by evaluating the effectiveness of a cleansing function. The authors realized their technique through model checking applied on a weakly structured data set describing the working careers of millions of people. Moreover, the authors made an anonymized version of their dataset and the analysis results publicly available to the community.

Longbing Cao from the Advanced Analytics Institute of the University of Technology in Sydney reports about complex applications such as big data analytics and their involvement of different forms of interactions (coupling relationships) from technical, business (domain-specific) and environmental (including socio-cultural and economic) aspects. Cao emphasizes that such couplings present complexities to learning systems far beyond what have been studied in statistics, mathematics and computer science, such as the typical dependency, association and correlation relationships. There are diverse types and forms of couplings embedded in poorly structured and ill-structured data. Modeling and learning such couplings thus become very fundamental but challenging. Focusing on couplings, this paper discusses the concept of coupling learning, to involve coupling relationships into learning systems. Coupling learning brings great potential in building a deep understanding of the essence of business problem and handling challenges that have not been addressed well by the existing learning theories and tools. Cao verfies his argument by several case studies of coupling learning, including coupling in recommender systems, and incorporating couplings into coupled clustering, coupling document clustering, coupled recommender algorithms and coupled behavior analysis.

Janez Kranjc, Jasmina Smailovic, Vid Podpecan, Miha Grcar, Martin Znidar, and Nada Lavrac describe in their paper "Active Learning for Sentiment Analysis on Data Streams: Methodology and Workflow Implementation in the ClowdFlows", a cloud-based scientific workflow platform, called: ClowdFlows, and its extensions enabling the analysis of data streams and active learning. The authors point out that by utilizing the data and workflow sharing in ClowdFlows, the labeling of examples can be distributed through crowdsourcing. The advanced features of ClowdFlows are demonstrated on a sentiment analysis use case, using active learning with a linear Support Vector Machine for learning sentiment classification models to be applied to microblogging data streams. Basically, sentiment analysis from data streams is aimed at detecting authors' attitude, emotions and opinions from texts in real-time. To reduce the labeling effort needed in the data collection phase, active learning is often applied in streaming scenarios, where a learning algorithm is allowed to select new examples to be manually labeled in order to improve the learner's performance. Even though there are many on-line platforms which perform sentiment analysis, there is no publicly available interactive on-line platform for dynamic adaptive sentiment analysis, which would be able to handle changes in data streams and adapt its behavior over time.

Ines Machado, Ana Gomes, Hugo F. Gamboa, Vítor B. Paixão and Rui M. Costa describe in their paper "Human Activity Data Discovery from Triaxial Accelerometer Sensor: non-supervised learning sensitivity to feature extraction parametrization" a human activity recognition framework based on feature extraction and feature selection techniques where a set of time, statistical and frequency domain features taken from 3-dimensional accelerometer sensors are extracted. This framework specifically focuses on activity recognition using on-body accelerometer sensors. Consequently, the authors present a novel interactive knowledge discovery tool for accelerometry in human activity recognition and study the sensitivity to the feature extraction parametrization.

References

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science, 323*(5919), 1297–1298.

Berka, P., Rauch, J., & Tomecková, M. (2007). Lessons learned from the ECML/PKDD discovery challenge on the atherosclerosis risk factors data. *Computing and Informatics, 26*(3), 329–344.

Cao, L., Joachims, T., Wang, C., Gaussier, E., Li, J., Ou, Y., et al (2014). Behavior informatics: A new perspective. *IEEE Intelligent Systems, 29*(4), 62–80.

Gorman, P., Ash, J., Lavelle, M., Lyman, J., Delcambre, L., Maier, D., et al (2000). Bundles in the wild: Managing information to solve problems and maintain situation awareness. *Library Trends, 49*(2), 266–289.

Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis. *Communications of the ACM, 55*(4), 45–54.

Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: Data-intensive scientific discovery*. Redmond, WA: Microsoft Research.

Hirsh, H. (2008). Data mining research: Current status and future opportunities. *Statistical Analysis and Data Mining, 1*(2), 104–107.

Holzinger, A. (2011). *Successful management of research and development*. Norderstedt: BoD.

Holzinger, A. (2013). Human–computer interaction & knowledge discovery (HCI–KDD): What is the benefit of bringing those two fields to work together? In A. Cuzzocrea, C. Kittl, D. E. Simos, E. Weippl, & L. Xu (Eds.), *Multidisciplinary research and practice for information systems. Springer lecture notes in computer science LNCS 8127* (pp. 319–328). Heidelberg, Berlin, New York: Springer.

Holzinger, A., Dehmer, M., & Jurisica, I. (2014). Knowledge discovery and interactive data mining in bioinformatics – State-of-the-art, future challenges and research directions. *BMC Bioinformatics, 15*(Suppl. 6), I1.

Holzinger, A., & Jurisica, I. (2014). Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. In A. Holzinger & I. Jurisica (Eds.), *Interactive knowledge discovery and data mining in biomedical informatics: State-of-the-art and future challenges. Lecture notes in computer science LNCS 8401* (pp. 1–18). Heidelberg, Berlin: Springer.

Kreuzthaler, M., Bloice, M. D., Faulstich, L., Simonic, K. M., & Holzinger, A. (2011). A Comparison of different retrieval strategies working on medical free texts. *Journal of Universal Computer Science, 17*(7), 1109–1133.

Mueller, H., Reihs, R., Zatloukal, K., & Holzinger, A. (2014). Analysis of biomedical data with multilevel glyphs. *BMC Bioinformatics, 15*(Suppl. 6), S5.

Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stříteský, V., & Holzinger, A. (2014). Computational approaches for mining user's opinions on the Web 2.0. *Information Processing & Management, 50*(6), 899–908.
Reddy, M. C., & Spence, P. R. (2008). Collaborative information seeking: A field study of a multidisciplinary patient care team. *Information Processing & Management, 44*(1), 242–255.
Sokol, L., & Chan, S. (2013). *Context-based analytics in a big data world: Better decisions*. IBM® Redbooks® Point-of-View Publication.
Turkay, C., Jeanquartier, F., Holzinger, A., & Hauser, H. (2014). On Computationally-enhanced visual analysis of heterogeneous data and its application in biomedical informatics. In *Lecture notes in computer science LNCS 8401* (pp. 117–140). Berlin, Heidelberg: Springer.

Andreas Holzinger
Medical University Graz & Graz University of Technology, Austria
E-mail address: a.holzinger@tugraz.at

Gabriella Pasi
Università degli Studi di Milano Bicocca, Milano, Italy
E-mail address: pasi@disco.unimib.it

Available online 8 December 2014