DEPARTMENT OF KNOWLEDGE TECHNOLOGIES E-8

The Department of Knowledge Technologies performs research in advanced information technologies aimed at acquiring, storing and managing knowledge to be used in the development of an information and knowledge-based society. Established areas include intelligent data analysis (machine learning, data mining, and knowledge discovery in databases), semantic data mining and the semantic web, language technologies and computational linguistics, decision support and knowledge management, while computational creativity is a novel research area. Apart from research in knowledge technologies, mostly based on artificial intelligence methods, we are also developing applications in environmental sciences, medicine and health care, biomedicine and bioinformatics, economy and marketing, linguistics and digital humanities.

Head:
Prof. Nada Lavrač

In 2014, we were involved in six national and eleven EU FP7 projects, two COST actions, two bilateral projects, one network financed by the European Science Foundation, one infrastructure project and two industry projects. The department hosts ten junior researchers working towards their PhDs.

In April 2014, the Department joined the Human Brain Project, which is one of the two mega projects funded under the EU FET Flagship Programme

In the area of **intelligent data analysis and data mining,** we have developed several new methods and used them in a number of application domains. We developed a method for semantic and visual explanation of mixture models, and a method for semantic explanation of automatically induced subgroups of data instances – this work was published in the Journal of intelligent information systems. The methodology NoiseRank for ensemble-based noise and outlier detection and the ViperCharts platform for visual results evaluation were published in the prestigious journal Data Mining and Knowledge Discovery. We developed a new methodology for multilayer clustering, which was successfully applied to financial market analysis. A new method for sentiment analysis of streams of Twitter messages was implemented in our web data mining platform ClowdFlows and successfully applied to stock market prediction - this work was published in two journal papers with a high impact factor: Information processing & management and Information sciences. We collaborated with the Department of Intelligent Systems, JSI, in the development of an interactive data mining method HMDM (Human-Machine Data Mining).

We have developed new methods for learning decision trees and ensembles for structured output prediction (multi-target classification, regression and (hierarchical) multi-label classification), and used them for searching and labeling images, modeling of dynamic systems, and for different problems in the area of environmental and life sciences. In particular, we clearly demonstrated the benefit of using a hierarchy in hierarchical multi-label classification, both for community structure prediction and for other tasks. We successfully applied the same approach to habitat modeling for extremophilic fungi, which can act as opportunistic pathogens. Finally, we developed the OntoDM ontology of data mining, which is much deeper and broader in scope in comparison to similar efforts, and describes well the tasks and algorithms for mining structured data.

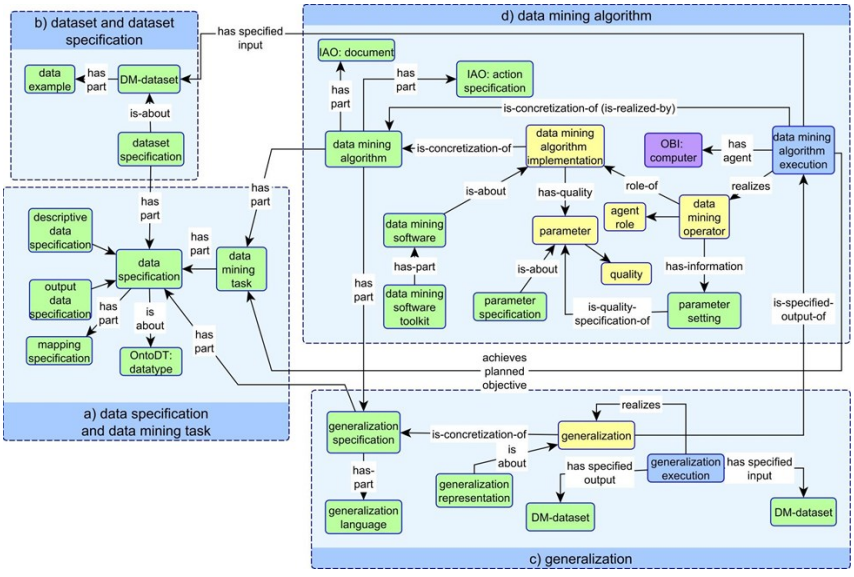

Figure 1: Part of the OntoDM-core ontology of data mining.

We have further developed methods for learning models of dynamical systems from data and domain background knowledge,

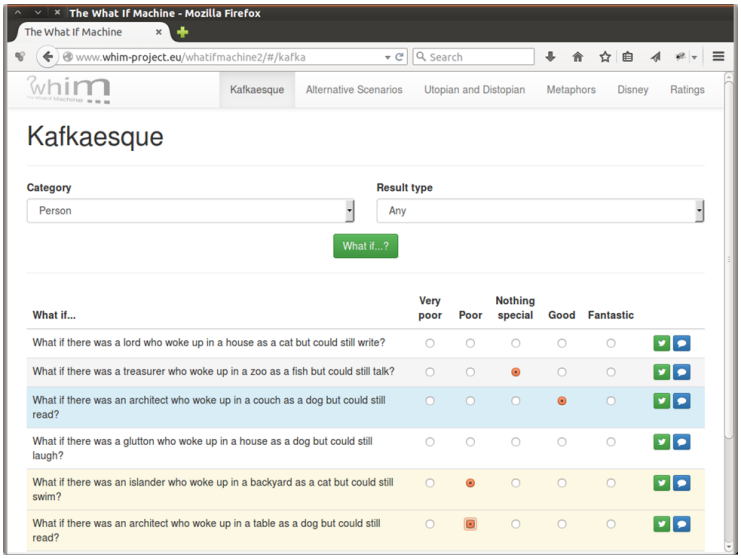

Figure 2: Web interface of "The What-if machine" and its components for crowdsourcing evaluation.

in particular for learning ensembles of such models. This research was part of the FP7 project SUMO (Supermodeling by combining imperfect models), which was successfully completed this year. These methods were also used for modeling of specific aquatic ecosystems and even entire watersheds. The FP7 project REWIRE (Rehabilitative Wayout In Responsive home Environments) was also completed, developing a rehabilitation system for post-stroke patients. Within REWIRE, we applied machine learning methods to analyze patient data and find relationships between patient state at admission and quality of life after rehabilitation (a year later).

We started coordinating the FET project MAESTRA, which addresses the task of analyzing complex data. It targets the breakthrough of developing predictive modeling methods capable of simultaneously addressing different data complexity aspects. These include structured output prediction in the context of massive, networked and incompletely labeled data. We joined the Human Brain Project (FET Flagship), where we take part in the development of the Medical Informatics Platform. The platform will enable the analysis of large quantities of data routinely collected in hospitals for diagnostic purposes. To this end, we are developing methods for rule-based clustering, based on subgroup discovery and predictive clustering, which produce understandable cluster descriptions.

In the area of **computational creativity**, the developed banded matrix approach to compound sentence ideation from two different corpora was evaluated and compared to other fictional narrative ideation approaches. In collaboration with the XLAB company, we have developed a methodology for automatic generation of slogans that is based on a combination of methods from computational linguistics, semantic resources and genetic algorithms.

The "NoiseRank" method for detection and ranking of noise and anomalies in data, was recognized as an outstanding scientific achievement by the Scientific Research Council for Engineering Sciences of ARRS and was presented at the event "Excellent in Science 2013" as part of the 9th Slovenian Innovation Forum 2014.

We are active in the FP7 project MUSE (Machine Understanding for interactive StorytElling) in the area of the computer understanding of natural language, the goal of which is to convert text into a 3D animation. Since 2012 we are involved in the FP7 project PROSECCO, whose goal is the promotion of activities in the area of computational creativity. Since 2013, we collaborate in two computational creativity projects: ConCreTe (Concept Creation Technologies) and WHIM (The What-If Machine); the latter was evaluated as excellent at its first project review. In 2014, we organized the Fifth International Conference on Computational Creativity (ICCC 2014) in Ljubljana, and the Seventeenth International Conference on Discovery Science (DS 2014) in Bled, which attracted about 100 participants each.

In the area of **text and web mining and heterogeneous information network analysis** we successfully concluded the FET project FOC (Forecasting Financial Crises) with an excellent final evaluation. We started work on two new projects, SIMPOL (Financial Systems Simulation and Policy Modelling) and MULTIPLEX (Foundational Research on MULTIlevel COmPLEX networks and systems).

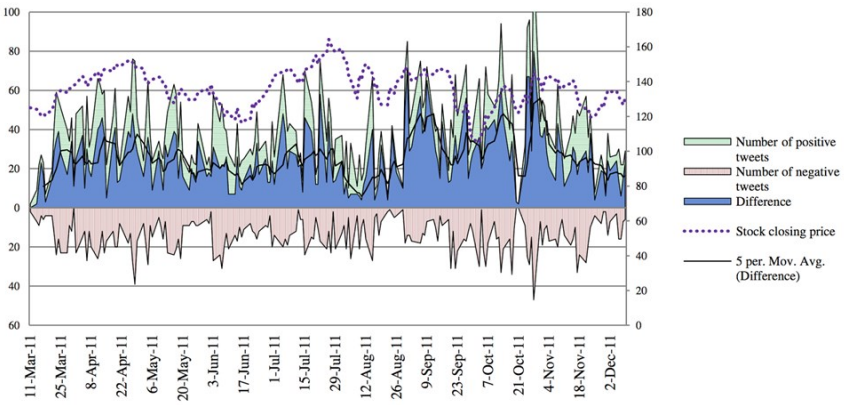

Figure 3: Visual representation of the sentiment of Twitter messages discussing stocks of the Baidu company.

The SIMPOL project aims to support decision makers and regulators in policy modeling and impact analysis, with the emphasis on financing sustainable and environmentally friendly projects. We have developed a knowledge sharing platform to support data collection and analysis. The knowledge representation is based on graphs and on Semantic Web principles. The data model is

stored in Neo4j, a high performance non-SQL database, designed to efficiently manipulate and query graphs. We have implemented the web portal http://simpol.ijs.si/ which uses Neo4j as its back-end. The portal supports activities to build socio-economic networks: overview, selection and inspection of existing networks in the database, extraction of networks from open data, import/export interface to a crowdsourcing platform, monitoring of news and blogs, and social media monitoring.

The goal of MULTIPLEX is the development of a mathematical framework of complex networks and algorithms aimed to establish a theoretical basis for the understanding, prediction and, possibly, the control of complex systems. Our role is to extract multilevel networks from textual data streams, i.e. news, blogs and Twitter. From tweets about environmental issues, we extracted a retweet network, assigned influence to the actors and sentiment to their tweets, and computed major network communities. Based on the sentiment assigned to the communities, we analyzed their leanings towards various environmental issues. From financial news, we created time-varying co-occurrence networks for 50 countries, and compared them to financial, trade, and geographical networks. It turns out that the news mostly reflect geographical proximity. However, positive sentiment news are most similar to the trade networks.

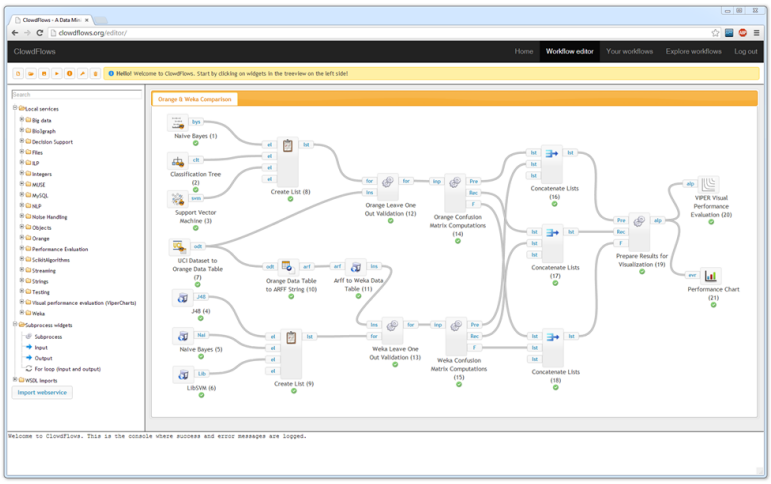

Figure 4: A typical data mining workflow in the web-based ClowdFlows platform.

In 2014, the Department organized three international conferences: the Fifth International Conference on Computational Creativity (ICCC 2014), the Seventeenth International Conference on Discovery Science (DS 2014), and the Ninth Language Technologies Conference at the Information Society Multiconference (LT@IS 2014).

In the area of **language technologies and digital humanities** we work mostly on producing language resources and methods to annotate text with linguistic information, with a focus on the Slovene language. In 2014 we started work on the national research project JANES (Resources, Tools and Methods for Research of Non-standard Internet Slovene), http://nl.ijs.si/janes/, which is led by the Dept. of Translation at the Faculty of Arts at the University of Ljubljana. We have compiled slWaC 2.0, a new corpus of Slovene gathered from Web documents, containing over one billion words. We have also developed a tool, called TweetCaT, for collecting tweets in a given language and a method using character-based statistical machine translation to normalize non-standard language. With these tools we produced the first large linguistically annotated corpus of Slovene tweets, which contains over 35 million words. We have also started work on extracting Slovene user comments, forums and blogs from the Web.

The recently concluded the IMP project, http://nl.ijs.si/imp/, produced a large set of resources of historical Slovene, including a digital library, dictionary and corpus, all available on the Web. We conducted a study to analyze the user experience with these resources, which showed that most users (predominantly teachers and students) consider the resources important for their work and for Slovene linguistics in general and gathered useful suggestions for improvements. We also continued our cooperation with the Slovenian Academy of Sciences and Arts on producing their Web-based Slovene Biographical Lexicon(s).

We were active in developing the concept of a planned new dictionary of Slovene language, in advocating open access to language resources, taken as scientific data, and in establishing the Slovene research infrastructure CLARIN.SI. This infrastructure is now organized as a consortium with, currently, nine partners: three Slovene universities (Ljubljana, Maribor, Primorska), two research institutes (SRC SASA and JSI), as well as societies (SDJT, Trojina, DZDR) and companies (Amebis, Alpineon) dealing with language technologies and resources. The CLARIN.SI infrastructure now meets the organizational and technical criteria so that it could in 2015 become a member of the European CLARIN ERIC, with Slovenia having a stable and permanent repository for language resources and language technology related web services.

In October we organized, in the scope of the JSI Information Society Meta-Conference, the 9th conference on Language Technologies. These biennial conferences have become the main forum for the presentation of research on language technologies and related fields in Slovenia, as well as Croatia. The proceedings of the 2014 conference contains two invited lectures and 29 reviewed regular papers with 180 pages.

We continued our work in the context of the ESF network NetWordS (European Network on Word Structure), in the COST action PARSEME (Parsing and Multi-word Expressions) and started work in the context of two bilateral project, one with Croatia and the other with Serbia. We collaborated in the work of the Slovene Institute of Standardization as the Slovene representatives in ISO/TC37/SC4 (Terminology and Other Language and Content Resources / Language Resources Management) by reviewing, translating and approving Slovene standards from this field. For the Slovenian Ministry of Culture we helped in preparing the Action Plan for the National Program for Language Policy 2014-2018.

In the area of **decision support,** our long-term goal is to develop methods and techniques of decision modeling, support them with software and integrate them with data-mining systems. In 2014, we focused on extending the qualitative multi-attribute method DEX in the directions of including numeric criteria, using probabilistic value distributions and considering relationally-defined decision alternatives, as well as developing corresponding software tools. We updated our computer program for multi-attribute modeling DEXi and added features for the visualization of utility functions. We investigated the potential of aggregation-disaggregation methods UTA and ACUTA for the representation of DEX utility functions. In the application area, we successfully completed two main projects: FP7 FIRST, in which we developed a system for the assessment of bank reputational risk, and a Slovenian project OVJE, in which we developed decision models and a decision support system for the assessment of sustainability of electric energy production in Slovenia until 2030. Other applications include emergency management, strategic management in companies, decision support in protected areas, and production of game devices.

Within the EVADIFF project (Evaluation et de développement et modčles outils d'aide ŕ la décision utilisés pour la Prevention des pollutions diffuses par les produits phytopharmaceutiques), commissioned by ARVALIS Institut du Vegetal, France, we completed the development of a decision-support system for the selection of mitigation measures for the protection of surface waters from pollution by phytopharmaceuticals. The system makes use of qualitative evaluation models based on expert knowledge and quantitative predictive models generated by data mining. After integration, customization and user training, the system will be deployed within ARVALIS during 2015.

IMPORTANT PUBLICATIONS

1. Ikonomovska, E., Gama, J., Džeroski, S. Online tree-based ensembles and option trees for regression on evolving data streams. Neurocomputing, ISSN 0925-2312, 2015, vol. 150, part. 150, pp. 458-470, doi: 10.1016/j.neucom.2014.04.076.
2. Levatić, J., Kocev, D., Džeroski, S. The importance of the label hierarchy in hierarchical multi-label classification. Journal of intelligent information systems, ISSN 0925-9902, [in press] 2014, 25 pp., doi: 10.1007/s10844-014-0347-y.
3. Miljković, D., Depolli, M., Stare, T., Mozetič, I., Petek, M., Gruden, K., Lavrač, N. Plant defence model revisions through iterative minimisation of constraint violations. International journal of computational biology and drug design, ISSN 1756-0756, 2014, vol. 7, no. 1, pp. 61-79, doi: 10.1504/IJCBDD.2014.058588.
4. Panov, P., Soldatova, L. N., Džeroski, S. Ontology of core data mining entities. Data mining and knowledge discovery, ISSN 1384-5810, [in press] 2014, 44 pp., doi: 10.1007/s10618-014-0363-0.
5. Piškorec, M., Antulov-Frantulin, N., Kralj Novak, P., Mozetič, I., Grčar, M., Vodenska I., Šmuc, T. Cohesiveness in Financial News and its Relation to Market Volatility, Scientific Reports 4, 5038-1-8, 2014.
6. Popovič, M., Štefančić, H., Sluban, B., Kralj Novak, P., Grčar, M., Mozetič, I., Puliga, M., Zlatić, V. Extraction of Temporal Networks from Term Co-occurrences in Online Textual Sources, PLoS ONE 9(12), e99515, 2014.
7. Ramšak, Ž., Baebler, Š., Rotter, A., Korbar, M., Mozetič, I., Usadel, B., Gruden, K. GoMapMan: integration, consolidation and visualisation of plant gene annotations within the MapMan ontology, Nucleic Acids Research 42(D1), D1167-D1175, 2014.
8. Sluban, B., Gamberger, D., Lavrač, N. Ensemble-based noise detection: noise ranking and visual performance. Data mining and knowledge discovery, ISSN 1384-5810, 2014, vol. 28, no. 2, pp. 265-303.
9. Smailović, J., Grčar, M., Lavrač, N., Žnidaršič, M. Stream-based active learning for sentiment analysis in the financial domain. Information sciences, ISSN 0020-0255, 2014, vol. 285, pp. 181-203, doi: 10.1016/j.ins.2014.04.034.
10. Vavpetič, A., Podpečan, V., Lavrač, N. Semantic subgroup explanations. Journal of intelligent information systems, ISSN 0925-9902, 2014, vol. 42, no. 2, pp. 233-254, doi: 10.1007/s10844-013-0292-1.