Lecture Notes in Artificial Intelligence 6581

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Karl Erich Wolff Dmitry E. Palchunov
Nikolay G. Zagoruiko Urs Andelfinger (Eds.)

Knowledge Processing and Data Analysis

First International Conference, KONT 2007
Novosibirsk, Russia, September 14-16, 2007
and First International Conference, KPP 2007
Darmstadt, Germany, September 28-30, 2007
Revised Selected Papers

Springer

Volume Editors

Karl Erich Wolff
Hochschule Darmstadt, Germany
E-mail: karl.erich.wolff@t-online.de

Dmitry E. Palchunov
Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
E-mail: palch@math.nsc.ru

Nikolay G. Zagoruiko
Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
E-mail: zag@math.nsc.ru

Urs Andelfinger
Hochschule Darmstadt, Germany
E-mail: u.andelfinger@fbi.h-da.de

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-22139-2 e-ISBN 978-3-642-22140-8
DOI 10.1007/978-3-642-22140-8
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011930320

CR Subject Classification (1998): I.2.4, I.2, H.2.8, F.4.1, H.3.1-3

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume collects the proceedings of two related international conferences on foundations and practical applications of mathematical methods of data analysis, of Formal Concept Analysis and of methods for information extraction from natural language texts. The first conference, named Knowledge - Ontology - Theory 2007 (KONT 2007), was held during September 14–16, 2007 in Novosibirsk (Russia) at the Sobolev Institute of Mathematics in cooperation with the Russian Foundation for Basic Research and the Association for Pattern Recognition and Image Analysis of the Russian Federation. The second conference, the International Conference on Knowledge Processing in Practice (KPP 2007), was held during September 28–30, 2007 in Darmstadt (Germany) at the University of Applied Sciences in cooperation with the Ernst Schröder Center for Conceptual Knowledge Processing and the Darmstadt University of Technology.

The aim of both conferences was to bring together practitioners and researchers in the interdisciplinary field of mathematical and concept-based knowledge processing. Knowledge processing today spans a broad spectrum of approaches and techniques from data analysis and pattern recognition over artificial intelligence with information retrieval and machine learning to conceptual knowledge processing with its graphical tools for knowledge visualization. From a lifecycle perspective, the field of knowledge processing covers the following main stages: data collection and pre-processing, discovery of regularities, creation of subject domain theories, application of formal knowledge structures, formal reasoning and interpretation in the application domain.

At both conferences, particular emphasis was placed on the technology and experience transfer aspects between practitioners and academic researchers in order to foster mutual learning and application-oriented research and development in the area of knowledge processing based on real empirical needs.

The contributions were all refereed and the accepted papers are collected in this volume. They cover four main focus areas which should, however, not be understood as mutually exclusive. Rather, all contributions share the common foundation that conceptual structures are essential to semantically meaningful and valid representation and processing of formal knowledge structures. The main focus areas are as follows: I: Applications of Conceptual Structures, II: Concept-Based Software, III: Ontologies as Conceptual Structures and IV: Data Analysis.

Part I: Applications of Conceptual Structures

This part on applications of conceptual structures brings together contributions that cover the history of conceptual knowledge processing (Wille) and conceptual extensions of Rough Set Theory (Ganter) as well as contributions from

modal logics (Shilov, Garanina) and applications of Temporal Concept Analysis (Wolff) for gene expression data of arthritic patients (Wollbold et al.) and information retrieval in science (Ponomaryov et al.). In most contributions, practical applications and experiences are also shown.

R. Wille outlines the vast experience base and major development steps that have been achieved since 1980 in the Darmstadt Research Group on Concept Analysis. The focus is on the application side and on historically reconstructing the evolutionary path of extending the mathematical theory of Formal Concept Analysis by a collection of methods for conceptual knowledge processing.

B. Ganter presents an application of Formal Concept Analysis in Rough Set Theory, introduced by Pawlak (1982), who based this theory on the idea of an approximation of a set by equivalence classes of an equivalence relation, called indiscernibility relation, on a given universe U of objects. Ganter generalizes this idea by introducing a preorder (reflexive and transitive) on the set U. The conceptual representation shows that this preorder expresses that one object concept is a subconcept of another one. The corresponding definable sets are the preorder ideals. They form a distributive lattice as in the case of the indiscernibility relation. Connections to functional dependencies, linguistic variables, and decision making are mentioned.

S.O. Kuznetsov presents an innovative algorithm which constructs the lattice of graph sets in a top-down way, starting from the smallest subgraphs of the graphs in a dataset. In contrast to previous bottom-up algorithms this approach allows one to stop at an appropriate level of approximation (projection), thus saving much time and space and making the computational complexity more tractable.

N.V. Shilov and N. Garanina present a summary of recent studies of the model-checking problem for certain combinations of propositional logics for knowledge, time, and actions. After a short overview of these logics they provide a brief introduction to modal logics and Kripke models using the example of Elementary Propositional Dynamic Logic (EPDL). Then they describe propositional logics for epistemic agents, branching temporal logic with actions, combined logics of knowledge, actions, and time, and finally model checking problems in combined logics.

K.E. Wolff presents new results in Temporal Concept Analysis. He starts with an example of a moving high-pressure zone and explains the basic notions in Temporal Conceptual Semantic Systems referring to this example. One of the main results is the purely conceptual introduction of the notion of a state of an object at some time granule (with respect to some view and some selection). The states are special cases of traces of an object. These traces generalize the notion of a volume of an object in some space. In practical applications the used semantic scales are a valuable tool for the representation of a suitable granularity. This is shown by an application in the chemical industry setting where the behavior of a distillation column over 20 days is represented by a lifetrack in a nested line diagram visualizing four many-valued attributes. The same technique is also used in the next article in the field of biomedicine.

J. Wollbold, R. Huber, R.W. Kinne, and K.E. Wolff present a cooperation between scientists from medicine, biology, and mathematics. They investigate disease processes of rheumatoid arthritis using time series of gene expression data. For the purpose of understanding their complicated temporal data they represent these data as a Temporal Conceptual Semantic System as introduced by Wolff (in the previous article). The application of transition diagrams which represent simultaneously the movement of several patients in some genetic space, represented by a concept lattice, yields valuable insight into genetic processes, for example, by finding new hypotheses concerning gene regulation. By applying concept-based analysis techniques the article points to innovative directions for research compared with the current state of the art in this field of biomedicine.

D. Ponomaryov, N. Omelianchuk, V. Mironova, E. Zalevsky, N. Podkolodny, E. Mjolsness, and N. Kolchanov present an innovative approach to automate the extraction process of formal knowledge structures from scientific publications. The goal is to contribute to a better exploitation and more comprehensive usage of already published research results in a specific scientific domain. The value of this work lies in accelerating the formation of hypotheses and theories, which is demonstrated with a concrete example in the field of botany.

Part II: Concept-Based Software

This part on concept-based software focuses around methodical challenges and approaches for supporting software development and software usage with concept-based tools and techniques.

A.S. Kleshchev explores the potential of ontologies as conceptual structures for supporting domain analysis and simulation in software development. He extends the approach also to technical applications, e.g., in optimizing compilers based on ontologies in formal knowledge structures.

G. Stumme discusses problems arising from the empirical success of social bookmarking. While social bookmarking allows for a bottom-up approach in jointly organizing and sharing knowledge assets and bookmarks on the Internet, users are increasingly getting lost in the tons of available meta-information that they are collectively generating. The author presents recent empirical research results that might help users in keeping control over the mass of poorly structured information elements in social bookmarking systems.

U. Priss and J. Old explore new ways of reducing large formal data structures in the field of lexical databases so that they are better understandable for humans. The challenge here is to keep as much as possible of all the relevant semantics of a specific data search while omitting as much as possible of the non-relevant or empty search results. To this purpose they borrow the term ‘weeding’ from biology: with every search they want to preserve as much of the valuable ingredients while collecting as few weeds as possible. They also present practical implementations for their ideas.

Part III: Ontologies as Conceptual Structures

This part on ontologies as conceptual structures focuses on developing taxonomies and conceptual structures, which are increasingly called 'ontologies,' for given problem domains. The aim is to support meaningful knowledge representation, knowledge communication and knowledge retrieval.

D.E. Palchunov describes the idea of a 'Virtual Catalog,' a synthesis of search engines and Internet catalogs. A Virtual Catalog can be considered as an ontology-based technology for information retrieval. To formulate and satisfy the information need of the users, three types of ontologies are created: an ontology of the subject domain of interest, an ontology for various types of Internet resources, and an ontology for the types of information search tasks (which is not yet implemented in the described system). The elaboration of the toolkit for the formalization of search queries is a very important part of this research. For that purpose, ontologies are developed as networks of sentences of First-Order-Logic. Two application fields are used to demonstrate the practicality of the approach.

I.L. Artemieva presents experiences and problems with the construction of ontologies for domains with complicated structure—as for example a multilevel ontology for chemistry, which serves as the main example in this paper to introduce a general method for constructing multilevel ontologies. This paper is part of a project of the Presidium of the Russian Academy of Sciences.

Y.A. Zagorulko and O.I. Borovikova describe a comprehensive approach to formalize knowledge structures in the field of humanities using ontologies. They also describe a method for building such an ontology and an ontology description logic, and they have developed an end-user-oriented editor to actively contribute to the further evolution of the ontology. The resulting IT system is accessible via a so-called knowledge portal.

A. Marchuk presents a dynamic information system on historical facts. The unique characteristic for historical facts is that a given topic (information about countries, persons, office documents) evolves over time. What is a historic fact today, is very often a different historic fact tomorrow as new developments have to be added to the fact to keep it up to date. The proposed approach is capable of automatically keep track of dynamic changes to historic facts. The benefit is that we can thus create databases which no longer become obsolete. The approach has been tested in practical applications at the A.P. Ershov Institute of Informatics Systems.

N. Loukachevitch discusses problems that arise when developing ontologies in the field of linguistics and natural languages. In (natural) languages taxonomic relationships cannot always be resolved in simple logical structures. For example, the basic rule "If class A is a subclass of class B, then each instance of class A is also an instance of B" may not be true in many cases for the conceptual structures in natural languages. The paper addresses ways for revealing typical mistaken taxonomic relationships in such situations. The findings have been derived from practical experiences with a large thesaurus and ontology on natural sciences and technologies.

Part IV: Data Analysis

This part on data analysis focuses on mathematical considerations, techniques, and algorithms for automated discovery of knowledge structures, data mining, similarity analysis, and approaches to visualize knowledge structures.

N.G. Zagoruiko introduces a formal representation of an empirical theory and applies it for tasks of several types in data mining, for example, the task of generating a data-dependent partition of a given set. Then the corresponding formal empirical theory represents (all) possible clustering methods as its formal objects, which are formally described in some language V by a set of characteristics (or attributes). Then a test algorithm T constructs for each clustering method a quality value. With respect to some chosen threshold the class of admissible clustering methods is introduced as the set of those clustering methods whose quality value exceeds the chosen threshold. In the second part of the paper the author reports on his approach for the construction of a measure of similarity, namely, his successful Function of Rival Similarity (FRiS). The paper finishes with some ideas concerning a strengthening of the empirical Data-Mining Theory.

I.A. Borisova, V.V. Dyubanov, N.G. Zagoruiko, and O.A. Kutnenko presents practical applications of the Function of Rival Similarity (FriS) to basic tasks of data mining. They first demonstrate the use of the FRiS-function for selecting typical representatives (stolps) of classes, which will be used for the recognition of new objects. Then, they demonstrate the application of FRiS to several tasks, for example, the tasks of type SDX, i.e., to the simultaneous construction of a classification (task S) of observable objects, building decision rules (task D) and the selection of informative subsets of attributes (task X). It is also shown that the FriS-based algorithms are invariant to the ratio of the number of objects to the number of attributes in the dataset and to the type of the probability distribution of the samples.

V.B. Barakhnin, V.A. Nekhayeva, and A.M. Fedotov analyze and compare three algorithms to solve the problem of computerized selection of Internet documents on scientific subjects based on similarity determination. It is concluded that the FRiS-algorithm which is based on Zagoruiko's principle of rival similarity is the best out of the analyzed algorithms, followed by the greedy algorithm.

E. Vityaev and S. Smerdov consider predictions provided by Inductive Statistical (I-S) inference which are, by Hempel, statistically ambiguous. To avoid this ambiguity, Hempel introduced the Requirement of Maximum Specificity (RMS). Vityaev and Smerdov introduce maximum specific (MS) rules and prove that they satisfy the RMS. The authors also prove that I-S inferences using MS rules avoid the problem of statistical ambiguity. To introduce a probability of events and sentences they use the product probability as in the case of statistical independence; using this product probability they introduce the notion of a probabilistic law and study semantic probabilistic inference, probabilistic Herbrand models, and predictions based on semantic probabilistic inference. Finally, they shortly describe a computer program, DISCOVERY, based on semantic probabilistic inference.

B. Kovalerchuk and A. Balinsky propose a novel approach for analyzing multidimensional data. They focus on the observation that often a human user can easily catch a border between patterns visually, but its analytical form can be quite complex to describe and difficult to discover by formal algorithms. The paper describes a new technique for extracting patterns and relations visually from multidimensional data using the process of data monotonization that gives the opportunity to use the theory of monotone Boolean and k-valued functions. The major novelty of the approach is in visualization of structural relations between n-dimensional objects instead of traditional attempts to visualize each attribute value of n-dimensional objects. Experiments with breast cancer data show the advantages of this approach in uncovering a visual border between benign and malignant cases in breast cancer diagnostics.

G. Lbov and V. Berikov present an algorithm for event tree construction on the basis of an analysis of expert knowledge and multivariate time series. The algorithm uses the Bayesian criterion for decision tree pruning. It can be used for the analysis of extreme events in the conditions of high a priori uncertainty about them. In this case an expert can contribute additional statistical information concerning the object under investigation. Experiments with artificial and real data sets confirm the usefulness of the proposed algorithm.

November 2010

Dmitry E. Palchunov
Nikolay G. Zagoruiko
Karl Erich Wolff
Urs Andelfinger

KONT 2007 Organization

Executive Committee

Conference Committee	
Conference Chair	Y.I. Zhuravlev, Full Member of the Russian Academy of Sciences
Co-chairs	D.E. Palchunov and N.G. Zagoruiko
Program Committee	Ablamejko S.V. (Minsk, Belarus) Cheremisina E.N. (Dubna, Russia) Fedotov A.M. (Novosibirsk, Russia) Gavrilova T.A. (St. Petersburg, Russia) Gladun V.P. (Kiev, Ukraine) Ganter B. (Dresden, Germany) Herre H. (Leipzig, Germany) Kleshchev A.S. (Vladivostok, Russia) Kolchanov N.A. (Novosibirsk, Russia) Kovalerchuk B. (Seattle, USA) Kuznetsov S.O. (Moscow, Russia) Lbov G.S. (Novosibirsk, Russia) Marchuk A.G. (Novosibirsk, Russia) Mihalskij R.S. (Washington, USA) Samokhvalov K.F. (Novosibirsk, Russia) Sviridenko D.I. (Moscow, Russia) Tselishchev V.V. (Novosibirsk, Russia) Vasiljev S.N. (Moscow, Russia) Vityaev E.E. (Novosibirsk, Russia) Wolff K.E. (Darmstadt, Germany) Zelger J. (Innsbruck, Austria)
Organizing Committee	Palchunov D.E., Chair Zagoruiko N.G., Chair Yakhyaeva G.E., Scientific Secretary Borisova I.A . Korovina M.V. Lbov G.S. Pavlovskij E.N. Rjaskin A.N. Salomatina N.V. Vasilenko N.M. Vityaev E.E. Vlasov D.J.

Sponsoring Institutions

Russian Fund of Basic Research
Russian Academy of Natural Sciences
Institute of Mathematics of the Siberian Branch of the
Russian Academy of Sciences
Novosibirsk State University

KPP 2007 Organization

Executive Committee

Conference Committee

Honorary Chair	Rudolf Wille University of Technology, Darmstadt, Germany
Conference Chair	Karl Erich Wolff University of Applied Sciences, Darmstadt, Germany
Co-chair	Urs Andelfinger University of Applied Sciences, Darmstadt, Germany
Program Committee	Bernhard Ganter, University of Technology, Dresden, Germany Wolfgang Hesse, University of Marburg, Germany Sergei O. Kuznetsov, Higher School of Economics, Moscow, Russia Dmitry E. Palchunov, Institute of Mathematics SB RAS, Novosibirsk State University, Novosibirsk, Russia Nikolay V. Shilov, Institute of Informatics Systems, Novosibirsk, Russia Uta Störl, University of Applied Sciences, Darmstadt, Germany Gerd Stumme, University of Kassel, Germany
Organizing Committee	Urs Andelfinger Karl Erich Wolff

Sponsoring Institutions

Darmstadt University of Applied Sciences
Center for Research and Development at
Darmstadt University of Applied Sciences
Mathematics and Science Faculty at Darmstadt University of Applied Sciences
Computer Science Faculty at Darmstadt University of Applied Sciences
Ernst Schröder Center for Conceptual Knowledge Processing

Table of Contents

Part I: Applications of Conceptual Structures

Conceptual Knowledge Processing: Theory and Practice 1
Rudolf Wille

Non-symmetric Indiscernibility 26
Bernhard Ganter

Computing Graph-Based Lattices from Smallest Projections 35
Sergei O. Kuznetsov

Combined Logics of Knowledge, Time, and Actions for Reasoning about Multi-agent Systems 48
Nikolay V. Shilov and Natalia O. Garanina

Applications of Temporal Conceptual Semantic Systems 59
Karl Erich Wolff

Conceptual Representation of Gene Expression Processes 79
Johannes Wollbold, René Huber, Raimund Kinne, and Karl Erich Wolff

From Published Expression and Phenotype Data to Structured Knowledge: The Arabidopsis Gene Net Supplementary Database and Its Applications 101
Denis Ponomaryov, Nadezhda Omelianchuk, Victoria Mironova, Eugene Zalevsky, Nikolay Podkolodny, Eric Mjolsness, and Nikolay Kolchanov

Part II: Concept-Based Software

How Can Ontologies Contribute to Software Development? 121
Alexander S. Kleshchev

A Comparison of Content-Based Tag Recommendations in Folksonomy Systems 136
Jens Illig, Andreas Hotho, Robert Jäschke, and Gerd Stumme

Data Weeding Techniques Applied to Roget's Thesaurus 150
Uta Priss and L. John Old

Part III: Ontologies as Conceptual Structures

Virtual Catalog: The Ontology-Based Technology for Information Retrieval 164
Dmitry E. Palchunov

Ontology Development for Domains with Complicated Structures 184
Irina L. Artemieva

Technology of Ontology Building for Knowledge Portals on Humanities 203
Yury Zagorulko and Olesya Borovikova

Methods and Technologies of Digital Historical Factography 217
Alexander Marchuk

Establishment of Taxonomic Relationships in Linguistic Ontologies 232
Natalia Loukachevitch

Part IV: Data Analysis

Problems in Constructing an Empirical Theory of Data Mining 243
Nikolay G. Zagoruiko

Use of the FRiS-Function for Taxonomy, Attribute Selection and Decision Rule Construction 256
Irina A. Borisova, Vladimir V. Dyubanov, Olga A. Kutnenko, and Nikolay G. Zagoruiko

Similarity Determination for Clustering Textual Documents 271
Vladimir Barakhnin, Vera Nekhaeva, and Anatolii Fedotov

On the Problem of Prediction 280
Evgenii Vityaev and Stanislav Smerdov

Visual Data Mining and Discovery in Multivariate Data Using Monotone n-D Structure 297
Boris Kovalerchuk and Alexander Balinsky

Construction of an Event Tree on the Basis of Expert Knowledge and Time Series 314
Gennady Lbov and Vladimir Berikov

Author Index 321