THE 18TH EUROPEAN CONFERENCE ON MACHINE LEARNING
AND
THE 11TH EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICE
OF KNOWLEDGE DISCOVERY IN DATABASES

PROCEEDINGS OF THE WORKSHOPS:

PRIOR CONCEPTUAL KNOWLEDGE IN MACHINE LEARNING AND DATA MINING AND WEB MINING 2.0

PriCKL'07 & Web Mining 2.0

September 21, 2007

Warsaw, Poland

Editors:
Bettina Berendt
Institute of Information Systems, Humboldt University Berlin, Germany
Dunja Mladenič
J. Stefan Institute, Ljubliana, Slovenia
Giovanni Semeraro
Department of Informatics, University of Bari, Italy
Myra Spiliopoulou
Faculty of Computer Science, Otto-von-Guericke-Univ. Magdeburg, Germany
Gerd Stumme
Knowledge and Data Engineering Group, University of Kassel, Germany
Vojtěch Svátek
University of Economics, Prague, Czech Republic
Filip Železný
Czech Technical University, Prague, Czech Republic

Typesetting:
Bettina Berendt

Preface

Introduction: PriCKL and Web Mining 2.0

This proceedings volume comprises the papers of two workshops held at ECML/PKDD 2007: PriCKL (Prior Conceptual Knowledge in Machine Learning and Knowledge Discovery) and Web Mining 2.0. Our own prior work (including the joint proceedings volume *Semantics, Web and Mining, Joint International Workshops, EWMF 2005 and KDO 2005* published as Springer LNAI 4289) made us expect that there would be areas of joint interest in these two workshops. This year's accepted papers and the topics of the invited talks corroborated this expectation. We therefore decided to partially merge the workshops themselves (by a joint session) and to fully merge the proceedings.

PriCKL : There is general agreement that the quality of ML and KDD output strongly depends not only on the quality of source data and sophistication of learning algorithms, but also on additional, task/domain specific input provided by domain experts for the particular session. There is however less agreement on whether, when and how such input can and should effectively be formalised and reused as explicit prior knowledge. In this workshop, we aimed to investigate current developments and new insights on learning techniques that exploit prior knowledge and on promising application areas. With respect to application areas, we invited – and received – in particular papers on bioinformatics / medical and Web data environments.

The workshop is part of the activities of the "SEVENPRO – Semantic Virtual Engineering for Product Design" project of the European 6th Framework Programme.

Web Mining 2.0 : The workshop "Web Mining 2.0" has been motivated by the specification of Web 2.0. We observe Web 2.0 as a powerful means of promoting the Web as a social medium, stimulating interpersonal communication and fostering the sharing of content, information, semantics and knowledge among people. The workshop hosts research on the role of web mining in and for the Web 2.0.

The workshop is part of the activities of the working groups "Ubiquitous Data – Interaction and Data Collection" and "Human Computer Interaction and Cognitive Modelling" of the Coordination Action "KDubiq – Knowledge Discovery in Ubiquitous Environments" of the European 6th Framework Programme.

PriCKL sessions

The contributions to PriCKL fell into four groups. The first three were ILP/MRDM and an application focus on bioinformatics; the role of the human user; and investigations of fully automated methods of integrating background knowledge. The last group focused on the use of background knowledge for Web mining; these papers were presented in the joint session of PriCKL and WebMining 2.0.

An overview of both the application area bioinformatics and computational techniques used in it was given by our first Invited Speaker, Stephen Muggleton, in his

talk on *Using prior knowledge in biological pathway discovery*. An important group of techniques (not only) in this domain are ILP/MRDM methods. *On Ontologies as Prior Conceptual Knowledge in Inductive Logic Programming* by Francesca A. Lisi and Floriana Esposito provides an overview of Inductive Logic Programming attempts at using Ontologies as prior conceptual knowledge. Specifically, they compare the proposals CARIN-ALN and AL-log. *Using Taxonomic Background Knowledge in Propositionalization and Rule Learning* by Monika Žáková and Filip Železný exploit explicit term and predicate taxonomies to improve relational learning. They speed up the process of propositionalization of relational data substantially, by exploiting such ontologies through a novel refinement operator used in the construction of conjunctive relational features. Subsequent search is also shown to profit from the taxonomic background knowledge.

Two contributions emphasize the role of the human expert in contributing background knowledge for data mining quality: In *A Knowledge-Intensive Approach for Semi-Automatic Causal Subgroup Discovery*, Martin Atzmueller and Frank Puppe present a method for identifying causal relations between subgroups to form a network of links between subgroups. Background knowledge is used to add missing links in this network, correct directionality, and remove wrong links. Their approach is semi-automatic: the network and the relations are visualized to allow a user to accept them into a final causal model or not. An example case study illlustrates how data mining can help to identify risk factors for medical conditions. In *Evaluation of GUHA Mining with Background Knowledge*, Martin Ralbovský evaluates results of the GUHA method (General Unary Hypotheses Automaton), one of the oldest data mining methods for hypothesis generation, against rules formulated by human domain experts. The rules concern relationships between health-related behavioural variables. He concludes that the semantics of the quantifiers need to be worked on, and the default quantitative parameters adjusted to the domain.

Fully automated uses of background knoweldge and their advantages, including speed and accuracy, are investigated with respect to different types of machine learning in four contributions. *A study of the SEMINTEC approach to frequent pattern mining* by Joanna Józefowska, Agnieszka Lawrynowicz, and Tomasz Lukaszewski describes an experimental investigation of various settings under an approach to frequent pattern mining in description logics (DL) knowledge bases. Background knowledge is used to prune redundant (partial) patterns, which substantially speeds up pattern discovery. *Quantitative association rule mining in genomics using apriori knowledge* by Filip Karel and Jiří Kléma addresses the problem of mining high-dimensional, quantitative, and noisy data like transcriptomic data. The quantitative AR approach is based on simple arithmetic operations with variables and it outputs rules that are syntactically like classical association rules. They use prior knowledge (expressed, for example, in a gene similarity matrix) to find promising rule candidates, thus pruning the search space and reducing the number of derived rules. *Conceptual Clustering Applied to Ontologies by means of Semantic Discernability* by Floriana Esposito, Nicola Fanizzi, and Claudia d'Amato proposes a way of clustering objects described in a logical language. The clustering method relies on a semi-distance measure and combines bisecting k-means and medoids into a hierarchical extension of the PAM algorithm (Partition Around

Medoids). *Nonlinear knowledge in learning models* by Mario R. Guarracino, Danilo Abbate, and Roberto Prevete proposes a method to include nonlinear prior knowledge in a Generalized Eigenvalues Support Vector Machine. Prior knowledge here is expressed as additional terms of the cost function of the SVM optimization problem. This improves both algorithmic complexity and prediction accuracy, as shown in a medical case study.

Web Mining 2.0 sessions

The workshop accommodates four papers and one invited talk. In his invited talk *Using context models and models for contextually instantiated social relations for mobile social computing services*, George Groh will discuss services that combine models of social structures and context awareness.

Two of the papers are on the dissemination of semantics in the Web, one of them dealing with automated semantic annotation and the other with the extraction of information from Web documents. The first (*Using Term-Matching Algorithms for the Annotation of Geo-Services* by Grcar and Klien) involves the use of prior conceptual knowledge and is therefore part of the joint session. In the second paper, Raeymaekers and Bruynooghe propose *A Hybrid Approach Towards Wrapper Induction*. They elaborate on wrapper induction for the extraction of information from Web documents: They point out that "tree-based" approaches, which observe a Web document as a tree structure, require less training examples than "string-based" approaches, which treat a document as a string of tokens. To achieve the flexibility and fine granularity possible with string-based approaches, they extend a tree-based wrapper induction method to a hybrid one.

Two further papers of the Web Mining 2.0 workshop deal with the dissemination of preferences about content. The study of Baltrunas and Ricci on *Dynamic Item Weighting and Selection for Collaborative Filtering* investigates the use of item ratings for collaborative filtering in a recommendation engine. The authors study different methods for item weighting and item selection, with the intention to increase recommendation accuracy despite data sparsity and high dimensionality of the feature space. In *Mining Music Social Networks for Automating Social Music Services*, Baccigalupo and Plaza study sequences of music songs, as found in music social networks, and propose a method for the prediction of the most appropriate next song in a playlist.

Joint session PriCKL / Web Mining 2.0

Prior conceptual knowledge has a large importance for the Web. Approaches range from the use of background knowledge (or "semantics") to improve the results of mining Web resources, to the use of background knowledge in mining various other resources to improve the Web. The two contributed papers from PriCKL in the joint session illustrate these two forms.

The Ex Project:Web Information Extraction using Extraction Ontologies by Martin Labský, Vojtěch Svátek, Marek Nekvasil and Dušan Rak addresses the problem of using background knowledge for extracting knowledge from the Web. They use richly-structured extraction ontologies to aid the Information Extraction task. The system also

makes it possible to re-use third-party ontologies and the results of inductive learning for subtasks where pre-labelled data abound.

Dealing with Background Knowledge in the SEWEBAR Project by Jan Rauch and Milan Šimůnek illustrates the use of data mining with background knowledge for creating the Semantic Web. The goal is to generate, in a decentralized fashion, local analytical reports about a domain, and then to combine them, on the Semantic Web, into global analytical reports. Background knowledge is applied on a meta-level to guide the functioning of the mining algorithms themselves (in this case, GUHA). An example of background knowledge are useful category boundaries / granularity for quantitative attributes. The case study concerns relationships between health-related behavioural variables.

In *Using Term-Matching Algorithms for the Annotation of Geo-Services*, Grcar and Klien study semantic annotation of spatial objects: Their objective is to associate terms that describe the spatial objects with appropriate concepts from a domain ontology. Their method achieves this by associating terms with documents fetched from the Web and then assessing term similarity (and term/concept similarity) on the basis of (a) document similarity, (b) linguistic patterns and (c) Google distance.

Research on the Semantic Web in particular and semantic technologies in general continues to profit from the support of the European Union. In his Invited Talk, Stefano Bertolo gives an overview of *EU funding opportunities in research of intelligent content and semantics in Call 3 of Framework Programme 7*.

We thank our reviewers for their careful help in selecting and improving submissions, the ECML/PKDD organizers and especially the Workshops Chairs for their support, our projects SEVENPRO and KDubiq, and the PASCAL project and the Czech Society for Cybernetics and Informatics for sponsoring.

August 2007

The Workshop Chairs:
Bettina Berendt
Dunja Mladenič
Myra Spiliopoulou
Gerd Stumme
Giovanni Semeraro
Vojtěch Svátek
Filip Železný

Workshop Organization

Workshop Chairs

PriCKL'07

Bettina Berendt (Institute of Information Systems, Humboldt University Berlin, Germany)
Vojtěch Svátek (University of Economics, Prague, Czech Republic)
Filip Železný (Czech Technical University, Prague, Czech Republic)

Web Mining 2.0

Bettina Berendt (Institute of Information Systems, Humboldt University Berlin, Germany)
Dunja Mladenič (J. Stefan Institute, Ljubliana, Slovenia)
Giovanni Semeraro (Department of Informatics, University of Bari, Italy)
Myra Spiliopoulou (Faculty of Computer Science, Otto-von-Guericke-Univ. Magdeburg, Germany)
Gerd Stumme (Knowledge and Data Engineering Group, University of Kassel, Germany)

ECML/PKDD Workshop Chair

Marzena Kryszkiewicz (Warsaw University of Technology)

Workshop Program Committees

PriCKL'07

Sarabjot Singh Anand
Martin Atzmueller
Laurent Brisson
Mario Cannataro
Martine Collard
Nicola Fanizzi
Peter Flach
Aldo Gangemi
Marko Grobelnik
Alipio Jorge
Nada Lavrac
Francesca Lisi
Bernardo Magnini
Stan Matwin
Dunja Mladenic
Bamshad Mobasher
Jan Rauch
Massimo Ruffolo
Myra Spiliopoulou
Steffen Staab
York Sure

Web Mining 2.0

Andreas Hotho
Maarten van Someren
Ernestina Menasalvas
Janez Brank
Michelangelo Ceci
Marco de Gemmis
Natalie Glance
Marko Grobelnik
Matthew Hurst
Pasquale Lops
Ion Muslea
Nicolas Nicolov
George Paliouras
Sarabjot Anand

PriCKL'07 Sponsoring Institutions

Czech Society for Cybernetics and Informatics

Web Mining 2.0 Sponsoring Institutions

EU Network of Excellence PASCAL
Pattern Analysis, Statistical Modelling, and Computational Learning

Table of Contents

I PricKL'07 Papers

II Web Mining 2.0 Papers

Part I

PricKL'07 Papers

A Knowledge-Intensive Approach for Semi-Automatic Causal Subgroup Discovery

Martin Atzmueller and Frank Puppe

University of Würzburg,
Department of Computer Science VI
Am Hubland, 97074 Würzburg, Germany
{atzmueller, puppe}@informatik.uni-wuerzburg.de

Abstract. In this paper, we present a methodological approach for knowledge-intensive causal subgroup discovery. We show how to identify causal relations between subgroups by generating an extended causal subgroup network utilizing background knowledge. Using the links within the network we can identify true causal relations, but also relations that are potentially confounded and/or effect-modified by external (confounding) factors. In a semi-automatic approach, the network and the discovered relations are then presented to the user as an intuitive visualization. The applicability and benefit of the presented technique is illustrated by examples from a case-study in the medical domain.

Full article in PDF

EU funding opportunities in research of intelligent content and semantics in Call 3 of Framework Programme 7 *(Invited Talk)*

Stefano Bertolo

European Commission, Brussels, Belgium
Stefano.BERTOLO@ec.europa.eu

Conceptual Clustering Applied to Ontologies by means of Semantic Discernability

Floriana Esposito, Nicola Fanizzi, and Claudia d'Amato

Dipartimento di Informatica – Università degli Studi di Bari
Campus Universitario, Via Orabona 4, 70125 Bari, Italy
{esposito|fanizzi|claudia.damato}@di.uniba.it

Abstract. A clustering method is presented which can be applied to relational knowledge bases to discover interesting groupings of resources through their annotations expressed in the standard languages of the Semantic Web. The method exploits a simple (yet effective and language-independent) semi-distance measure for individuals, that is based on the semantics of the resources w.r.t. a number of dimensions corresponding to a set of concept descriptions (discriminating features). The algorithm adapts the classic BISECTING K-MEANS to work with medoids. A final experiment demonstrates the validity of the approach using absolute quality indices.

Full article in PDF

Nonlinear knowledge in learning models

Mario R. Guarracino[1], Danilo Abbate[1], and Roberto Prevete[2]

[1] High Performance Computing and Networking Institute
{mario.guarracino,danilo.abbate}@na.icar.cnr.it
National Research Council, Italy
[2] University of Naples Federico II
prevete@na.infn.it

Abstract. For most real life problems it is difficult to find a classifier with optimal accuracy. This motivates the rush towards new classifiers that can take advantage of the experience of an expert. In this paper we propose a method to include nonlinear prior knowledge in Generalized Eigenvalues Support Vector Machine. The expression of nonlinear kernels and nonlinear knowledge as a set of linear constraints allows us to have a nonlinear classifier which has a lower complexity and halves the misclassification error with respect to the original generalized eigenvalues method. The Wisconsin Prognostic Breast Cancer data set is used as a case study to analyze the performance of our approach, comparing our results with state of the art SVM classifiers. Sensitivity and specificity results for some publicly available data sets well compare with the other considered methods.

A study of the SEMINTEC approach to frequent pattern mining

Joanna Józefowska, Agnieszka Lawrynowicz, and Tomasz Lukaszewski

Institute of Computing Science, Poznan University of Technology,
ul. Piotrowo 2, 60-965 Poznan, Poland
{jjozefowska, alawrynowicz, tlukaszewski}@cs.put.poznan.pl

Abstract. This paper contains the experimental investigation of various settings under an approach to frequent pattern mining in description logics (DL) knowledge bases that we coined SEMINTEC. Frequent patterns in this approach are the conjunctive queries to DL knowledge base. First, we prove that the approach introduced in our previous publication, for the DLP fragment of DL family of languages, is also valid for more expressive languages. Next, we present the experimental results on knowledge bases of different sizes and complexities.

Full article in PDF

Quantitative association rule mining in genomics using apriori knowledge

Filip Karel and Jiří Kléma

Department of cybernetics, Czech Technical University in Prague,
Technická 2, Praha 6, 166 27
karelf1@fel.cvut.cz, klema@labe.felk.cvut.cz

Abstract. Regarding association rules, transcriptomic data represent a difficult mining context. First, the data are high-dimensional which asks for an algorithm scalable in the number of variables. Second, expression values are typically quantitative variables. This variable type further increases computational demands and may result in the output with a prohibitive number of redundant rules. Third, the data are often noisy which may also cause a large number of rules of little significance. In this paper we tackle the above-mentioned bottlenecks with an alternative approach to the quantitative association rule mining. The approach is based on simple arithmetic operations with variables and it outputs rules that do not syntactically differentiate from classical association rules. We also demonstrate the way in which apriori genomic knowledge can be used to prune the search space and reduce the amount of derived rules.

Full article in PDF

The *Ex* Project: Web Information Extraction using Extraction Ontologies

Martin Labský, Vojtěch Svátek, Marek Nekvasil, and Dušan Rak

Department of Information and Knowledge Engineering,
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
e-mail: {labsky,svatek,nekvasim,rakdusan}@vse.cz

Abstract. Extraction ontologies represent a novel paradigm in web information extraction (as one of 'deductive' species of web mining) allowing to swiftly proceed from initial domain modelling to running a functional prototype, without the necessity of collecting and labelling large amounts of training examples. Bottlenecks in this approach are however the tedium of developing an extraction ontology adequately covering the semantic scope of web data to be processed and the difficulty of combining the ontology-based approach with inductive or wrapper-based approaches. We report on an ongoing project aiming at developing a web information extraction tool based on richly-structured extraction ontologies and with additional possibility of (1) semi-automatically constructing these from third-party domain ontologies, (2) absorbing the results of inductive learning for subtasks where pre-labelled data abound, and (3) actively exploiting formatting regularities in the wrapper style.

Full article in PDF

On Ontologies as Prior Conceptual Knowledge in Inductive Logic Programming

Francesca A. Lisi and Floriana Esposito

Dipartimento di Informatica, Università degli Studi di Bari
Via E. Orabona 4, 70125 Bari, Italy
{lisi, esposito}@di.uniba.it

Abstract. In this paper we provide a survey of Inductive Logic Programming (ILP) attempts at using Ontologies as prior conceptual knowledge. In particular, we take a critical look at two ILP proposals based on knowledge representation frameworks that integrate Description Logics and Horn Clausal Logic and draw from them general conclusions that can be considered as guidelines for an upcoming Onto-Relational Learning aimed at extending Relational Learning to account for Ontologies.

Full article in PDF

Using prior knowledge in biological pathway discovery *(Invited Talk)*

Stephen Muggleton

Imperial College London, London, UK
shm@doc.ic.ac.uk

Abstract. Within the new area of Systems Biologists there is widespread use of graph-based descriptions of bio-molecular interactions which describe cellular activities such as gene regulation, metabolism and transcription. Biologists build and maintain these network models based on the results of experiments in wild-life and mutated organisms. This presentation will provide an overview of recent ILP research in Systems Bioloogy, concentrating on the use of encoded prior knowledge. Indeed one of the key advantages of ILP in this area is the availability of background knowledge on existing known biochemical networks from publicly available resources such as KEGG (used in data sets such as those in the Nature paper by Bryant, King, Muggleton, etc). The topic has an inherent importance owing to its application in biology and medicine. Moreover, descriptions have an inherent relational structure in the form of spatial and temporal interactions of the molecules involved. We will argue the requirements for rich probabilistic logic-based representations which can be machine learned. From a logical perspective the objects include genes, proteins, metabolites, inhibitors and cofactors. The relationships include biochemical reactions in which one set of metabolites is transformed to another mediated by the involvement of an enzyme. One of the representational challenges is that within various databases the same object can be referred to in several ways, which brings in the problem of identity uncertainty. The available genomic information is also very incomplete concerning the functions and even the existence of genes and metabolites, leading to the necessity of techniques such as logical abduction to introduce novel functions and even invention of new objects.

Evaluation of GUHA Mining with Background Knowledge

Martin Ralbovský

Department of Information and Knowledge Engineering,
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
martin.ralbovsky@gmail.com

Abstract. Background knowledge is used for evaluation of specific KDD technique – GUHA method. This is done by verification of verbal background knowledge rules on a medical STULONG dataset. Formalization for the verbal rules was developed and tools for verification of the rules against output of GUHA procedures implemented. We conducted experiments that and drew conclusions about the mostly used settings of GUHA procedures.

Full article in PDF

Dealing with Background Knowledge in the SEWEBAR Project

Jan Rauch and Milan Šimůnek

Faculty of Informatics and Statistics, University of Economics, Prague nám W. Churchilla 4,
130 67 Prague, Czech Republic,
rauch@vse.cz, simunek@vse.cz

Abstract. SEWEBAR is a research project the goal of which is to study possibilities of dissemination of analytical reports through Semantic Web. We are interested in analytical reports presenting results of data mining. Each analytical report gives answer to one analytical question. Lot of interesting analytical questions can be answered by GUHA procedures implemented in the LISp-Miner system. The project SEWEBAR deals with these analytical questions. However the process of formulating and answering such analytical questions requires various background knowledge. The paper presents first steps in storing and application of several forms of background knowledge in the SEWEBAR project.

Full article in PDF

Using Taxonomic Background Knowledge in Propositionalization and Rule Learning

Monika Žáková and Filip Železný

Czech Technical University
Technická 2, 16627 Prague 6, Czech Republic
zakovm1@fel.cvut.cz, zelezny@fel.cvut.cz

Abstract. Knowledge representations using semantic web technologies often provide information which translates to explicit term and predicate taxonomies in relational learning. Here we show how to speed up the process of propositionalization of relational data by orders of magnitude, by exploiting such ontologies through a novel refinement operator used in the construction of conjunctive relational features. Moreover, we accelerate the subsequent search conducted by a propositional learning algorithm by providing it with information on feature generality taxonomy, determined from the initial term and predicate taxonomies but also accounting for traditional θ-subsumption between features. This information enables the propositional rule learner to prevent the exploration of useless conjunctions containing a feature together with any of its subsumees and to specialize a rule by replacing a feature by its subsumee. We investigate our approach with a propositionalization algorithm, a deterministic top-down propositional rule learner, and a recently proposed propositional rule learner based on stochastic local search. Experimental results on genomic and engineering data [**?**] indicate striking runtime improvements of the propositionalization process and the subsequent propositional learning.

Full article in PDF

Part II

Web Mining 2.0 Papers

Mining Music Social Networks for Automating Social Music Services

Claudio Baccigalupo and Enric Plaza

IIIA - Artificial Intelligence Research Institute
CSIC - Spanish Council for Scientific Research
Campus UAB, 08193 Bellaterra, Catalonia (Spain)
Vox: +34-93-5809570, Fax: +34-93-5809661
Email: {claudio,enric}@iiia.csic.es

Abstract. Community-driven services compile data provided by the community members, for instance playlists in Web 2.0 music sites. We show how this data can be analysed and knowledge about sequential associations between songs and artists can be discovered. While most of this kind of analysis focus on (symmetric) similarity measures, we intend to discover which songs can "musically follow" others, focusing on the sequential nature of this data in a database of over 500,000 playlists. We obtain a song association model and an artists association model, we evaluate these models comparing the results with other similarity-based analysis, and finally we show how these models can be used to automatically schedule sequences of songs in a social Web radio service.

Full article in PDF

Dynamic Item Weighting and Selection for Collaborative Filtering

Linas Baltrunas and Francesco Ricci

Free University of Bozen-Bolzano
Domeninkanerplatz 3, Bozen, Italy
{lbaltrunas,fricci}@unibz.it

Abstract. User-to-user correlation is a fundamental component of Collaborative Filtering (CF) recommender systems. In user-to-user correlation the importance assigned to each single item rating can be adapted by using item dependent weights. In CF, the item ratings used to make a prediction play the role of features in classical instance-based learning. This paper focuses on item weighting and item selection methods aimed at improving the recommendation accuracy by tuning the user-to-user correlation metric. In fact, item selection is a complex problem in CF, as standard feature selection methods cannot be applied. The huge amount of features/items and the extreme sparsity of data make common feature selection techniques not effective for CF systems. In this paper we introduce methods aimed at overcoming these problems. The proposed methods are based on the idea of dynamically selecting the highest weighted items, which appear in the user profiles of the active and neighbor users, and to use only them in the rating prediction. We have compared these methods using a range of error measures and we show that the proposed dynamic item selection performs better than standard item weighting and can significantly improve the recommendation accuracy.

Full article in PDF

Using Term-Matching Algorithms for the Annotation of Geo-Services

Miha Grcar[1] and Eva Klien[2]

[1] Jozef Stefan Institute, Dept. of Knowledge Technologies, Jamova 39, 1000 Ljubljana, Slovenia
miha.grcar@ijs.si
[2] Institute for Geoinformatics, Robert-Koch-Str. 26-28, 48149 Münster, Germany
klien@uni-muenster.de

Abstract. This paper presents an approach for automating semantic annotation within service-oriented architectures that provide interfaces to databases of spatial-information objects. The automation of the annotation process facilitates the transition from the current state-of-the-art architectures towards semantically-enabled architectures. We see the annotation process as the task of matching an arbitrary word or term with the most appropriate concept in the domain ontology. The term matching techniques that we present are based on text mining. To determine the similarity between two terms, we first associate a set of documents [that we obtain from a Web search engine] with each term. We then transform the documents into feature vectors and thus transition the similarity assessment into the feature space. After that, we compute the similarity by training a classifier to distinguish between ontology concepts. Apart from text mining approaches, we also present two alternative techniques, namely hypothesis checking (i.e. using linguistic patterns such as "term1 is a term2" as a query to a search engine) and Google Distance.

Full article in PDF

Using context models and models for contextually instantiated social relations for mobile social computing services *(Invited Talk)*

Georg Groh

Technische Universität München, München, Germany
grohg@in.tum.de

Abstract. Social network analysis and models for social structures have gained substantial interest in connection with Web 2.0, communities and other social computing paradigms. While numerous platforms provide means to manage personal social networks of simple kinds, few approaches have been investigated that aim at modelling instantiations of social relations and subsequently using these models for services which are socially- *and* context-aware at the same time. In contrast to the simple models of relations which always represent an average with respect to contextual parameters such as time and space, we will investigate models for describing the instantiations of these relations in time and space and discuss ideas for heuristic methods for identifying these instantiated relations algorithmically.
These models can be used for a broad spectrum of context-aware mobile services in the fields of Contextual Social Awareness, Contextual ocial Recommenders and Information Exchange as well as Context-sensitive Authorization and we will suggest ideas of how such services can be designed to effectively use instantiated social relation models.

A Hybrid Approach Towards Wrapper Induction

Stefan Raeymaekers and Maurice Bruynooghe

K.U.Leuven, Dept. of Computer Science, Celestijnenlaan 200A, 3001 Leuven, Belgium
{stefanr,maurice}@cs.kuleuven.ac.be

Abstract. The approaches to learn wrappers for extraction from semi-structured documents (like HTML documents) are divided into string based ones, and tree based ones. In previous papers we have shown that tree based approaches perform much better and need less examples than string based approaches, but have the disadvantage that they can only extract complete text nodes, whereas string based approaches can extract within text nodes. In this paper we propose a hybrid approach that combines the advantages of both systems. We compare this approach experimentally with a string based approach on some sub node extraction tasks.

Full article in PDF

Author Index