Introduction to RSFDGrC/PReMI-2011 Invited Sessions

This volume contains the extended abstracts of presentations gathered within two invited sessions – *Search and Analytics: Methodologies and Technologies* and *Machine Intelligence and Soft Computing: Expert Foresight* – included into the programmes of the *13th International Conference on Rough Sets, Fuzzy Sets and Granular Computing* (RSFDGrC-2011) and the *4th International Conference on Pattern Recognition and Machine Intelligence* (PReMI-2011), held jointly at the National Research University Higher School of Economics (NRU HSE) in Moscow, Russia, on June 25-30, 2011.

RSFDGrC and PReMI are the two series of scientific events spanning the last 15 and five years, respectively. They investigate the meeting points of six major disciplines outlined in their titles, as well as such areas as *Soft Computing*, *Artificial Intelligence*, *Knowledge Discovery*, *Knowledge Management* and many others. The invited sessions were organized in order to strengthen interactions between researchers and practitioners interested in these major topics, as well as to encourage young people to look more intensively towards interdisciplinary projects related to real-life challenges.

Given a remarkable feedback from the key persons and organizations involved in RSFDGrC-2011 and PReMI-2011, we collected and scheduled 15 talks related to the latest trends in such areas as *Image and Character Recognition*, *Semantic Search and Modeling*, *Modern Database Technologies*, *Decision Support for Commerce* and *Bio-medical Data Analysis*. The volume also includes the abstracts of three PReMI-2011 tutorial lectures.

We are grateful to all speakers who contributed to this volume. We would also like to acknowledge the following organizations and sponsoring institutions: NRU HSE, Laboratoire Poncelet, International Rough Set Society, International Fuzzy Systems Association, Center for Soft Computing Research at Indian Statistical Institute, Russian Foundation for Basic Research, Dynasty Foundation, ABBYY Software House, Yandex and Witology.

June 25, 2011 — Sergei O. Kuznetsov
Moscow, Warsaw — Dominik Ślęzak

Table of Contents

Search and Analytics: Methodologies and Technologies

Entity-oriented Search Result Diversification 1
Andrey Plakhov

High-tech Innovative Appliances (Soft Hardware) for Effective Management of Enterprises and Industries 2
Oleg V. Ena

Semantic Search over Large Repository of Scientific Articles: An Overview of SONCA 4
Marcin Szczuka

Formal Concept Analysis for Knowledge Discovery in Structured and Unstructured Data 5
Sergei O. Kuznetsov

New Features of Infobright's RDBMS: Rough SQL and Seamless Handling of Unstructured Attributes in Machine-generated Data Sets 6
Dominik Ślęzak

Pattern Recognition and Machine Intelligence: Tutorials

Time Series Analysis 8
Roberto Baragona

Problems of Ontology Development for a Broad Domain 9
Natalia Loukachevitch

Recent Advances on Generalization Bounds 10
Konstantin Vorontsov

Machine Intelligence and Soft Computing: Expert Foresight

Reliability Analysis for Aerospace Applications: Reducing Over-Conservative Expert Estimates in the Presence of Limited Data 11
Vladik Kreinovich

Data Analysis Projects for Business: Challenges, Approaches and Cases .. 13
Daniel Kanevskiy

Application of the Integrity, Purposefulness and Adaptivity Principle in OCR and Applied Linguistics 14
Aram Pakhchanian

Image Processing and Analysis Using Soft Computing Tools 15
Malay K. Kundu

New Histogram-like Techniques for Cardinality Estimation in Database Query Optimization 17
B. John Oommen

Research Challenges of Dynamic Socio-Semantic Networks 18
Rostislav Yavorsky

Trust Networks for Recommender Systems 21
Chris Cornelis

Ontology for Multimedia Information Processing 22
Santanu Chaudhury

Gaining Insight into Clinical Pathway with Process Discovery Techniques 25
Jonas Poelmans

Determining Pattern of Variation in Gene Expression Profiles Using Correlation Clustering Algorithms 26
Rajat K. De

Entity-oriented Search Result Diversification

Andrey Plakhov

Yandex, Russia

Abstract. In my talk, I'm going to discuss the approach to search results diversification currently used at Yandex, which is named Spectrum. Large number of queries sent to Yandex are highly ambiguous and mention a specific entity or a class of entities. A query might refer to several objects of the same name (like [apple] might mean either a fruit or a consumer electronics company). More importantly, a query might represent an underlying intent from a large spectrum: e.g. someone searching for [pizza] might want either a restaurant offering delivery service, or a recipe, or even images of pizza.

Spectrum is based on analyzing click-through statistics. The system first identifies objects in queries. Each object is then classified into one or more categories, e.g. "cities", "humans", "cars", "medicines" etc. Based on the object's category, our mined knowledge about typical information needs related to the object and relevant pages available on the Web, Spectrum determines the share of users looking for this object in relation to each of the potential intents. The search engine then uses this information to rank its results for ambiguous queries using the probabilistic model of SERP perception. Target ranking is exactly the one that maximizes the user's chance to find a relevant answer.

High-tech Innovative Appliances (Soft Hardware) for Effective Management of Enterprises and Industries

Oleg V. Ena

Avicomp Services, Moscow, Russia
oleg.ena@avicomp.ru

Today's business environment is characterized by an ever increasing number of news items and pieces of analytical information that need to be processed. Daily volume of a news items is estimated as millions of information articles. A great part of these news items contain valuable facts and primary data that may be analyzed and customized for usage by enterprises and governmental decision-makers. These facts and figures may be indirectly related to a particular field of economic analysis (i.e. merges and acquisitions, new appointments in corporate sector, financial claims, etc.), or be very specific (oil extraction volumes, building of oil derricks for petroleum industry, etc.). Obviously, the greatest daily volume of news comes from the Internet. The Internet has gradually evolved from its initial entertaining function, gaining social significance and becoming an abundant source of information, used by enterprises and their consumers.

One of the major obstacles for effective use of Internet is the absence of software that would allow processing and undertaking semantic analyses of sector-specific texts. ICT studies related to understanding of Internet texts, evolution from Web of documents to Web of data (Web 3.0) is the main concept, on which the Future Internet is based. A special W3C committee, Semantic Web Activity,[1] of the W3C consortium works towards standardization and defining the strategic directions for evolution of Internet. One of them relates to knowledge extraction and classification, which has become rather fashionable worldwide, including the many European programs on the subject.[2]

The main goal of the project outlined in the talk is centered on applying Semantic Web technologies for broad market application, i.e., in-depth analysis of Internet and corporate archives for analytical and strategic analysis at the level of enterprises. Enterprises are interested in constant information accumulation and capturing data, which may be used for increasing their business effectiveness. They are interested in a wide variety of facts and figures: legislation in their spheres of activity, new products appearing on the market, newly registered patents and technologies, news from their competitors (i.e. new sales areas, new contracts, appointments of top management, meetings and negotiations with investors). At the moment there are no solutions on the market to

[1] http://www.w3.org/2001/sw/

[2] See, e.g., Theseus, which is ICT component of FP7.

satisfy this demand. A corporate analyst cannot receive a satisfactory and comprehensive answer through any of existing search engines, or through placing inquiry (search mode) at an IT vendor system (e.g.: ”Tell me about all facts of last year negotiations with the western investors in Russian oil companies”).

The project requires research at the level of computerized extraction of facts, events, data, as well as cause-and-effect relation at micro and micro-economic level with a view to enrich and expand information base for corporate decision-making. For example, extraction should be performed automatically from huge volume of texts (tens of millions of texts) with the use of the newest semantic technologies. In the talk, we focus on describing algorithms, models, methods and tools applied to create an effective combination of the required conceptual and software components, such as:

- The above-mentioned semantic technologies - the support for extraction of knowledge from texts;
- Modern economic models - new models that may to take into consideration new kinds of economic facts of semantic nature;
- Modern mathematical methods - new methods and algorithms that may take into consideration new kind of economic facts of semantic nature.

The soft hardware product will collect and process billions facts, events and the indicators in 24x7 automatic mode that are directly and indirectly related to an enterprise’s economic activity. Data accumulation will occur through all information channels - the Internet, television, documents databases of enterprises, press archives etc. The product will gather resources in different European languages. The accumulated data will be objectively interpreted with the use of new economic models developed by economists.

The innovative application combines economic theory and best practice, serves as an innovative tool for analyzing the external information environment for enterprises. On the one hand, the product shall provide methodical support, and on the other hand, serve as a tool for instant reception of important classified information necessary for speedy decision-making.

Furthermore, classification of billions the economic facts, gathered in line with the verified and tested methodology, will allow, in the long run, feed back into existing economic models and advance them through identifying additional indicators which characterize an enterprise and its development stage and corresponding needs more accurately.

In conclusion, we believe that the methodology and technology resulting from the presented project will allow the enterprises to make grounded economic decisions, taking into consideration a greater number of factors and using the best economic practice.

Semantic Search over Large Repository of Scientific Articles: An Overview of SONCA

Marcin Szczuka

Institute of Mathematics
University of Warsaw
ul. Banacha 2, 02-097 Warsaw, Poland
szczuka@mimuw.edu.pl

Abstract. We outline the architecture of the SONCA system (abbreviation for **S**earch based on **ON**tologies and **C**ompound **A**nalytics) aimed at search and construction of information within document repositories originating from different sources, including libraries and publishers. Documents can be provided in various incomplete formats. The system should be able to use various knowledge bases related to the investigated areas of science. It should also allow for independent sources of information about the analyzed objects, such as, e.g., information about scientists who may be identified as the stored articles' authors.

In particular, we present the relational data schema aimed at efficient storage and querying parsed scientific articles, as well as entities corresponding to authors, institutions, references, scientific concepts, et cetera. An important requirement of the proposed model is to be able to query about all possible entities that may be interesting for users, therefore, to be able to add new types of entities with no need of adding new data tables and increasing the schema's complexity. Another important aspect is to store detailed information about parsed articles in order to conduct analytic operations over large volumes of articles in combination with the domain knowledge about scientific topics, by means of standard SQL and RDBMS management. The overall goal is to support computation of semantic indexes related to entities considered most frequently by end users, as well as ad-hoc SQL access by more advanced users.

SONCA's development is supported by the Polish National Centre for Research and Development (NCBiR) under the grant SP/I/1/77065/10 by the Strategic Scientific Research and Experimental Development program: "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information".

Formal Concept Analysis for Knowledge Discovery in Structured and Unstructured Data

Sergei O. Kuznetsov

School for Applied Mathematics and Information Science
National Research University Higher School of Economics
Moscow, Russia

Methods for analyzing data based on Formal Concept Analysis (FCA) fit well to the paradigm of knowledge discovery, which is interactive and iterative, involving numerous steps with many decision being made by the user.

Techniques related to extraction of knowledge from data, like attribute exploration, generation of implication bases and partial implication bases, were in the mainstream of FCA research from the very beginning.

Moreover, certain well-known methods of data analysis and machine learning, such as clustering and biclustering, induction of decision trees, generation of version spaces, JSM-method and generating bases of association rules are naturally expressed in terms of FCA, which helps their understanding and realization. For example, one has usually an enormous number of association rules valid in an arbitrary dataset. To represent all the rules, one can define a very natural lattice-based small subset of them (rule base), from which all other rules can be derived.

In practical applications objects and situations are often described not by binary tables, but in more complex ways, e.g., by graphs with labeled vertices and edges. This kind of descriptions underly applications in various domains, like analysis of properties of chemical molecules, human interaction in groups, and collection of texts. We show how these cases can be efficiently treated by FCA-based methods. We show a natural generalization of this approach, where objects can be described by arbitrarily ordered structures, e.g., by numerical intervals. We consider applications in pharmacology and bioinformatics.

Formal concepts may be considered as clusters of objects described by clusters of attributes, i.e., FCA is a natural tool for biclusterization, where each bicluster is described not by means of a distance metric, but sets of objects sharing sets of common attributes. This often gives much better interpretation of a group of similar objects. We show various applications of FCA-based biclustering, from gene expression analysis to analysis of collections of texts (text mining).

New Features of Infobright's RDBMS: Rough SQL and Seamless Handling of Unstructured Attributes in Machine-generated Data Sets

Dominik Ślęzak

Institute of Mathematics
University of Warsaw
ul. Banacha 2, 02-097 Warsaw, Poland
slezak@mimuw.edu.pl

Infobright Inc., Canada/Poland
47 Colborne St. #403, Toronto, ON M5E1P8 Canada
ul. Krzywickiego 34 lok. 219, 02-078 Warsaw, Poland
slezak@infobright.com

The relational model has been present in research and applications for decades, inspiring a number of RDBMS products based on entirely different architectures, but sharing the same way of understanding and representing the data [2]. Given 40 years of history, it is clear that the relational paradigms should not be blindly followed in all situations [1]. On the other hand, given its popularity, the relational framework is usually the easiest one to accept by database end-users and the most convenient one for interfacing with various tools [5].

An important trend in databases relates to the analytic engines aimed at advanced reporting and ad hoc querying. Such engines are usually used at the level of data marts, in the market segments where rapid data growth is expected, particularly for machine-generated data sets.[1] Originally, they have been technically complex and difficult to maintain. However, they have evolved toward solutions such as, e.g., Infobright's Community/Enterprise Editions[2] (ICE/IEE) capable of handling terabytes of data on a single off-the-shelf box.

Infobright's engine is a fully functional RDBMS product with internals based on columnar storage [3], adaptive compression [8], as well as compact *rough* information that replaces standard database indexes [7]. In this talk, we outline foundations of two new features of the latest ICE/IEE release[3] – fast computation of approximate results of SQL statements and the usage of domain knowledge about data to optimize string compression and query performance.

In particular, we investigate practical inspirations and opportunities for enriching standard SQL language with approximation aspects, assuming minimum impact on query syntax, as well as maximum easiness of interpreting and employing inexact query answers. We also discuss the idea of expressing data hierarchies independently from both logical and physical modeling layers, taking

[1] en.wikipedia.org/wiki/Machine-generated_data

[2] www.infobright.{org.com}

[3] www.dbms2.com/2011/06/14/infobright-4-0/

into account that machine-generated data providers and domain experts may need interfaces other than those designed for database users [4, 6].

References

1. Agrawal, R., et al.: The Claremont Report on Database Research. SIGMOD Rec. 37(3): 9–19 (2008).
2. Codd, E.F.: Derivability, Redundancy and Consistency of Relations Stored in Large Data Banks. SIGMOD Rec. 38(1): 17–36 (2009). (Originally: IBM Research Report RJ599, 1969.)
3. Hellerstein, J.M., Stonebraker, M., Hamilton, J.R.: Architecture of a Database System. Foundations and Trends in Databases 1(2): 141–259 (2007).
4. Kowalski, M., Ślęzak, D., Toppin, G., Wojna, A.: Injecting Domain Knowledge into RDBMS – Compression of Alphanumeric Data Attributes. In: Proc. of ISMIS, Springer, LNAI 6804 (2010) pp. 386–395.
5. Moss, L.T., Atre, S.: Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications. Addison-Wesley (2003).
6. Ślęzak, D., Toppin, G.: Injecting Domain Knowledge into a Granular Database Engine – A Position Paper. In: Proc. of CIKM, ACM (2010) pp. 1913–1916.
7. Ślęzak, D., Wróblewski, J., Eastwood, V., Synak, P.: Brighthouse: An Analytic Data Warehouse for Ad-hoc Queries. PVLDB 1(2): 1337–1345 (2008).
8. Wojnarski, M., et al.: Method and System for Data Compression in a Relational Database. US Patent Application 2008/0071818 A1 (2008).

Time Series Analysis

Roberto Baragona

University of Rome, Italy

Abstract. Time series analysis is a large research field that finds applications in several disciplines. Essential to time series analysis theory and practice are the three steps of model identification, estimation and diagnostic checking. Techniques are available that allow both the researchers and the practitioners to develop novel devices to analyze time series data, as well as to select and test the effectiveness of models for explanatory or forecasting purpose. Theory will be presented in three parts, in ascending order of difficulty, from basic linear models to testing and modeling non-linear and non-stationary behavior, up to vector models and time series interaction problems as the obvious consequence. Whatever would be the model complexity or the difficulty of taking properly into account unusual or unexpected time series sequences, identification, estimation and diagnostics will have to be performed, adopting the techniques that will best fit the data.

Problems of Ontology Development for a Broad Domain

Natalia Loukachevitch

Moscow State University, Russia

Abstract. Working in various applications one often needs to create an ontology of the domain for description of domain knowledge. In many cases the domain is not well-structured, its boundaries are not clearly defined, a lot of knowledge should be extracted from textual sources. In the tutorial I will discuss some typical problems and solutions when developing ontology in such complicated circumstances.

Recent Advances on Generalization Bounds

Konstantin Vorontsov

Dorodnycin Computing Center RAS, Moscow, Russia
voron@forecsys.ru

Abstract. Bounding generalization ability of learning algorithms remains a challenging open problem in Statistical Learning Theory for more than 40 years, starting with pioneer works of Vapnik and Chervonenkis. Many improvements has been made recently in this field and many mathematical techniques are developed to apply *generalization bounds* for learning algorithms design. Most important issues are briefly reviewed in the first part of the tutorial.

In the second part a *permutational probabilistic framework* is introduced and a combinatorial technique is presented for obtaining tight data dependent generalization bounds. It is based on a detailed representation of the internal structure of the classifier set via a multipartite graph called the *splitting and connectivity (SC) graph*. Combinatorial generalization bounds based on statistical properties of the SC-graph are shown to be exact in some nontrivial particular cases. Some practical applications of SC-bounds are considered.

Reliability Analysis for Aerospace Applications: Reducing Over-Conservative Expert Estimates in the Presence of Limited Data

Vladik Kreinovich

University of Texas at El Paso, El Paso, TX 79968, USA
vladik@utep.edu

Abstract. Unique highly reliable components are typical for aerospace industry. For such components, due to their high reliability and uniqueness, we do not have enough empirical data to make statistically reliable estimates about their failure rate. To overcome this limitation, the empirical data is usually supplemented with expert estimates for the failure rate. The problem is that experts tend to be – especially in aerospace industry – over-cautious, over-conservative; their estimates for the failure rate are usually much higher than the actual observed failure rate. In this presentation, we describe a new fuzzy-related statistically justified approach for reducing this over-estimation. Our preliminary results are described in [2].

Keywords: reliability analysis, aerospace industry, expert estimates, over-conservative expert estimates

Reliability: how it is usually described and evaluated. Failures are ubiquitous. As a result, reliability analysis is an important part of engineering design.

In reliability analysis of a complex system, it is important to know the reliability of its components; see, e.g., [1]. Reliability of a component is usually described by an *exponential model*, in which the probability $P(t)$ for a system to be intact by the time t is equal to $\exp(-\lambda \cdot t)$ for some constant λ. For this model, the average number of failures per unit time is equal to λ; as a result, this value is called a *failure rate*. Another important characteristic – mean time between failure (MTBF) θ – is, in this model, equal to $1/\lambda$.

Usually, the failure rate λ (or, equivalently, the MTBF θ) are determined by analyzing the records of actual failures. When we observe a sufficient number of failures, we can then take an arithmetic average of the observed times between failures – and this average is a statistically justified estimate for θ.

Reliability estimates in aerospace industry: a challenge. In aerospace industry, especially in designing spaceships for manned flights, reliability is extremely important. Because of this importance, aerospace systems use unique, highly reliable components.

This reliability, however, leads to a challenge: since the components are unique and highly reliable, we do not have enough failure records to make statistically reliable estimates about their failure rate: in most cases, we have up to

5 failures. This scarcity of data is especially critical on the stage when we are still designing a spaceship.

Need to use expert estimates. To overcome this limitation, the empirical data is usually supplemented with expert estimates for the failure rate.

Expert estimations are over-conservative: a problem. A problem with expert estimates is that experts tend to be – especially in aerospace industry – over-cautious, over-conservative. The experts' estimates for the failure rate are usually much higher than the actual observed failure rate.

New approach: main idea. Our main idea is to use the fact that here, we have *two* sources of knowledge: (1) the empirical failure data t_{ij}, and (2) the expert estimates $e_1, \ldots, e_n$ for the failure rates of different components.

For each individual component i, we do not have enough data to provide us with a meaningful statistically significant estimate for its failure rate λ_i. However, when we combine all these data together, we will get enough data points to gauge the accuracy of an expert – as an instrument for estimating the failure rates. As a result of this statistical analysis, an expert becomes a statistically justified estimation tool; so, we can add the expert estimates to the observed times t_{ij}; this additional data allows us to get better estimates for λ_i.

Discussion. If the empirical data t_{ij} are not sufficient to make statistically reliable estimates about the failure rate, why these data are considered sufficient for gauging/reducing the over-conservativeness of experts' estimates?

Failures of different components are considered statistically independent. Thus, *in the absence of expert estimates*, to find the failure rate λ_i, we can only use the values t_{ij} corresponding to this component. Since we have few such values, this data is not sufficient.

On the other hand, the over-conservativeness of *an expert* is reflected in the expert's estimates of the failure rates of all the components. Thus, to estimate this over-conservativeness, we can use the data from all the components. We may have about 5 measurement values for each component, but since we have dozens of components, we thus have hundreds of values t_{ij} that can be used to estimate this over-conservativeness – enough to make statistically reliable estimates.

Acknowledgments. This work was supported by a contract from the Boeing Company, by the National Science Foundation grants HRD-0734825 and DUE-0926721 and by Grant 1 T36 GM078000-01 from the National Institutes of Health.

References

1. Barlow, E. E.: Engineering Reliability, SIAM Publ., Philadelphia (1998)
2. Ferregut, C., Campos, F. J., Kreinovich, V.: Reducing over-conservative expert failure rate estimates in the presence of limited data: a new probabilistic/fuzzy approach, Proc. 30th Annual Conf. of the North American Fuzzy Information Processing Society NAFIPS'2011, El Paso, Texas, March 18–20, 2011 (2011)

Data Analysis Projects for Business: Challenges, Approaches and Cases

Daniel Kanevskiy

Forecsys, Russia

Abstract. Real-world data analysis applications involve a host of non-research, though research-related, issues. This talk is an attempt to summarize research-intensive projects experience, with the emphasis on the connection between science and business from the scientist' point of view. This context allows one to formulate several necessary, but not sufficient conditions of a successful data analysis project, such as business-oriented results, penetration into data-flow processes and effective project management in a highly risky environment. General observations on the topic are supplemented with real cases having classification or forecasting problems at the core.

Application of the Integrity, Purposefulness and Adaptivity Principle in OCR and Applied Linguistics

Aram Pakhchanian

ABBYY Software House, Russia

Abstract. In my talk I will be describing the general principle used by ABBYY across many of its core technologies that deal with recognition and semantic processing. The principle was introduced by Alexander Shamis years ago and is called IPA (Integrity, Purposefulness and Adaptivity). This core idea of IPA was borrowed from the observation of how live species detect and identify objects around them. Object identification process goes through the following steps: the attempt identification the object in general, as an integral set of its features. At that step we use only features that apply to the object as a whole. The result is not yet a final decision but a set of hypotheses. As each of them is being addressed, the system purposefully uses only detectors that measure features that could help to support or decline the specific hypothesis. This process leads to each of detectors assigning an integral measure of probability. The final decision is made by comparing weighted average probabilities coming from each of detectors. As the system goes through this process, it learns how to better fit its recognition process into specific overall conditions and that allows to adapt the system to improve its behaviour in these conditions.

FineReader OCR technology is based on IPA and many of its components, which include character recognition, page analysis, word recognition, etc. The same applies even to greater extent to ICR technology developed by ABBYY, as it is based on IPA as its core principle. The FlexiCapture technology that is used to detect and capture field from form documents, also uses IPA in its core.

As we further pursued technology goals outside of optical recognition area, we have found the IPA to be a universal principle that can be applied in many other areas. One major area where IPA can be efficiently applied is linguistics. The application of IPA to the process of syntax analysis will be described.

Image Processing & Analysis Using Soft Computing Tools

Malay K. Kundu

Machine Intelligence Unit
Indian Statistical Institute
Kolkata, India

Abstract. Soft computing is a consortium of methodologies which work synergistically and provides, in one form or another, flexible information processing capabilities for handling real life ambiguous situations. Its aim is to tolerate the imprecision, uncertainty, approximate reasoning and partial truth in order to achieve tractability, robustness, low solution cost, and close resemblance with human like decision making. In other words, it provides the foundation of the conception and design of high machine IQ (MIQ) systems, and therefore forms the basis for future generation computing systems. At this juncture, fuzzy logic (FL), artificial neural networks (ANN), genetic algorithms (GAs) and Rough sets (RS) are the four principal components where FL provides algorithms for dealing with imprecision and uncertainty, and computing with words, ANN the machinery for learning and adaptation; GA is used for optimization and searching and RS for rule induction from incomplete data sets.

In an image analysis system, uncertainties can arise at any phase resulting from incomplete or imprecise input information, ambiguity or vagueness in input images, ill-defined and/or overlapping boundaries among the classes or regions, and indefiniteness in defining/extracting features and relations among them. Any decision taken at a particular stage will have an impact on the subsequent stages. It is therefore required for an image analysis system to have sufficient provision for representing the uncertainties involved at every stage, so that the ultimate output (results) of the system can be associated with the least uncertainty.

The utility of fuzzy and rough sets theory in handling uncertainty, arising from deficiencies of information available from a situation (as mentioned above) in image processing and recognition problems, have adequately been addressed in the literature. These theories provide approximate, yet effective and more flexible means of describing the behavior of systems which are too complex or too ill-defined to admit precise mathematical analysis by classical methods and tools. Since both the theory of fuzzy sets rough sets are generalization of classical set of theory, both have greater flexibility to capture faithfully the various aspects of incompleteness or imperfection (i.e., deficiencies) in information of a situation. They also have capability to mimic human reasoning process for decision making.

On the other hand, to achieve robustness of performance with respect to random noise and failure of components, and to obtain output in real

time, a system can be made artificially intelligent if it is able to emulate some aspects of human information processing system. Artificial neural network (ANN) based approaches are attempts to achieve these goals. One may also note that, most of the image analysis operations are co-operative in nature and the tasks of recognition mostly need formulation of complex decision regions. ANN models have the capability of achieving these properties. All these characteristics, therefore, suggest that image processing and recognition problems can be considered as prospective candidates for neural network implementation.
It is well known that the methods developed for image processing and recognition are usually problem dependent. Moreover, many tasks involved in the process of analyzing/identifying a pattern need appropriate parameter selection and efficient search in complex spaces in order to obtain optimal solutions. This makes the process not only computationally intensive, but also leads to a possibility of losing the exact solution. Genetic Algorithms (GAs), another biologically inspired technology, are randomized search and optimization techniques guided by the principles of natural evolution and natural genetics. They are efficient, adaptive and robust search processes, producing near optimal solutions and have a large amount of implicit parallelism. Therefore, the application of genetic algorithms for solving certain problems of image processing/pattern recognition, which need optimization of computational requirements, robust, fast and approximate solution, appears to be appropriate and natural. The presentation deals with the relevance and feasibility of soft computing tools like FL and GAs in few areas of image processing and analysis.

References

1. S. K. Pal, A. Ghosh and M. K. Kundu, Soft Computing for Image Processing, Physica Verlag, Heidelberg, 2000.
2. M. Banerjee, M. K. Kundu and P. Maji "Content-based image retrieval using visually significant point features", Fuzzy Sets and Systems, vol. 160, no. 23, pp. 3323-3341, 2009.
3. M. Acharyya, M. K. Kundu and R. K. De " Segmentation of remotely sensed images using wavelet features and their evaluation in soft computing framework" , IEEE Trans. on Geoscience and Remote Sensing, Vol.41, No. 12, pp. 2900-2905, 2003.
4. M. Acharyya, M. K. Kundu and R. K. De " Extraction of features using M- band wavelet packet frames and their neuro-fuzzy evaluation for multi-texture segmentation", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.25, No. 12, pp. 1639-1644, 2003
5. S.P. Maity and M. K. Kundu "Genetic Algorithms for Optimality of Data Hiding in Digital Images", Soft Computing, 13, pp. 361-373, 2009.
6. M. Banerjee and M. K. Kundu , "Handling of impreciseness in gray level corner detection using fuzzy set theoretic approach", Applied Soft Computing, 8, pp 1680-1691, 2008.
7. S. K. Pal, D. Bhandari and M. K. Kundu "Genetic algorithms for image enhancement ", Pattern Recognition Letters, vol. 15; pp 261–271, 1994.

New Histogram-like Techniques for Cardinality Estimation in Database Query Optimization

B. John Oommen

School of Computer Science, Carleton University, Ottawa, Canada**

Abstract

Problem Studied: The problem that we have studied involves the design and implementation of new histogram methods that have direct and immediate applications in Query Optimization (QO) in databases. The problem is inherently an extremely hard problem, and its complexity is termed to be "NP-Hard". Quite simply stated, given a database with various relations, each with its own attributes, the QO problem involves determining the optimal way of searching the database to yield the result of a given query. The problem is "NP-Hard" essentially because there are a "large" (typically exponential) number of possible Query Evaluation Plans (QEPs) for a given query.

Solution Methodology: Rather than search all possible QEPs, traditional and current database systems try to estimate the efficiency of various QEPs, and thus discard the ones which are not promising. To achieve this, these systems rely on the fact that the complexity of a QEP depends on the size of the intermediate relations that the QEP generates. Thus, rather than test a particular QEP, current systems avoid calculating the size of the intermediate relations, but estimate it using a structure called a "Histogram". These Histograms do not deal with the actual physical data, but rather process the "meta-data" stored in the data dictionary. The histograms currently in use are the Equi-depth and the Equi-width Histograms.

Current Technology: All the current database systems utilize the histogram in one of its forms. The IBM-DB2, ORACLE, Sybase, and NCR's Teradata, INGRES etc. all use the Equi-depth histogram.

Our Strategy: We have proposed two new histogram-like techniques refereed to as the Attribute Cardinality Maps (ACMs). These are the Rectangular Attribute Cardinality Map (R-ACM) and the Trapezoidal Attribute Cardinality Map (T-ACM). The technology has been patent protected (US Patent No. 6,865,567; Issued on March 8, 2005).

** *Chancellor's Professor*; *Fellow : IEEE* and *Fellow : IAPR*. The Author also holds an *Adjunct Professorship* with the Dept. of ICT, University of Agder, Norway.

Research Challenges of Dynamic Socio-Semantic Networks

Rostislav Yavorskiy

Witology, Models and Algorithms
Kapranova str. 3, Moscow, RUSSIA
Rostislav.Yavorskiy@witology.com
http://www.witology.com

Abstract. A general model of a socio-semantic network is presented in terms of state-transition systems. We provide some examples and indicate research directions, which seem to us the most important from the application point of view.

Keywords Social Network, Semantic Network, Socio-Semantic Network

1 Model of a socio-semantic network

1.1 Social network

A social network is usually modeled as a weighted multi-graph

$$G = \{V, E_1, \ldots, E_k; \pi, \delta_1, \ldots, \delta_k\},$$

where

- V represents members of the network,
- $E_1, \ldots, E_k \subset V \times V$ denote different relations between the members, e.g. being a friend, follower, relative, co-worker etc.
- $\pi : V \to \Pi$ is a *user profile* function, which stores personal information about the network members.
- $\delta_i : E_i \to \Delta_i$ $(i \in \{1, \ldots, k\})$ keeps parameters and details of the corresponding relation.

1.2 Content

The model of the content has a very similar definition. It is a multi-graph

$$C = \{T, R_1, \ldots, R_m; \theta, \gamma_1, \ldots, \gamma_m\},$$

where

- T stands for the set of all elements of the generated content, e.g. posts, comments, evaluations, tags etc.
- $R_1, \ldots, R_m \subset T \times T$ denote different relations on the content, e.g. being a reply on, have the same subject, etc.
- $\theta : T \to \Theta$ stores parameters of the content;
- $\gamma_i : R_i \to \Gamma_i$ $(i \in \{1, \ldots, k\})$, similarly, keeps parameters and details of the corresponding relation.

1.3 Authorship and other relations between the users and the content

The basic connections between the social graph and the content are defined by the authorship relation A,

$$A \subset V \times T.$$

One can also consider other kinds of connections of this kind, but usually all of them could be modeled via introducing a new type of content. For example, the relation *John is interested in post "Announcement"* could be modeled by introducing a new content node *interest evidence*, which points to "Announcement" (use the corresponding relation R_i here) and is authored by John.

1.4 The context

Before we turn to description of the socio-semantic network dynamics there is one more important parameter not to be missed. It is external context, Ω, which may include different parameters like project or campaign phase, flag for a bank holiday, or a maintenance status of the network.

2 The socio-semantic network dynamics

Now, when all the components of the network have been defined, the list of possible system updates, which determine the network evolution, is rather evident:

- addition of new members to V;
- changes in user profiles π;
- updates of social relations $E_1, \dots, E_k$ and their parameters $\delta_1, \dots, \delta_k$;
- creation and update of content nodes in T, (also affects the authorship relation A);
- changes in properties of and relations between the content nodes $\theta, \gamma_1, \dots, \gamma_m$;
- changes in context Ω.

3 Examples

3.1 School or a training center

A training center usually has a standardized set of reading materials, textbooks, tasks, assignments and exam tests. At the same time, the students' network evolves permanently. In terms of the definition above one can say that the content part of this socio-semantic network is rather stable while the social network is very dynamic.

3.2 Research or analytic team

This example resides on the opposite side of the spectra. The team (the social network part) is rather stable while the content is actively processed and generated.

3.3 Fixed term project

A targeted crowdsourcing project or a collective intelligence venture provide an example of a dynamic socio-semantic network, which is created from the scratch and is aimed at solving a particular task or a problem. New content is generated and new members join the network at all stages of its lifecycle.

4 Research challenges

Assume that we have all necessary data about the network dynamics available. In all the examples mentioned above one can identify two principal tasks for analysis:

- Given all the data about the users activities and the content discover the right people (knowing, capable, skillfull etc.)
- Given all the data about the social network dynamics and the content evolution discover the right texts (interesting, influential, prominent etc.)

Many promising approaches and useful algorithms have already been developed during the last decades [1–4], several new ideas are implemented in the Witology platform [5]. Some have proved to be quite efficient, although most of them are based on fairly simple mathematical tools. Still, the field is rather in its rudimentary phase. We believe that the next breakthrough lies in interdisciplinary research covering sociology, psychology, linguistics and other related fields.

References

1. Sergey Brin, Lawrence Page. *The anatomy of a large-scale hypertextual Web search engine.* Computer Networks and ISDN Systems 30: 107117, (1998).
2. Damon Horowitz, Sepandar D. Kamvar. *The Anatomy of a LargeScale Social Search Engine.* Proceedings of WWW'2010, April 26–30, 2010, Raleigh, North Carolina.
3. Camille Roth, Jean-Philippe Cointet. *Social and Semantic Coevolution in Knowledge Networks,* Social Networks, 32(1):16-29 (2009)
4. Jean-Philippe Cointet, Camille Roth. *Socio-semantic Dynamics in a Blog Network,* IEEE SocialCom International Conference on Social Computing, Vancouver, Canada, August 2009.
5. Witology, "Search of good ideas through search of people, and search of right people through search of ideas", http://www.witology.com

Trust Networks for Recommender Systems

Chris Cornelis

Department of Applied Mathematics and Computer Science
Ghent University, Gent, Belgium
Chris.Cornelis@UGent.be

E-commerce companies rely on personalized recommendations to promote their products and to strengthen customer loyalty. Specifically, recommender systems cater to customers' information needs proactively by making them aware of "items" (books, movies, web pages, files, jobs, etc.) that may interest them. The success of recommendation techniques depends on a good knowledge of the user's interests. Such information can be obtained e.g. from browsing or buying behavior, but more and more information is also drawn from the social networks in which the users participate. For instance, if a user indicates that he values the opinions of a peer highly, this may be taken into account when forming recommendations. Trust networks quantify such knowledge: they allow users to express their trust and distrust in their peers through numerical trust scores. Since most trust networks are sparse, determining trust scores for user pairs that are indirectly connected is an important problem, which can be solved through defining adequate propagation and aggregation operators.

In this talk, I will review the bilattice-based model from [1] to represent trust and distrust as separate, gradual concepts, along with the associated propagation and aggregation operations defined in [1] and [2]. I will also recall different trust-enhanced recommendation strategies that were introduced in the literature [3,4,5], and discuss how they can be combined with distrust information [6].

References

1. P. Victor, C. Cornelis, M. De Cock, P. Pinheiro Da Silva: Gradual Trust and Distrust in Recommender Systems. Fuzzy Sets and Systems 160(10): 1367-1382 (2009).
2. P. Victor, C. Cornelis, M. De Cock, E. Herrera-Viedma: Practical Aggregation Operators for Gradual Trust and Distrust. To appear in: Fuzzy Sets and Systems.
3. J. Golbeck: Computing and applying trust in web-based social networks. PhD thesis (2005).
4. P. Massa, A. Avesani: Trust-aware recommender systems. In: Proc. ACM Recommender Systems Conference (2007) pp. 17-24.
5. J. O'Donovan, B. Smyth: Trust in recommender systems. In: Proc. IUI (2005) pp. 167-174.
6. P. Victor, C. Cornelis, M. De Cock, A. Teredesai: Trust- and Distrust-Based Recommendations for Controversial Reviews. IEEE Intelligent Systems 26(1): 48-55 (2011).

Ontology for Multimedia Information Processing

Santanu Chaudhury

Indian Institute of Technology
Hauz khas, Delhi 110 016, India

Machine interpretation of documents and services in Semantic Web environment is primarily enabled by (a) the capability to mark documents, document segments and services with semantic tags and (b) the ability to establish contextual relations between the tags with a domain model, which is formally represented as ontology. Human beings use natural languages to communicate an abstract view of the world. Natural language constructs are symbolic representations of human experience and are close to the conceptual model that Semantic Web technologies deal with. Thus, natural language constructs have been naturally used to represent the ontology elements. This makes it convenient to apply semantic web technologies in the domain of textual information.

In contrast, media documents are perceptual recording of human experience. An attempt to use the conceptual model to interpret the perceptual records gets severely impaired by the semantic gap that exists between the perceptual media features and the conceptual world. Notably, the concepts have their roots in perceptual experience of human beings and the apparent disconnect between the conceptual and the perceptual world is rather artificial. The key to semantic processing of media data lies in harmonizing the seemingly isolated conceptual and the perceptual worlds. Ontological description of a domain needs to be extended to enable perceptual modeling, over and above conceptual modeling that is presently supported. The perceptual model of a domain primarily comprises observable media properties of the concepts. Such perceptual models will be useful for semantic interpretation of media documents, just as the conceptual models help in the semantic interpretation of textual documents.

Concepts are formed in human minds through a complex refinement process of personal experiences [1]. An observation of the real world object amounts to reception of perceptual data through our sense organs. The raw data goes through a process of refinement depending on the personal viewpoint of the observer, and gets encoded in our minds to give rise to a mental model. An abstraction of many such mental models, arising out of many observations of the real world, gives rise to concepts, which are labeled with some linguistic constructs to facilitate communication using spoken or written text. Further, the similarities between the perceptual properties of concepts give rise to concept taxonomy, on which domain ontology is built. For example, observation of a number of monuments, and analysis of their structural similarities, gives rise to concepts like the "monument" and its sub-classes, such as "fort", "palace" and "tomb". The fact that concepts are abstractions of perceptual observations has an interesting consequence. When we look for an instance of a concept in the

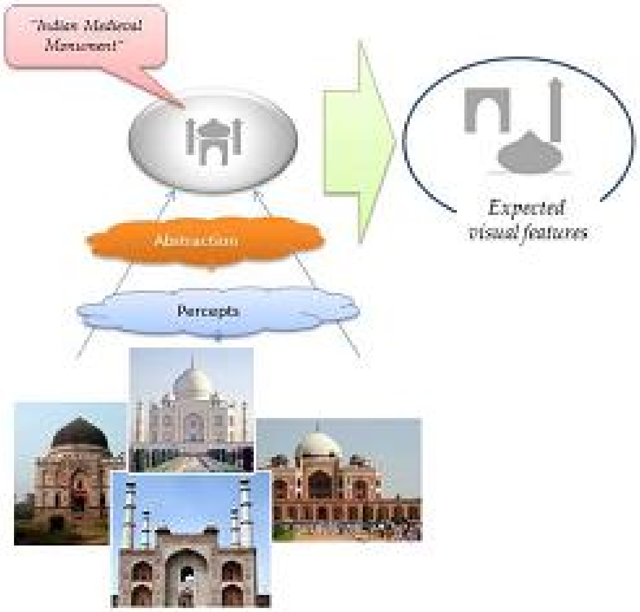

Fig. 1. Perceptual Modeling of a Concept "Indian Medieval Monument".

real world or in multimedia documents, we expect some perceptual patterns to appear.

A concept is recognized on the basis of evidential strength of some of these patterns which are actually observed. For instance, we recognize a monument, when we observe one or more of its characteristic visual patterns such as a dome, the minarets and the facade in still images or videos. Thus, concepts in the real world are characterized by observation models, which comprise expected manifestations of the concept in the perceptual space and are the key to their recognition through perceptual evidences. Figure 1 illustrates the abstraction process and some expected media patterns for a concept labeled "Indian Medieval Monuments". It may also be noted that the observation model of a concept is influenced by the related concepts also. The observation model of a monument, e.g. the Tajmahal, will comprise the color and texture of its building material, namely marble. On the other hand, Tajmahal being an instance of a tomb, an example media instance of the former, e.g. a photograph, is a valid media instance of the latter. Figure 2 illustrates such media property propagation across related concepts in a domain.

Creation of an observation model for a concept requires domain knowledge that embeds perceptual property descriptions of concepts and media illustrations. For example, a part of the observation model for the monument "Tajmahal" can be derived from the knowledge that the Tajmahal is an instance

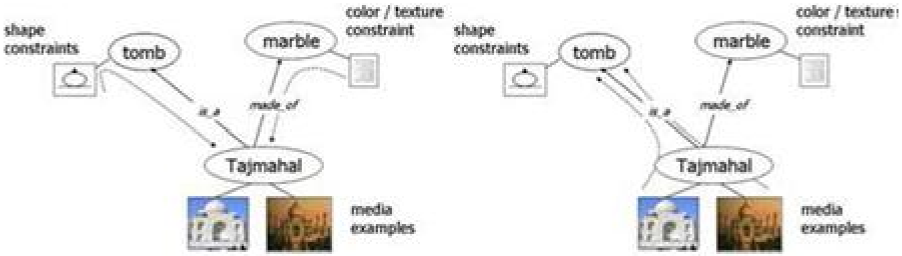

Fig. 2. Media property propagation across related concepts.

of a "medieval monument" and that medieval monuments comprise "domes", "minarets" and "arches", which can be recognized by their characteristic shapes. Multimedia ontologies need to encode this information.

To express such ontologies representing perceptual models, we need a multimedia ontology language that provides a means for expressing media properties explicitly for the concepts in a domain. It should have enough flexibility to capture the encoding of different types of media feature descriptions and different media formats at varying levels of abstractions. It should be possible to specify the uncertainties associated with the media property of concepts. The language should provide the capability of expressing spatial and temporal relations between the media properties that often characterize multimedia objects and events. To enable formulation of observation model for the domain concepts, it should support reasoning with the relations between the concepts for media property propagation. The ontology language needs to be complemented with a scheme for reasoning with uncertainties for concept recognition.

Existing works on knowledge representation for multimedia data processing lack the rigor of modern ontology languages and associated reasoning. With this background, we present a new ontology representation language Multimedia Web Ontology Language (MOWL) for multimedia data processing.

References

1. H. Kangassalo: Conceptual level user interfaces to data bases and information systems. In: Advances in information modelling and knowledge bases, H. Jaakkola, H. Kangassalo, and S. Ohsuga (eds.), IOS Press (1991) pp. 66–90.

Gaining Insight into Clinical Pathway with Process Discovery Techniques

Jonas Poelmans

K.U. Leuven, Faculty of Business and Economics
Naamsestraat 69, 3000 Leuven, Belgium

Abstract. The literature on management of health care organisation has identified clinical pathways as a valuable methodology for structuring and improving clinical care processes. This methodology has received considerable interest over the last years and is used to impose a care plan designed by domain experts that should be followed by all stakeholders to provide optimal care to the patient.

However, if this methodology is used without bottom up data analysis techniques that give insight into what is really happening at the operational working floor, it is possible there are serious shortcomings that remain undetected. We analysed the breast cancer care process in our hospital and found for example that 25 % of our breast conserving therapy patients did not receive a revalidation although it was prescribes as a key intervention in the care pathway primary operable breast cancer. Activities performed to patients are logged to a database by the patient care system. We used process and data discovery techniques that visualized the underlying concepts of these data and found that they are powerful instruments for the multidisciplinary team to gain insight in the organisation of their care processes.

Determining Pattern of Variation in Gene Expression Profiles Using Correlation Clustering Algorithms

Rajat K. De

Machine Intelligence Unit, Indian Statistical Institute
203 Barrackpore Trunk Road, Kolkata 700108, India

Cluster analysis (of gene-expression data) is a useful tool for identifying biologically relevant groups of genes that show similar expression patterns under multiple experimental conditions. A number of methods are already in literature for clustering gene-expression data. However most of these algorithms group the genes based on the similarity of their expression values.

Here we present a correlation clustering algorithm, called Divisive Correlation Clustering Algorithm (DCCA) developed recently, which is suitable for finding a group of genes having similar pattern of variation in their expression values. In order to detect clusters with high correlation and biological significance, we use the concept of correlation clustering introduced by Bansal et al. The algorithm DCCA provides clusters of genes without taking number of clusters to be created as an input. DCCA uses the correlation matrix in such a way that all the genes in a cluster have the highest average correlation among them. The clustering results of the DCCA were found to be more significantly relevant to the biological annotations than those of the other methods.

But this algorithm may also fail for certain cases. In order to overcome these situations, we have developed another correlation clustering algorithm, called average correlation clustering algorithm (ACCA), which is able to provide better clustering solution than that produced by some others including DCCA. ACCA is able to find groups of genes having more common transcription factors and similar pattern of variation in their expression values. Moreover, ACCA is more efficient than DCCA with respect to the time of execution. Analysis of the results shows the superiority of ACCA over some others including DCCA, in determining a group of genes having more common transcription factors and with similar pattern of variation in their expression profiles.