Final report for CENIIT project 02.05 Transparent access to multiple databanks in bioinformatics

Patrick Lambrix

Researchers in various areas, e.g. medicine, agriculture and environmental sciences, use biomedical data sources and tools to answer different research questions or to solve various tasks, for instance, in drug discovery or in research on the influence of environmental factors on human health and diseases. Due to the recent explosion of the amount of on-line accessible data and tools, finding the relevant sources and retrieving the relevant information is not an easy task. Further, often information from different sources needs to be integrated. In this project we have dealt with these problems by investigating a number of issues in relation to search and organization of data on the Web with a focus on biomedical data.

Highlights:

- Winner of Ontology Alignment Evaluation Initiative 2008 Anatomy track.
- Best paper selections.
- Publications in high impact journals in bioinformatics and semantic web fields.
- Production of 2 PhDs.
- Best IDA undergraduate thesis awards.

1 Project Description and Results

The original application focused on defining methods and developing systems for search in multiple data sources. In the early stages we tackled the problem of integrating biological data sources. We studied existing biological data sources regarding their content, data quality, updates, consistency, data models, semantic heterogeneity, and access and retrieval methods [LJ03,Jak05,Jak06]. Based on this we defined requirements for information integration systems in this area and discussed how existing systems conform to these requirements [Jak06]. We also developed the BioTRIFU (*Bio-T*he *R*ight *I*nformation *F*or yo*U*) system that addresses some of these requirements [LJ03,Jak05]. Further, we investigated the use of ontological information in data source integration [JL05,DJLSW06,LS07].

In the latter stages we have re-focused the project into two parts. The first part generalizes the original setting to data sources in a Semantic Web. The Semantic Web can be seen as an extension of the current Web in which information is given a well-defined meaning by annotating Web content with ontology terms. This part of the project was related to our cooperation in the EU Network of Excellence REWERSE. A vision for a Semantic Web in which integration is a major task, is given in [Lam05]. One of the major technologies in the Semantic web are ontologies [Lam04,LTJS07] and the investigation on the use of ontological information for data integration is an important aspect. Further, we studied state of the art ontology engineering tools and evaluated them with respect to a number of criteria including functionality, data models, reasoning support, user interface, visualization, customization and extendibility [LHP03]. We also evaluated two of the main ontology merging tools [LE03]. These evaluations were among the first in its kind and have been well cited. Based on the lessons learned in these evaluations we developed different alignment algorithms and SAMBO (*S*ystem for *A*ligning and *M*erging *B*iomedical *O*ntologies), a state-of-the-art system for aligning and merging ontologies [LEMT03,LT05,LT06,CTL06,WTWLS06, LT07,LTL08,LTX08]. In evaluations we have shown that our system performs very well compared to other systems. Further, we also participated in the Ontology Alignment Evaluation Initiative 2007 and 2008 where we focused on the 'anatomy' task. SAMBO performed well in 2007 and won the track in 2008. SAMBO's successor, SAMBOdtf, obtained second place in 2008 [TL07b,LTL08]. In [LT05] we proposed a general framework for aligning and merging ontologies. Most current alignment systems can be seen as instances of our framework. We also showed how the framework can be used to experiment with different alignment strategies and their combinations. Based on our experience using and evaluating SAMBO, we developed a framework for evaluating ontology alignment strategies and their combinations. We also implemented a state-of-the-art tool, KitAMO (Tool*Kit* for *A*ligning and *M*erging *O*ntologies) [LT07,LT08], that is based on the framework and supports the study, evaluation and comparison of alignment strategies and their combinations based on their performance and the quality of their alignments on test cases. It also provides support for the analysis of the evaluation results. Finally, we are one of the first groups to tackle the problem of recommending ontology alignment strategies for a given alignment task [TL07a].

The second part of the project looks at the problem of finding similar data. This is an important basic task in data management, and in

particular in data integration. In this part of the project we proposed a method for similarity-based grouping [JRL06]. Further, we developed the KitEGA (Tool*Kit* for *E*valuation *G*rouping *A*lgorithms) framework for evaluating similarity-based grouping strategies and implemented a tool based on the framework [JL07].

Homepages for the different parts of the project:
http://www.ida.liu.se/∼iislab/projects/BioTRIFU/
http://www.ida.liu.se/∼iislab/projects/SAMBO/
http://www.ida.liu.se/∼iislab/projects/KitAMO/
http://www.ida.liu.se/∼iislab/projects/KitEGA/

2 Promotions and Degrees

Patrick Lambrix became professor in 2007.
Two students received a PhD degree during the project period:

Vaida Jakonienė, PhD 2006, Lic 2005

He Tan, PhD 2007, Lic 2006

3 Cooperation

Industry contacts

We have contacts with AstraZeneca (main contacts: Bo Servenius, PhD molecular biology, former chair of the Nordic Society for Bioinformatics and Marcus Bjäreland, head of text mining group). We also developed contacts with the Institute for Infocomm Research, Singapore.

Academic contacts

We were involved in the REWERSE network of excellence (Sixth Framework of the European Union). We collaborated in the 'semantic web for bioinformatics' work package. In this work package there is cooperation between researchers from Bucharest, Edinburgh, Jena, Lisbon, London (City), Manchester, Paris (INRIA), Skövde, Sankt Gallen. The closest cooperation is with the group in Dresden with which we have co-authored several papers.

Other CENIIT projects

We cooperated with Lena Strömbäck's CENIIT project and journal articles and conference articles were published.

4 Use of Resources

The CENIIT funding supported the project leader:
Patrick Lambrix, PhD, docent/professor.
Other people in the project were funded by other sources:
Vaida Jakonienė, PhD student/PhD (supported by CUGS),
He Tan, PhD student/PhD (supported by VR).

5 Project Publications

We only list the publications strictly covered by this CENIIT project.

Journal publications

[JL07] Jakoniene V, Lambrix P, 'A Tool for Evaluating Strategies for Grouping of Biological Data', *Journal of Integrative Bioinformatics*, 4(3):83, 2007.
[LHP03] Lambrix P, Habbouche M, Pérez M, 'Evaluation of ontology development tools for bioinformatics', *Bioinformatics*, 19(12):1564-1571, 2003. *Also best paper selection for and re-publication in the 2005 edition of the Yearbook of the International Medical Informatics Association, pp 547-554.*
[LT06] Lambrix P, Tan H, 'SAMBO - A System for Aligning and Merging Bio-Ontologies', *Journal of Web Semantics, Special issue on Semantic Web for the Life Sciences*, 4(3):196-206, 2006.
[LT07] Lambrix P, Tan H, 'A Tool for Evaluating Ontology Alignment Strategies', *Journal on Data Semantics*, VIII:182-202, 2007.

Book chapters

[Lam04] Lambrix P, 'Ontologies in Bioinformatics and Systems Biology', chapter 8 in Dubitzky W and Azuaje F (ed.) *Artificial Intelligence Methods and Tools for Systems Biology*, pp 129-146, Springer, 2004.
[LT08] Lambrix P, Tan H, 'Ontology alignment and merging', chapter 6 in Burger, Davidson, Baldock (eds) *Anatomy Ontologies for Bioinformatics: Principles and Practice*, pp 133-150, Springer, 2008.
[LTJS07] Lambrix P, Tan H, Jakonienė V, Strömbäck L, 'Biological Ontologies', chapter 4 in Baker, Cheung (eds) *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*, pp 85-99, Springer, 2007.

Magazine publication

[LS07] Lambrix P, Strömbäck L, 'Where is my protein? - Issues in Information Integration', *BIOforum Europe*, 7-8/07:24-26, 2007. Invited contribution. +it Also republished in the 2007 Highlights issue.

Conference and workshop publications

[CTL06] Chen B, Tan H, Lambrix P, 'Structure-based filtering for ontology alignment', *Proceedings of the IEEE WETICE Workshop on Semantic Technologies in Collaborative Applications*, pp 364-369, 2006.

[DJLSW06] Doms A, Jakonienė V, Lambrix P, Schroeder M, Wächter T, 'Ontologies and Text Mining as a Basis for a Semantic Web for the Life Sciences', *Reasoning Web, Second International Summer School*, LNCS 4126, pp 164-183, 2006.

[JL05] Jakonienė V, Lambrix P, 'Ontology-based Integration for Bioinformatics', *Proceedings of the VLDB Workshop on Ontologies-based techniques for DataBases and Information Systems - ODBIS 2005*, pp 55-58, Trondheim, Norway, 2005.

[JRL06] Jakonienė V, Rundqvist D, Lambrix P, 'A method for similarity-based grouping of biological data', *Proceedings of the 3rd International Workshop on Data Integration in the Life Sciences - DILS06*, LNBI 4075, pp 136-151, Hinxton, UK, 2006.

[Lam05] Lambrix P, 'Towards a Semantic Web for Bioinformatics using Ontology-based Annotation', *Proceedings of the 14th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises*, pp 3-7, Linköping, Sweden, 2005. Invited talk.

[LE03] Lambrix P, Edberg A, 'Evaluation of ontology merging tools in bioinformatics', *Proceedings of the Pacific Symposium on Biocomputing - PSB03*, pp 589-600, Kauai, Hawaii, USA, 2003.

[LEMT03] Lambrix P, Edberg A, Manis C, Tan H, 'Merging DAML+OIL bio-ontologies', *Proceedings of the International Workshop on Description Logics*, Rome, Italy, 2003.

[LJ03] Lambrix P, Jakonienė V, 'Towards transparent access to multiple biological databanks', *Proceedings of the First Asia-Pacific Bioinformatics Conference*, pp 53-60, Adelaide, Australia, 2003.

[LT04] Lambrix P, Tan H, 'Merging DAML+OIL Ontologies', *Proceedings of the Sixth International Baltic Conference on Databases and Information Systems*, pp 425-435, Riga, Latvia, 2004.

[LT05] Lambrix P, Tan H, 'A Framework for Aligning Ontologies', *Proceedings of the Third Workshop on Principles and Practice of Semantic Web Reasoning*, LNCS 3703, pp 17-31, Dagstuhl, Germany, 2005.
[LTL08] Lambrix P, Tan H, Liu Q, 'SAMBO and SAMBOdtf results for the Ontology Alignment Evaluation Initiative 2008', *Proceedings of the Third International Workshop on Ontology Matching*, pp 190-198, Karlsruhe, Germany, 2008.
[LTX08] Lambrix P, Tan H, Xu W, 'Literature-based alignment of ontologies'. *Proceedings of the Third International Workshop on Ontology Matching*, pp 219-223, Karlsruhe, Germany, 2008.
[TJLAS06] Tan H, Jakonienė V, Lambrix P, Aberg J, Shahmehri N, 'Alignment of Biomedical Ontologies using Life Science Literature', *Proceedings of the International Workshop on Knowledge Discovery in Life Science Literature*, LNBI 3886, pp 1-17, Singapore, 2006.
[TL07a] Tan H, Lambrix P, 'A method for recommending ontology alignment strategies', *Proceedings of the 6th International Semantic Web Conference*, LNCS 4825, pp 494-507, Busan, Korea, 2007.
[TL07b] Tan H, Lambrix P, 'SAMBO results for the Ontology Alignment Evaluation Initiative 2007', *Proceedings of the Second International Workshop on Ontology Matching*, pp 236-243, Busan, Korea, 2007.
[WTWLS06] Wächter T, Tan H, Wobst A, Lambrix P, Schroeder M, 'A Corpus-driven Approach for Design, Evolution and Alignment of Ontologies', *Proceedings of the Winter Simulation Conference*, pp 1595-1602, 2006. Invited contribution.

PhD theses

[Jak06] Jakonienė V, *Integration of Biological Data*, PhD Thesis 1035, Department of Computer and Information Science, Linköpings universitet, Linköping, 2006.
[Tan07] Tan H, *Aligning Biomedical Ontologies*, PhD Thesis 1110, Department of Computer and Information Science, Linköpings universitet, Linköping, 2007.

Lic theses

[Jak05] Jakonienė V, *A study in integrating multiple biological data sources*, Lic Thesis 1149, Department of Computer and Information Science, Linköpings universitet, Linköping, 2005.

[Tan06] Tan H, *Aligning and Merging Biomedical Ontologies*, Lic Thesis 1225, Department of Computer and Information Science, Linköpings universitet, Linköping, 2006.

MSc and BSc theses

[Xu08] Xu W, LIU-IDA/LITH-EX-A–08/058–SE, SVM-cased algorithms for aligning ontologies using literature.
[Run06] Rundqvist D, LITH-IDA-EX–06/029–SE, Grouping biological data. *Best Undergraduate Thesis Award 2006 at the Department of Computer and Information Science, Linköpings universitet.*
[Che06] Chen B, LITH-IDA-EX–06/019–SE, Structure-based ontology alignment.
[Zha06] Zhang P, LITH-IDA-EX–06/018–SE, A user interface for SAMBO using ontology visualization tools.
[Che04] Chétrit H, LiTH-IDA-EX-04/017-SE, A Tool for Facilitating Ontology Construction from Texts.
[Tan03] Tan H, LiTH-IDA-Ex-03/51-SE, Merging DAML+OIL Bio-Ontologies.
[Her03] Hernandez Lopez R, LiTH-IDA-Ex-03/27, Semi-automatic wrapper generation for biological databanks.
[Bre02] Bresell A, LiTH-IDA-Ex-02/76 (AstraZeneca, Lund), Interpretation of microarray expression data using ontology browsing.
[Edb02] Edberg A, LiTH-IDA-Ex-02/62, Förening av ontologier inom bioinformatik. *Best Undergraduate Thesis Award 2002 at the Department of Computer and Information Science, Linköpings universitet.*
[AL04] Abdulahad B, Lounis G, LITH-IDA-EX-ING-04/020-SE, A user interface for the ontology merging tool SAMBO.
[Man03] Manis C, LiTH-IDA-Ex-Ing-02/32, Ontology Merge - a web-based application for merging ontologies.
[HP02] Habbouche M, Pérez M, LiTH-IDA-Ex-Ing-02/10, Utvärdering av ontologiverktyg för bioinformatik.

Proceedings

[Jak03] Jakonienė V, Nilsson R, *Abstract Book of the Fourth Swedish Bioinformatics Workshop for PhD students and PostDocs*, Linköping, Sweden, 2003.