Lecture Notes in Bioinformatics 3886

Edited by S. Istrail, P. Pevzner, and M. Waterman

Subseries of Lecture Notes in Computer Science

Eric G. Bremer Jörg Hakenberg
Eui-Hong (Sam) Han Daniel Berrar
Werner Dubitzky (Eds.)

Knowledge Discovery in Life Science Literature

PAKDD 2006 International Workshop, KDLL 2006
Singapore, April 9, 2006
Proceedings

Springer

Series Editors

Sorin Istrail, Brown University, Providence, RI, USA
Pavel Pevzner, University of California, San Diego, CA, USA
Michael Waterman, University of Southern California, Los Angeles, CA, USA

Volume Editors

Eric G. Bremer
Children's Memorial Hospital, Brain Tumor Research Program
2300 Children's Plaza, Chicago, Illinois 60614, USA
E-mail: egbremer@northwestern.edu

Jörg Hakenberg
Humbold-Universität zu Berlin
Computer Science Department, Knowledge Management in Bioinformatics
Unter den Linden 6, 10099 Berlin, Germany
E-mail: hakenberg@informatik.hu-berlin.de

Eui-Hong (Sam) Han
iXmatch Inc.
5555 West 78th Street Suite E, Minneapolis, MN 55439-2702, USA
E-mail: han@cs.umn.edu

Daniel Berrar
Werner Dubitzky
University of Ulster
School of Biomedial Sciences, Bioinformatics Research Group
Cromore Road, Coleraine BT52 1SA, Northern Ireland, UK
E-mail: {dp.berrar,w.dubitzky}@ulster.ac.uk

Library of Congress Control Number: 2006921543

CR Subject Classification (1998): H.2.8, J.3, I.2, H.3, I.5, I.4, F.1

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743
ISBN-10 3-540-32809-2 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-32809-4 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11683568 06/3142 5 4 3 2 1 0

Preface

This volume of the Springer Lecture Notes in Computer Science series contains the contributions presented at the International Workshop on Knowledge Discovery Life Science Literature 2006 (KDLL 2006) held in Singapore, 9 April 2006, in conjunction with the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2006).

The life sciences encompass research and development in areas such as biology, pharmacology, biophysics, biochemistry, neuroscience, medicine, and environmental sciences. A common theme among life science disciplines is the desire to understand the stimuli-response mechanisms of biological entities, systems, and processes at different levels of organization—from molecules to organisms to ecosystems. As natural phenomena are being probed and mapped in ever-greater detail, life scientists are generating an increasingly growing amount of textual information in the form of full-text research articles, abstracts, Web content, reports, books, and so on. Even in well-focused subject areas it is becoming more and more difficult for researchers and practitioners to find, read, and process all textual information relevant to their tasks. Knowledge discovery in text (KDT) is a fast-developing field that encompasses a variety of methodologies, methods and tools, which facilitate automated processing of text information stored in electronic format. KDT tasks that are particularly interesting to life science include:

- Identification and retrieval of relevant documents from one or more large collections of documents;
- Identification of relevant sections in large documents (passage retrieval);
- Co-reference resolution, i.e., the identification of expressions in texts that refer to the same biological, medical, or biotechnological entity, process, or activity;
- Extraction of life science entities (e.g., genes, proteins, agonists, antagonists, mechanisms, disease, etc.) or relationships (e.g., gene-function, drug–gene interactions, protein–protein interactions, diseases and disease states, etc.) from text collections;
- Automated characterization of biological, biomedical, and biotechnological entities and processes (e.g., annotation of genes or proteins);
- Extraction and characterization of more complex patterns and interaction networks (e.g., biological pathways, topologies, reaction networks, drug-response patterns);
- Automated generation of text summaries;
- Automated construction, expansion, and curation of ontologies for different domains (e.g., characterization of genes, proteins, medical terms);
- Construction of controlled vocabularies from fixed sets of documents for particular domains in biology and medicine.

KDT approaches in the life sciences are faced with a number of challenges that make such endeavors much more complicated than KDT studies in classical application areas such as retail, marketing, customer relationship management, and finance. Important challenges of KDT approaches in biology, biochemistry, biotechnology, medicine, and other life science areas include:

- The need for a mechanistic understanding of biology at different levels of organization and therefore the need for descriptive, predictive as well as for explanatory models;
- The requirement to handle large terminologies characteristics for life science areas. Such terminologies are often redundant, inconsistent, and are constantly evolving;
- The necessity to process and analyze life science texts at different levels of unit granularity, e.g., abstract, full text, passage, section of text such as results, discussion, conclusion sections;
- The management and handling of KDT data and KDT results – this includes the access to and integration of text collections in particular in heterogeneous and distributed computing environments such as the Internet, intranets and grids;
- The complex issue of pre-processing and transforming life science texts using statistical, natural language processing, and other techniques. This also involves issues such as combination of life science text with other forms of data and information, e.g., data from biomedical experiments and information from ontologies, thesauri, dictionaries, warehouses and similar systems;
- The adaptation and improvement of existing and development of new methodologies, algorithms, tools, and systems for different KDT tasks, such as text clustering, classification, entity and relationship extraction, template-based approaches, etc. relevant to life science R&D problems;
- Both the statistical as well as the epistemological (knowledge-based) validation and interpretation of KDT results;
- The constraints posed by computational resources (memory, storage, processor, network bandwidth) arising from large-scale KDT tasks in the life sciences;
- The standardization of KDT approaches in the life sciences.

The objective of the KDLL 2006 Workshop was to bring together scientists who have researched and applied KDT methodologies and techniques in the context of biology, biotechnology, medicine, and other life science areas. The workshop was conceived as a forum facilitating the discussion of innovative work in progress and of important new KDT directions in the life sciences. In addition to life science areas typically associated with bioinformatics (i.e., molecular and cell biology), a specific intention of the workshop was to discuss KDT developments in biochemistry, pharmacology, medicine, neuroscience, environmental sciences, and so on. By sharing the insights, discussing ongoing work and the results that have been achieved, the workshop participants gleaned a comprehensive view of the state of the art in this area and were able to identify emerging and future research issues.

The workshop was structured into a one-day session consisting of two invited talks and 12 presentations of the papers selected for the workshop. Below we briefly summarize the contributions to KDLL 2006.

The contribution of **Tan et al.** addresses the importan t problem of 'aligning' multiple and partially overlapping ontologies. For this they propose an algorithm capable of taking into account ontology structure. **Mathiak et al.** present an interesting picture search engine for life science literature and show how it can be used to improve literature pre-selection. By looking for papers with images (and their textual annotations) concerning the biomedical experiments, they could considerably improve the precision of the retrieval system. **Torii et al.** explore biomedical named entity tagging and evaluate the performance of headwords and suffixes using names from the Unified Medical Language System and incorporating the GENIA ontology. Their study sheds new light on how named entity tagging performs under different conditions and assumptions. **Eom et al.** present a tree kernel-based method to mine protein–protein interactions from text. Their results suggest that this method learns protein interaction information through structure patterns and achieves promising results. **Dimililer et al.** investigate a support vector machine approach to identify and automatically annotate named biomedical entities as an extension of the traditional named entity recognition task to special domains. Specifically, they study the effects of using word formation patterns, lexical, morphological, and surface words for this task. **Huang et al.** explore the problem of discovering potential biomedical relationships from text data. To do so, they follow a study that involves 'temporal topic profiles.' Their approach uses MeSH terms from MEDLINE resources. **Jang et al.** study protein name and protein interaction extraction by using an existing full parser without training or tuning. Their approach is based on a sophisticated substitution-of-words technique and shows that parsing errors can be reduced and parsing precision increased by this sentence simplification method. **Wu et al.** present a robust named entity recognition system based on support vector machines. Testing their system on biomedical data sets, their results show that their approach outperforms relevant competitor methods and, because of its fast execution time, is suitable for real-time applications. **Wang et al.** look at text classification problems involving examples where a small set of labeled positive examples and a very large set of unlabeled examples exist. They present a weighted voting classifier scheme for tackling this problem. **Takeuchi et al.** investigate the problem of mapping different keywords representing the same entity of concept to a canonical form. Such a dictionary with canonical entries may contain many invalid entries. The paper presents methods for detecting invalid entries in such a dictionary. The investigation of **Ning et al.** revolves around automatically filtering and assessing erroneous entries in protein databases. This approach is an important contribution to tackling the problem of handling errors in biomedical databases. **Natarajan et al.** present a download agent and pre-processing tool, which facilitates the task of accessing, downloading and pre-processing full-text articles. Once fully developed, this tool will be useful

for many applications requiring the handling and processing of large full-text research article collections.

We believe that the KDLL 2006 Workshop has made a valuable contribution towards shaping future work in the field of knowledge discovery in life science literature.

Singapore, April 2006

Eric Bremer
Jörg Hakenberg
Eui-Hong (Sam) Han
Daniel Berrar
Werner Dubitzky

Organization

Acknowledgments

KDLL 2006 was sponsored and organized by SPSS Inc.; The DataMiningGrid Consortium (EC grant IST-2004-004475), Northwestern University, Chicago, IL, USA; Humboldt-Universität zu Berlin, Berlin, Germany; iXmatch Inc., Minneapolis, MN, USA; University of Ulster, Coleraine, Northern Ireland, UK. A special thanks goes to the invited speakers—George Karypis, University of Minnesota, MN, USA and Olivier Jouve, SPSS Inc., Paris, France—who reminded us of the mind-boggling breadth and depth of modern life science informatics and the KDT challenges involved. We are indebted to the PAKDD workshop organizers Ah-Hwee Tan, Nanyang Technological University, Singapore and Huan Liu, Arizona State University, USA for their patience and support with all logistical aspects of the workshop. Finally, we would like to extend our gratitude to the members of the KDLL 2006 International Program Committee.

International Program Committee

Jürgen Franke, DaimlerChrysler, Ulm, Germany

Lynette Hirschman, Information Technology Center, The MITRE Corporation, Bedford, MA, USA

Olivier Jouve, SPSS Inc., Paris, France

Min-Yen Kan, National University of Singapore, Singapore

Harald Kirsch, European Bioinformatics Institute, Hinxton, UK

Irena Koprinska, University of Sydney, School of Information Technologies, Sydney, Australia

Patrick Lambrix, Linköpings Universitet, Sweden

Simon Lin, Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL, USA

Michael N Liebman, Windber Research Institute, Windber, PA, USA

Eric Martin, SPSS Inc., Paris, France

Adeline Nazarenko, LIPN Institut Galilée, University of Paris-Nord, Paris, France

See-Kiong Ng, Institute for Infocomm Research, Singapore

Gerhard Paaß, Fraunhofer Institute for Autonous Intelligence Systems, St. Augustin, Germany

Patrick Ruch, University Hospital of Geneve, Switzerland; National Library of Medicine, Bethesda, USA

Alexander K. Seewald, Austrian Research Institute for Artificial Intelligence, Vienna, Austria

Shusaku Tsumoto, Department of Medical Informatics, Shimane Medical University, School of Medicine, Izumo, Japan

Karin Verspoor, Computational Linguist Knowledge and Information Systems Science Team Computer & Computational Science Division, Los Alamos National Laboratory, Los Alamos, NM, USA

Paul van der Vet, Computer Science Department, University of Twente, Enschede, Netherlands

Sponsors

SPSS, Inc.

The DataMiningGrid Consortium

The University of Minnesota, Minneapolis, MN, USA

iXmatch Inc., Minnesota, Minneapolis, MN, USA

Children's Memorial Hospital, Chicago, IL, USA

Northwestern University, Chicago, IL, USA

Humboldt-Universität zu Berlin, Berlin, Germany

The University of Ulster, Coleraine, Northern Ireland, UK

Table of Contents