Applications and Evaluation: Overview

Igor Mozetič and Nada Lavrač

Jožef Stefan Institute, Ljubljana, Slovenia

1 Introduction

This part of the book presents several applications which were motivated by the concept of bisociation, and to some extent exploited the notions of heterogeneous information networks, explicit contextualization and/or context crossing.

The main goals of these applications are:

- to verify if the principles of heterogeneous information networks and bisociation, and their computational realization, can lead to new discoveries,
- to test the software platforms developed for bisociative knowledge discovery, and
- to find actual new discoveries in at least some application domains.

Most of the applications are in the area of biology, but in addition there are interesting digressions to finance, improvements of business processes, and music recommendations.

2 Contributions

Eronen at al. [1] discusses Biomine as a BisoNet which integrates heterogeneous biological databases. It consists of over 1 million nodes, representing biological entities (genes, proteins, ontology terms, ...), and over 8 million edges, representing weighted relations of different types. Biomine search algorithms implement link discovery between distant nodes in the graph, and can be exploited for context crossing in bisociative reasoning.

Biomine is an essential component of SegMine, described by Mozetič et al. [8], which implements a form of bisociative reasoning for the analysis of microarray data. SegMine first performs explicit contextualization by subgroup discovery (implemented by the SEGS algorithm), where sets of enriched genes are found. Context crossing is then triggered by queries to Biomine which discovers long range links between distant sets of genes. In the analysis of senescence in human stem cells [10] (not described in this book), a biology expert used SegMine to formulate three new hypotheses which can improve the understanding of the underlying mechanisms in senescence.

A novel application of SegMine, extended to plant biology, is described by Langohr et al. [5]. The problem addressed is the analysis of plant response to a virus attack, from a series of microarray datasets. All human related databases and ontologies in SEGS and Biomine were replaced by plant related data, and

M.R. Berthold (Ed.): Bisociative Knowledge Discovery, LNAI 7250, pp. 359–363, 2012.

subgroup discovery is used again in the later stage of analysis to characterize contrasts between different microarray datasets. The bisociative component in this study consists of the transfer of knowledge about a well-understood plant, namely A. thaliana, to investigate a less well-understood plant, in this case the potato.

Bisociations between related organisms are also exploited in the approach by Kimming and Costa [4]. They investigate metabolic pathways with the goal of automatic pathway curration by link and node prediction, similar to Biomine. However, they make use of information on related organisms to suggest filling of incomplete pathways.

An exploration of textual resources to get an insight into a biological domain is described by Miljković et al. [7]. The ultimate goal is to develop a dynamic model of plant defense to a virus attack. Scientific literature read by domain experts is automatically analyzed to extract triplets of the form <node1, edge, node2> which then form a heterogeneous information network. The network can be explored to find cross-context links between different bodies of human expertise, or eventually (not described in this book) to find novel cross-talk links between different submodels of the plant defense response.

Another analysis of textual resources is described by Schmidt et al. [11]. The goal is to better understand a biological bile acid and xenobiotic system (BAXS) by bisociative hints from a drastically different domain of finance. The idea was to retreive several thousands of scientific papers from both domains, cluster them, and then identify outlier documents. Outliers are biological papers more similar to financial papers than to the rest of biological papers, or vice versa. From the outliers, some interesting bridging terms can be identified which connect the two disparate domains.

An application to business process modelling is presented by Martin and He [6]. The goal is to improve business processes by discovery of process models, their analysis, extension and mining. Process instances concerning repair and call-center data were used to define different contexts, and bisociative reasoning suggested three possible routes which could lead to process improvement.

Finally, Stober et al. [12] present an advanced user interface for music recommendation. Here, the concept of bisociation provides motivation for unexpected and fortunate (serendipitous) recommendations. The article demonstrates how the separation of the similarity measures for projection and distortion makes it possible to link two distinct views on a music collection. As a consequence, it creates a setting where serendipitous recommendations become more likely.

The main conclusions of the applications presented in this volume are the following:

- The concepts of bisociation, explicit contextualization, and context crossing have the potential to help formulate research hypotheses which lead to new discoveries.
- Several software platforms for bisociative reasoning were developed; at least two of them are used regularly by domain experts in human and plant biology (SegMine and Biomine).

– At least in two domains (microarray analysis of human stem cells, autism) biology and medical experts formulated significant new research hypotheses which facilitate novel insights into the domains.

3 Lessons Learned

3.1 The BISON Software for Applications Development

There are numerous lessons learned from the applications described in this volume. In addition to the BISON platform and the software developed within the BISON project, a lesson learned is that in bisociative discovery tasks we were often able to use also other existing software tools, which were used beyond their original scope and purpose. For example, Ontogen [2] is an interactive tool for the construction of topic ontologies, but was used in several applications described in this book for outlier detection and b-term identification. A lesson learned from using Ontogen's similarity graph is, however, that it needs to be used with care. Although two documents appear to be close to each other in the similarity graph, actually they can be distant but at a similar distance from the centroid of the given document cluster. Another tool successfully used in the BISON project was Biomine which was designed for link discovery in biological domains, but in the stem cells microarray analysis its visualization facility enabled the biology expert to identify "gene hubs" (nodes with a large number of edges) and "outlier genes" (nodes with a few edges and of low strength). These concepts are known in social network analysis, but were not exploited in the Biomine context before. Another success, for which Biomine and SEGS were designed but not actually used before, was the relative ease with which we replaced human related databases with plants related databases.

3.2 Application Potential of the BISON Methodology

A major lesson learned from the microarray analysis applications described in this part of the book is that a huge amount of effort is needed to develop a software platform to be used by biology experts. On one hand, the software must match state-of-the-art biological software tools to be competitive. On the other hand, it must address a large number of data management requirements (e.g., different and sometimes inconsistent formats) which are important for routine biological research, but are largely uninteresting and irrelevant from a computer science and knowledge discovery perspective.

The role of bisociations was mainly conceptual but played a crucial role in the formulation of new hypotheses. It made us aware of the need for explicit definition of distinct contexts, for the search of links between them, and for intentional jumps "out-of-the-context". These were initially accomplished by ingenious connection of seemingly unrelated tools (SEGS and Biomine), but later evolved into a novel, interactive, service-oriented platform with a set of SegMine workflows, implemented in a principled way. This led to a natural extension to contrasting coSegMine [5] which opens exciting opportunities for future research.

3.3 Evaluation of the BISON Methodology and the Potential for Triggering Creativity

The problem of cross-context link discovery from scientific papers (presented in Part IV: Exploration) is that in new domains the success criteria are unclear and that only the expert's evaluation is possible. However, in the document analysis case studies on migraine-magnesium and autism-calcineurin [3] this problem did not occur since the task was to evaluate the method by rediscovering known b-terms.

In these cross-context link discovery applications we reused Ontogen in a novel, unforeseen way. The Ontogen approach may not be seen as an approach that triggers creativity, but still it is a useful tool for cross-context discovery. The strongest novelty and lesson learned is that indeed outliers are very useful means of speeding up link discovery in cross-context domains, which was confirmed experimentally in the migraine-magnesium and autism-calcineurin domains [9]. The utility of Ontogen was further proven in the completely new BAXS-finance domain pair as well.

4 The Future of Bisociative Reasoning and Cross-Context Data Mining

Computational creativity community is aware of Koestler's work, but this community can now establish a clear link with the data mining community through the results presented in this book. The BISON project has identified a novel cross-context data mining task which could be of large interest to the data mining community. The investigated research topic, cross-context data mining and knowledge discovery, is not yet part of mainstream data mining research. By further raising awareness of this cross-context/cross-domain knowledge discovery paradigm, the work presented in this book has the potential to ensure that cross-context knowledge discovery will become a recognized topic and thus a first class citizen of major machine learning, data mining and knowledge discovery conferences.

References

1. Eronen, L., Hintsanen, P., Toivonen, H.: Biomine: A Network-Structured Resource of Biological Entities for Link Prediction. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 364–378. Springer, Heidelberg (2012)
2. Fortuna, B., Grobelnik, M., Mladenic, D.: OntoGen: Semi-automatic Ontology Editor. In: Smith, M.J., Salvendy, G. (eds.) HCII 2007. LNCS, vol. 4558, pp. 309–318. Springer, Heidelberg (2007)

3. Juršič, M., Sluban, B., Cestnik, B., Grčar, M., Lavrač, N.: Bridging Concept Identification for Constructing Information Networks from Text Documents. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 66–90. Springer, Heidelberg (2012)
4. Kimmig, A., Costa, F.: Link and Node Prediction in Metabolic Networks with Probabilistic Logic. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 407–426. Springer, Heidelberg (2012)
5. Langohr, L., Podpečan, V., Petek, M., Mozetič, I., Gruden, K.: Subgroup Discovery from Interesting Subgroups. In: Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 390–406. Springer, Heidelberg (2012)
6. Martin, T., He, H.: Bisociation Discovery in Business Process Models. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 452–471. Springer, Heidelberg (2012)
7. Miljković, D., Podpečan, V., Grčar, M., Gruden, K., Stare, T., Petek, M., Mozetič, I., Lavrač, N.: Modelling a Biological System: Network Creation by Triplet Extraction from Biological Literature. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 427–437. Springer, Heidelberg (2012)
8. Mozetič, I., Lavrač, N., Podpečan, V., Novak, P.K., Motaln, H., Petek, M., Gruden, K., Toivonen, H., Kulovesi, K.: Semantic Subgroup Discovery and Cross-context Linking for Microarray Data Analysis. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 379–389. Springer, Heidelberg (2012)
9. Petrič, I., Cestnik, B., Lavrač, N., Urbančič, T.: Bisociative Knowledge Discovery by Literature Outlier Detection. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 313–324. Springer, Heidelberg (2012)
10. Podpečan, V., Lavrač, N., Mozetič, I., Kralj Novak, P., Trajkovski, I., Langohr, L., Kulovesi, K., Toivonen, H., Petek, M., Motaln, H., Gruden, K.: SegMine workflows for semantic microarray data analysis in Orange4WS. BMC Bioinformatics 12, 416 (2011)
11. Schmidt, O., Kranjc, J., Mozetič, I., Thompson, P., Dubitzky, W.: Bisociative Exploration of Biological and Financial Literature Using Clustering. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 438–451. Springer, Heidelberg (2012)
12. Stober, S., Haun, S., Nürnberger, A.: Bisociative Music Discovery & Recommendation. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 472–483. Springer, Heidelberg (2012)