

Schema-less Data-gathering, Integration, and Processing: Towards Support for Knowledge Discovery Process in Heterogeneous Environments

Abraham Bernstein

Dynamic and Distributed Information Systems, University of Zürich, Switzerland

Research question

Sara, a compliance officer in a bank, is analyzing a case flagged as suspicious to inform the authorities of a crime (if necessary) and use the gathered data to refine the fraud-detection system. To that end she needs to access a variety of different data sources such as the transaction-chains connected to the case from the transaction processing systems, information from other intuitions about in-/outgoing transactions, case notes from the police, marketing databases, etc., in order to understand the case's nature and refine the knowledge discovery process used by the automated fraud-detection system.

Peter would like to predict if a new protein might be able to curb the growth of cancer cells. To that end he needs to determine the structure of the protein using NMR spectrometer, process the specimens' gene sequences, combine them with data from a number of protein, cross-reference these with references from Medline, and use the gathered data to build a predictor for the likelihood that the protein will curb cancer-cell growth.

Sara's and Peter's problem is complex and ubiquitous to knowledge-intensive industries. To achieve their goal they need to:

1. Access a number of data-sources, each having its own scheme in different formalisms and gather them for processing
2. Integrate these different data, and generate an interoperable "data-could".
3. Devise a suitable inference experiment to induce/test an inference model.

Indeed, these hurdles reflect the overall knowledge discovery process as outlined Fayad *et al* (1996, see also Figure 1). Unfortunately, however, whilst much work has gone into each of the steps in this process support for the overall process has been largely neglected: enormous advances have been made in the areas such as feature selection, development of advanced induction algorithms, visualizations, data integration techniques, etc. – an integrated environment, that guides real-world users in the overall process from the heterogeneous data sources to the results is still missing.

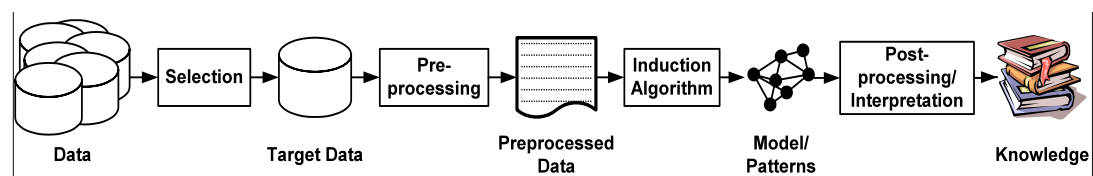


Figure 1: The Knowledge Discovery Process as presented by Fayyad *et al* (1996)

The goal of this study is to present and evaluate the suitability of three technologies to overcome the abovementioned hurdles and serve as the major building blocks for such an integrated environment. Specifically, we propose (i) the use of schema-less, typed graph knowledge bases (Weiss *et al*, 2008; Stocker *et al*, 2008) as an efficient means to store gathered heterogeneous information, (ii) the use of approximate/probabilistic

reasoning techniques on top of these kinds of knowledge bases (Ziegler *et al*, 2006; Kiefer *et al*, 2007) as an appropriate means to address the integration issue, and (iii) the development of advanced Intelligent Discovery Assistants (Bernstein *et al*, 2005; Kalousis *et al*, 2008) to support the design, execution, and exploration of KD processes.

Approach

This study explores the suitability of the three techniques as major building blocks for an integrated knowledge discovery environment. As such it will aim to (i) identify the requirements for each of the building blocks, (ii) assess the suitability of the proposed technologies to fulfill the requirements qualitatively, and (iii) evaluate the technologies quantitatively using benchmarks. As such it represents a typical technical/design science approach.

Main findings/expected contributions

We expect three levels of contributions of this study.

First, the proposition and evaluation of the three major building blocks is a major contribution in itself. Each of them has been the subject of very active research over the last decades and is still very much under investigation. As such, any insight into how the development of each of them can be improved is a contribution in itself.

Second, the investigation of how to combine the components into an integrated system is still in its beginnings. Indeed both the NSF and the EU are funding projects to start exploring such approaches. Furthermore, workshops on the subject are currently being held at major conferences. As such, any insight in how such systems can be developed/deployed would be a substantial contribution.

Lastly, we hope that this study would serve as a call to arms and contribution from the IS research community.

Current status of the manuscript

Whilst some of the studies on the building blocks are very mature (some even published) others are new and evaluated to different degrees. Any statement on the integrated system is very new and not yet evaluated. Input from the community is highly valued at this point in the investigation.

References

- A. Bernstein, F. Provost, S. Hill, "Towards Intelligent Assistance for a Data Mining Process: An Ontology-based Approach for Cost-sensitive Classification", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 4, April 2005, p. 503-518.
- Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, The KDD process for extracting useful knowledge from volumes of data, *Communications of the ACM*, v.39 n.11, p.27-34, Nov. 1996
- A. Kalousis, A. Bernstein, M. Hilario, "Meta-learning with kernels and similarity functions for planning of data mining workflows", *Proceedings of the ICML/COLT/UAI 2008 Planning to Learn Workshop*, Editor(s): Brazdil, Bernstein, Hunter; July 2008.
<http://www.ifi.uzh.ch/ddis/staff/goehring/btw/files/KalousisBernsteinHilario.pdf>

C. Kiefer, A. Bernstein, M. Stocker, "The Fundamentals of iSPARQL - A Virtual Triple Approach For Similarity-Based Semantic Web Tasks", *Proceedings of the 6th International Semantic Web Conference (ISWC)*, 2007; *Lecture Notes in Computer Science*, Springer, p. 295--309.

M. Stocker, A. Seaborne, A. Bernstein, C. Kiefer, D. Reynolds, "SPARQL Basic Graph Pattern Optimization Using Selectivity Estimation", *Proceedings of the 17th International World Wide Web Conference (WWW)*, April 2008, ACM Press, New York, NY, USA.

C. Weiss, P. Karras, A. Bernstein, "Hexastore: Sextuple Indexing for Semantic Web Data Management", *Proc. of the 34th Intl Conf. on Very Large Data Bases (VLDB)*, 2008

P. Ziegler, C. Kiefer, C. Sturm, K. Dittrich, A. Bernstein, "Detecting Similarities in Ontologies with the SOQA-SimPack Toolkit", *10th International Conference on Extending Database Technology (EDBT 2006)*, Editor(s): Ioannidis, Scholl, Schmidt, Matthes, Hatzopoulos, Boehm, Kemper, Grust, Boehm; March 2006; *Lecture Notes in Computer Science*, Vol. 3896, Springer, p. 59--76.