Roberto Bruni, Stefano Chessa, Antonio Cisternino,
Gianna M. Del Corso, Nadia Pisanti (Eds.)

WiGoWiN

1st Internal Conference on

What is Going on and What is Next?

Pisa, Italy, May 26–27, 2010

Proceedings

Dipartimento di Informatica, Università di Pisa

Preface

This volume contains the proceedings of WiGoWiN, the "1st Internal Conference on What is Going on and What is Next?", held in Pisa, Italy on May 26–27, 2010.

The event aimed at establishing a forum for spreading knowledge on current research activity and main achievements carried out at the Computer Science Department, promoting the exchange of ideas about future trends/perspectives for research in computer science, and strengthening collaborations between different research groups.

Presentations on the above issues have been given by faculty members of the Computer Science Department to a wide, heterogeneous audience that included MSc and PhD students, postdocs, and Computer Science related researchers from Academy, Industry and other research Institutions in Pisa.

The event was initially planned to last one day, but since the call for abstracts was very successful it needed to be expanded on one day and a half. The Organizing Committee received 29 abstracts. Due to time constraints, it was possible to allocate 25 presentation slots only, and therefore some members of the Organizing Committee volunteered not to present some activities they co-authored to leave room for those of other colleagues. This volume collects all 29 abstracts, listed according to the alphabetical order of authors. Although some abstracts reported work carried out in collaborations with researchers from other Universities and Institutions, for conciseness in the Table of Contents we list only authors from the Computer Science Department of the University of Pisa. Full credits can be found in each abstract.

The topics of the event included talks on: motivations, descriptions and applications of recent results; methodologies and tools for addressing novel challenges; key achievements and main open problems of specific research fields; experimental teaching experiences; interdisciplinary approaches; visionary perspectives on where computer science should head to in a particular area, in the near future.

We want to thank all faculty members who (co-)authored the abstracts collected in the proceedings. Very special thanks go to all the young researchers, PhD students, PostDocs and temporary research-contract holders whose enthusiasm and efforts form the backbone of most research activities, despite the fact that the national and local financial policies do not encourage at all their permanence in the italian academy. We wish that their careers will progress and receive the due honors in the same country where the investments on their scientific education were made.

Our best thanks go to Giorgio Levi and the direction of the Dipartimento for being the main sponsor of this event. We also thank our Preside Umberto Mura for making Aula Magna available for presentations. Finally, we want to thank Vincenzo Ciancia, Andrea Corradini, Giovanni Giuseppe Mandorino and Stefano Palla for their technical support during the preparation of the conference web page, and our colleagues Chiara Bodei, Gianluigi Ferrari, Ornella Menchi, and Susanna Pelagatti who, jointly with ourselves, made it possible to cover event fees.

Pisa, May 2010

Antonio *Gianna* *Nadia* *Roberto* *Stefano*

Table of Contents

From Data to Decision

Antonio Albano

If I Had Enough Time . . . Research Suggestions Given Away

Context. Organizations collect data in large quantities, and they are aware that data by themselves are not very useful if they are not condensed in appropriate forms for human interpretation. Moreover, organizations are becoming increasingly aware that nowadays the solution is to invest in strategies based on Business Intelligence tools in order to obtain from the available data useful and timely information to assist with decision-making at all levels: strategic, managerial and operational. Business Intelligence tools can be divided into the following types: Multidimensional Analysis, to carry out an interactive analysis of data; Data Mining, to carry out exploratory analysis of data.

However, the effective use of Business Intelligence tools require a solution of several problems. The focus of the presentation will be only on two of them: (1) how to develop a technology to analyze quickly huge amount of data organized in a Data Warehouse, a database specifically designed to make data easier to analyze in order to answer business questions using Business Intelligence tools, (2) how to facilitate the use of Data Mining, which currently is not as widely used as it could be in the business world.

A Column-Oriented DBMS for Data Warehouses. (*Joint work with Advanced Systems*) A truly change in price/performance characteristics of Business Intelligence applications can be achieved only by a fundamental rethinking of Data Warehouse systems architecture. The main commercial Data Warehouse systems available today are based on row-oriented relational technology optimized for *On Line Transaction Processing* applications, i.e. they store database tables by rows, while substantial improvements in query performance for *On Line Analytical Processing* applications can be achieved by systems based on column-oriented technology, i.e. by storing database tables by columns rather than by rows. This is because the dominant queries require grouping and aggregations of large amounts of data using only a few columns. This new assumption means that several aspects of storage structures and query optimization of star queries with group-by need to be reconsidered to produce efficient query execution plans.

A brief overview is presented of the SADAS system and of some open problems. SADAS is the result of an industrial research project sponsored by the Italian Ministry of Education, University and Research (MIUR) to support the cooperation of universities and industries in prototyping innovative systems.

How to exploit data mining without becoming aware of it. (*Joint work with Nicola Ciaramella*) Data Mining has proved to be a valuable tool in discovering non-obvious information from a large collection of data, however in the business world is not as widely used as it could be. Common reasons include the

following: (a) Data Mining process requires an *unbounded rationality*; (b) potential end users may not be available to inform developers on what problems they are interested in having solved or what their requirements might be; (c) high costs in the use of Data Mining experts; (d) the actual result of Data Mining may be irrelevant or simply cannot be used.

A brief overview is presented of a methodology, a system and the open problems to exploit Data Mining in traditional businesses using the following approach based on a context of *bounded rationality* and *long tail economy*: many models are automatically generated and stored in a database; when the end users specify some features of the model they are looking for, a search engine then retrieves any relevant models.

As a simple analogy, consider data as grapes and models as wine; Data Mining is then like the wine making process. Although users can make wine from their own grapes, this takes both time and know-how, and naturally most business users prefer bottled wine. Data Mining still takes place behind the scenes, but the business user is unaware of it.

A Universal Language Engine for Machine Reading

Giuseppe Attardi and Maria Simi

joint work with Niladri Chatterjee, Stefano Dei Rossi, Zahurul Islam, Haoyuan Li
Dipartimento di Informatica
Università di Pisa

Context. Traditionally, research in Natural Language Processing has divided the activity of language analysis into a series of simpler tasks, in order to tackle such a complex problem by splitting it into more affordable separate sub-tasks. In particular, tasks such as Part of Speech Tagging or Chunking purportedly can be carried out in isolation using only shallow linguistic information, for instance superficial aspects related to the morphological structure of words (prefix, suffix, capitalization, etc.). Processing can be done with Markov models, assuming a behavior that only depends on the current state and is independent from the past and the future. Exploiting techniques of Machine Learning, tools have been developed that perform such tasks with good accuracy (close to 98%) and high efficiency.

Sentence analysis requires performing a sequence of tasks, using linguistic pipelines consisting of tools such as: Sentence Splitter, Tokenizer, Lemmatizer, POS tagger, Chunker, Parser. Additional tools for semantic analysis include Named Entity Recognizer, Super Sense Tagger, Semantic Role Labeling, Coreference Resolution [20] and Machine Translation [13].

It is questionable though whether the human mind performs linguistic analysis in separate stages: in fact in many situations it would appear quite beneficial to exploit information produced by one component while performing another task.

It has been recently shown [9] that a dependency parser can subsume other tasks such as chunking with no loss of efficiency, by exploiting new algorithms and deterministic parsing techniques [1, 3, 4, 5, 6]. Semantic Role labeling can be done while parsing with minor extensions to the parser algorithm. We also showed that a Named Entity Recognizer can avoid the use of POS tags [7], extracting directly simple features from words.

Hence it possible to reconsider the architecture of Natural Language Processing systems, reducing the number of pipeline stages, as some recent research on multitask systems [18] suggests.

Improving the effectiveness of text analytics technologies may have significant impact in the way information is processed to extract knowledge, in the way computers can support higher level of interaction with humans and in the way they can assist in performing complex tasks.

Research Goals. Machine Reading [19] involves the ability of capturing knowledge from naturally occurring texts and transforming it into suitable representations for use by reasoning systems, analysis systems or any other processing tools.

Our main research goal is to design and develop the architecture for a *Universal Integrated Natural Language Analytics Engine for Machine Reading*, i.e. capable of performing complex semantic analysis of texts, exploiting multiple layers of features extracted simultaneously from the input documents. Such engine can be trained jointly on all tasks and produce a *common model* consisting of shared weights that allows obtaining a *reading of a document* from multiple perspectives. In particular the engine could be trained to perform parsing, semantic role labeling, information extraction and syntax-based translation.

Linguistic knowledge will be incorporated into special document indexes that will enable semantic search. The *semantic index* will represent linguistic information in a redundant way, spread across each occurrence of a word [2], in order to be quickly accessible where needed for search, while exploiting compression techniques for space saving. Such redundancy is the opposite of database normalization and is akin to *denormalization*.

The language model and semantic index will represent billions of linguistic knowledge items stored in a form resembling an *artificial brain* with a huge number of connections. The language engine will exploit this structure and provide effective linguistic abilities.

Research Challenges. A relevant aspect of the integrated architecture will be the ability of creating *a single model encompassing all knowledge about a language*, instead of relying on half a dozen models, one for each task.

The unified model instead will be learned automatically by absorbing large quantities of text and it will be enriched and grow continuously by the addition of new textual documents, obtained from the many available sources both public and private (e.g. mail).

Suitable machine learning techniques will be required, for instance *Deep Learning* techniques for building the engine models, extending those used in the DeSR parser [6].

A second essential aspect of a language engine should be the ability of *continuously learning* from additional sources. While an initial model can be learned from labeled corpora, the engine should be capable of absorbing new information from unlabeled data that it reads subsequently. A suitable approach for learning from unlabeled data is by means of *Self Training*. Self Training is a semi-supervised learning method where the training corpus is extended with selected unannotated data that the system itself has tagged and that are considered as sufficiently accurate.

The research would aim to develop an efficient *incremental model representation* so that the model can be updated with new evidence, rather than having to retrain a model from scratch.

Achievements. Our team has developed state of the art techniques and solutions for text analytics, information extraction, indexing and search and question answering.

Our systems have achieved top scores in competitions such as CoNLL 2006, 2007, 2008, TREC Question Answering 2000, 2001, 2002, TREC Terabyte 2004, 2005, TREC Blog Mining 2006, Evalita 2007, 2009.

In particular we mention:

- The DeSR dependency parser.
- The Tanl linguistic pipeline, including POS, NER, Super Sense taggers and Coreference resolution.
- The DeepSearch semantic search engine.
- The semantically annotated Wikipedia and applications using it, e.g. Yahoo! Correlator.
- Linguistically directed browsing: Yahoo! Quest.

Demos of these systems are available from http://www.di.unipi.it/~attardi.

References

1. G. Attardi. Experiments with a Multilanguage Non-projective Dependency Parser. In Proc. of the Tenth CoNLL. 2006.
2. G. Attardi, M. Simi, Blog Mining Through Opinionated Words, *Proc. of The Fifteenth Text Retrieval Conference (TREC 2006)*, NIST, Gaithersburg (MD), 2006.
3. G. Attardi, A. Chanev, M. Ciaramita, F. Dell'Orletta and M. Simi. Multilingual Dependency Parsing and Domain Adaptation using DeSR. *Proc. the CoNLL Shared Task Session of EMNLP-CoNLL 2007*. 2007.
4. G. Attardi, M. Simi. DeSR at the Evalita Dependency Parsing Task *Proc. of Workshop Evalita 2007. Intelligenza Aritificiale*, 4(2). 2007.
5. G. Attardi, F. Dell'Orletta. Reverse Revision and Linear Tree Combination for Dependency Parsing. *Proc. of NAACL HLT 2009*. 2009.
6. G. Attardi, F. Dell'Orletta, M. Simi, J. Turian. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. *Proc. of Workshop Evalita 2009*. 2009.
7. G. Attardi, S. Dei Rossi, F. Dell'Orletta, E.M. Vecchi. The Tanl Named Entity Recognizer at Evalita 2009. *Proc. of Workshop Evalita 2009*. 2009.
8. G. Attardi, S. Dei Rossi, F. Dell'Orletta, E.M. Vecchi. Experiments in tagger combination: arbitrating, guessing, correcting, suggesting. *Proc. of Workshop Evalita 2009*. 2009.
9. G. Attardi, F. Dell'Orletta. Chunking and Dependency Parsing. *Proc. of LREC 2008 Workshop on Partial Parsing*, Marrakech, 2008.
10. C. Bosco, A. Mazzei, V. Lombardo, G. Attardi, A. Corazza, A. Lavelli, L. Lesmo, G. Satta, M. Simi. Comparing Italian parsers on a common treebank: the Evalita experience. *Proc. of LREC 2008*, Marrakech, 2008.
11. M. Ciaramita, G. Attardi, F. Dell'Orletta and M. Surdeanu. DeSRL: A Linear-Time Semantic Role Labeling System. *Proceedings the Twelfth Conference on Natural Language Learning*, Manchester, 2008.
12. H. Zaragoza, J. Atserias, M. Ciaramita, and G. Attardi. Semantically annotated snapshot of the English Wikipedia, *Proceedings of LREC 2008*, Marrakech, 2008.
13. Attardi, G., et al.: Tanl (Text Analytics and Natural Language Processing): Analisi di Testi per il Semantic Web e il Question Answering. http://medialab.di.unipi.it/wiki/SemaWiki. 2008.

14. G. Attardi, S. Dei Rossi, G. Di Pietro, A. Lenci, S. Montemagni, M. Simi. A Resource and Tool for Super-sense Tagging of Italian Texts. *Proceedings of LREC 2010*, Malta. 2010.
15. J. Atserias, G. Attardi, M. Simi, H. Zaragoza. Active Learning for Building a Corpus of Questions for Parsing. LREC 2010.
16. G. Attardi, D. Li. Extending a Dependency Treebank with Self-Training. Submitted.
17. G. Attardi, S. Dei Rossi, G. Di Pietro, A. Lenci, S. Montemagni, M. Simi. A Resource and Tool for Super-sense Tagging of Italian Texts. LREC 2010.
18. R. Collobert, J. Wetson. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. Proc. of the 25th Int. Conference on Machine Learning, Helsinki, Finland. 2008.
19. DARPA. DARPA-BAA-09-03 Machine Reading Broad Agency Announcement (BAA), http://www.darpa.mil/IPTO/solicit/baa/BAA-09-03_PIP.pdf. 2009.
20. G. Attardi, S. Dei Rossi, M. Simi. The Tanl Coreference Tagger at SemEval 2010. *Proc. of SemEval 2010*, 2010.

Modelling, Simulation and Verification of Biological Systems

Roberto Barbuti, Giulio Caravagna, Chiara Bodei, Andrea Bracciali, Pierpaolo Degano, Peter Drabik, Roberta Gori, Francesca Levi, Andrea Maggiolo–Schettini, Roberto Marangoni, Paolo Milazzo, Giovanni Pardini, and Aureliano Rama

Dipartimento di Informatica, Università di Pisa,
Largo B. Pontecorvo, 3, I-56127, Pisa, Italy

Cell biology, the study of the morphological and functional organization of cells, is now an established field in biochemical research. Computer Science can help research in cell biology in several ways. For instance, it can provide biologists with models and formalisms capable of describing and analyzing complex systems such as cells. In the last few years many formalisms originally developed to model systems of interacting components, such as Petri Nets, Hybrid Systems, and the π-calculus, have been applied to Biology. New formalisms, such as the Brane calculi, the *kappa* calculus, the Beta binders and Bioambients, have been defined for describing biomolecular and membrane interactions. The P systems, proposed as biologically inspired computational models, have been later applied to the description of biological systems.

CLS, the Calculus of Looping Sequences, has been proposed as a formalism with a simple notation, having the ability of describing biological systems at different levels of abstraction, having some notion of compositionality and being flexible enough to allow the description of new kind of phenomena without being specialised to the description of a particular class of systems Some variants and properties of CLS have been studied in [18, 10, 17].

P systems allow describing biological membranes and the movement of molecules across membranes. Variants of P systems and their properties have been studied in [8, 11–16].

As regards process calculi, a great effort has been devoted to adapt traditional models for describing the molecular and biochemical aspects of biological systems. There is a particular interest in the design of calculi able to capture the *quantitative* aspects (both time and probability) of real life applications, in the style of stochastic π-calculus.

The use of the mentioned formalisms offers the possibility of applying to models of biological systems a variety of automated methods for analysis and verification. The information on system dynamics obtained in this manner could be profitably used for testing hypotheses of biologists and for guiding *in vitro* experiments. Simulators explore a given behaviour of the biological system model over a given time interval, thus realizing a *virtual experiment*. In this approach, statistically relevant results can be deduced by running several simulations.

An example of this is the formal specification of VICE (VIrtual simplified prokaryotic Cell), a hypothetical prokaryote with a minimal genomic set, but

close to a real cell, whose metabolic pathways are specified *in silico*, in terms of a π-calculus with a stochastic semantics [23]. Simulation results are coherent with *in vitro* experiments.

Similarly, a stochastic and discrete model of the Calyx of Held synapse has been proposed and simulated, that it is shown to be expressive enough to accurately capture the modelled phenomena [5].

Another approach is based on the verification of (quantitative) temporal properties by means of model checking techniques. The tool PRISM, for example, supports the validation of probabilistic or stochastic models formalised as a *Discrete-Time Markov Chain* (DTMC) or as *Continuous-Time Markov Chain* (CTMC). Quantitative temporal properties over all the runs of a biological system model can be expressed in the probabilistic logic PCTL or in the continuous-time logic CSL. Properties are able to capture important aspects of the behaviuor of biological systems.

One specific feature of biological processes is that they are composed by a huge number of processes with identical behaviour, such as thousands of molecules of the same type. Thus, the state space of the model is often very large (or even infinite). Moreover, the experimental data concerning the biological model are typically not precisely known.

Approximation techniques, based on the Abstract Interpretation theory, have been established to be one of the most effective ways for overcoming these limitations. Specifically, abstraction techniques able to deal with the typical uncertainty of biological systems, have been investigated. The abstractions proposed in [26, 22] support the validation of probabilistic reachability properties.

Another approximation technique is Control Flow Analysis (CFA), that is applied to calculi for modelling biological systems, such as Beta-binders and Brane calculi [2, 3]. As far as biological properties are concerned, a taxonomy of properties of metabolic networks, based on causality relationships, has been proposed in [4].

An extension of the approach for modelling biological systems where the rates of reactions are not precisely known but may vary over intervals, has been proposed in [1].

References

1. R. Barbuti, F. Levi, P. Milazzo and G. Scatena. *Probabilistic Model Checking of Biological Systems with Uncertain Kinetic Rates.* Proc. of RP '09, LNCS 5797, pp. 68–74, 2009.
2. C. Bodei. A Control Flow Analysis for Beta-binders with and without Static Compartments. Theoretical Computer Science 410(33-34): 3110-3127, 2009.
3. C. Bodei, A. Bracciali, D. Chiarugi. Control Flow Analysis for Brane Calculi. Proc. of MeCBIC'08, ENTCS 227: 59-75, 2009.
4. C. Bodei, A. Bracciali, D. Chiarugi, R. Gori. A Taxonomy of Causality-Based Biological Properties. To appear in Proc. of FBTC'10.
5. A. Bracciali, M. Brunelli, E. Cataldo and P. Degano. Formal models of the calyx of Held. In A. Condon, D. Harel, J.N. Kok, A. Salomaa, and E.Winfree (Eds.) Algorithmic Bioprocesses. Natural Computing Series, 2009, 331–366.

6. R. Barbuti, G. Caravagna, A. Maggiolo-Schettini, P. Milazzo. An intermediate language for the simulation of biological systems. Proc. of FBTC'07. ENTCS(194): 19–34, 2007.
7. R. Barbuti, G. Caravagna, A. Maggiolo-Schettini, P. Milazzo. An Intermediate Language for the Stochastic Simulation of Biological Systems. Theoretical Computer Science 410 (2009), pp. 3085-3109.
8. R. Barbuti, G. Caravagna, A. Maggiolo-Schettini, P. Milazzo. P Systems with Endosomes. Int. J. of Computers, Comm. & Control, IV(3): 214–223, 2009.
9. R. Barbuti, S. Cataudella, A. Maggiolo-Schettini, P. Milazzo, A. Troina. A probabilistic model for molecular systems. Fundamenta Informaticae **67**: 13–27, 2005.
10. R. Barbuti, A. Maggiolo-Schettini, P. Milazzo. Extending the calculus of looping sequences to model protein interaction at the domain level. Proc. of ISBRA'07. LNBI 4463, pp.638–649, 2006.
11. R. Barbuti, A. Maggiolo-Schettini, P. Milazzo, G. Pardini, L. Tesei. Spatial P Systems. Natural Computing, to appear.
12. R. Barbuti, A. Maggiolo-Schettini, P. Milazzo, L. Tesei. Timed P Automata. Fundamenta Informaticae, 94(1), 1–19, 2009.
13. R. Barbuti, A. Maggiolo-Schettini, P. Milazzo, S. Tini. A P Systems Flat Form Preserving Step-by-step Behaviour. Fundamenta Informaticae, 87(1): 1–34, 2008.
14. R. Barbuti, A. Maggiolo-Schettini, P. Milazzo, S. Tini. Compositional semantics and behavioral equivalences for P Systems. Theoretical Computer Science 395(1): 77–100, 2008.
15. R. Barbuti, A. Maggiolo-Schettini, P. Milazzo, S. Tini. On the Efficiency of Promoters and Cooperating Rules in P Systems, in: G. Paun, M.J. Perez-Jimenez, A. Riscos Nunes (Eds.), Tenth Workshop on Membrane Computing, 543–546, 2009.
16. R. Barbuti, A. Maggiolo-Schettini, P. Milazzo, S. Tini. P Systems with Transport and Diffusion Membrane Channels. Fundamenta Informaticae 93(1-3): 17–31, 2009
17. R. Barbuti, A. Maggiolo-Schettini, P. Milazzo, P. Tiberi, A. Troina. Stochastic CLS for the Modeling and Simulation of Biological Systems. Transactions on Computational Systems Biology IX, 5121, 86-113, 2008.
18. R. Barbuti, A. Maggiolo-Schettini, P. Milazzo, A. Troina. A calculus of looping sequences for modelling microbiological systems. Fundamenta Informaticae **72**: 21–35, 2006.
19. R. Barbuti, A. Maggiolo-Schettini, P. Milazzo, A. Troina. Bisimulation congruences in the calculus of looping sequences. Proc. ICTAC'06. LNCS 4281, 93–107, 2006.
20. R. Barbuti, A. Maggiolo-Schettini, P. Milazzo, A. Troina. Bisimulations in Calculi Modelling Membranes. Formal Aspects of Computing, 20, 351–377, 2008
21. L. Brodo, P. Degano, C. Priami. A stochastic semantics for bioambients. Proc. of PaCT'07. LNCS 4671, 22–34, 2007.
22. A. Coletta, R. Gori, F. Levi. Approximating probabilistic behaviours of biological systems using abstract interpretation. Proc. of FBTC '08, ENTCS 229 (1): 165–182, 2009.
23. D. Chiarugi, P. Degano, R. Marangoni. A computational approach to the functional screening of genomes. PLoS Computational Biology, 3(9), 2007.
24. M. Curti, P. Degano, C. Priami, C. Baldari. Modelling biochemical pathways through enhanced pi-calculus. Theoretical Computer Science **325**(1): 111–140, 2004.
25. P. Degano, D. Prandi, C. Priami, P. Quaglia. Beta-binders for biological quantitative experiments. Proc. of QAPL 2006. ENTCS 164: 101–117, 2006.
26. R. Gori and F. Levi. Abstract Interpretation for Probabilistic Termination of Biological Systems. Proc. of MeCBIC'09, EPTCS (11), pp. 137-153, 2009.

Quantitative and Automatic Analysis of Neurological Signals and Images of Cognitive Interest

Umberto Barcaro

joint work with Ovidio Salvetti[1]
[1]Institute of Information Science and Technologies, C.N.R., Pisa

Abstract. A research project is briefly described aiming at the introduction and application of mathematical and automatic methods to the study of the different levels of consciousness in the human brain. The first part of this communication includes a short review of two approaches recently proposed in the field of neurosciences. The second part summarizes three data analysis methods that appear as possibly powerful tools for the investigation. These methods regard one-dimensional signals, images, and textual data, respectively.

Keywords: Signal Processing, Image Processing, Text Analysis, Cognitive Science, Electroencephalogram, Consciousness.

This communication briefly introduces a research project which involves researchers of the Group of Computational Intelligence & Machine Learning of the Computer Science Department and researchers of the Signals & Images Laboratory of the Institute of Information Science and Technologies of the National Research Council. This project aims at the proposal and implementation of mathematical and automatic methods for the study of different levels of consciousness in the human brain.

This communication is divided into two parts.

The first part is dedicated to a brief review of two approaches that have been recently proposed in the rapidly advancing research field regarding the neural basis of consciousness. Among the numerous important contributions, we have chosen these two approaches because we consider them as particularly significant for our project. We will add a short reflection on the primary role played by methods proper of computer science for the cognitive study of consciousness levels.

The second part summarizes three data analysis techniques that we have implemented and can be viewed as promising tools for the research project. These techniques were proposed in three different environments: (a) the "Biopatterns" Project, in which the Group of Computational Intelligence & Machine Learning participated under the guide of Professor Antonina Starita; (b) the "Heartfaid" Project, whose consortium included the Signals & Images Laboratory; (c) collaborations in cognitive research established within the International Association for the Study of Dreams.

We now shortly describe the two chosen instances of important recent approaches.

The first example is given by a study of the brain responses to sensory stimuli, specifically acoustic stimuli, performed by a French group [1]. The Authors analyzed the responses to rare deviant stimuli delivered within a serial flow of frequent standard stimuli. They simultaneously recorded Functional Magnetic Resonance Images (fMRI), Intracerebral Local Field Potentials, and Event-Related Potentials. The experimental protocol aimed at separating responses due to neural systems characterized by different levels of consciousness. Violations of local regularity elicited effects "suggestive of an automatic, nonconscious, and encapsulated mode of processing". Violations of global regularity evoked a very different response given by conscious brain subsystems. The differences regarded waveform as well as time latency.

The other article we now consider is a review paper by Raichle, a neurologist at the Washington University in St. Louis, describing a "paradigm shift in functional brain imaging" [2]. This paper focuses on researches that showed, by means of fMRI, that a set of brain regions, called "Default Mode Network" (DNN), decreased their activity across a wide array of task conditions with respect to passive control conditions. Blood-Oxygen-Level Dependent (BOLD) fluctuations recorded from distant brain areas belonging to the DMN presented a high correlation. The DMN system therefore appears as providing a background for the emergence of conscious activity.

For both of these examples, the processing of data required advanced methods for feature extraction, data combination, data representation, and statistical analysis.

We now mention three methods for signal and image processing that can be useful for our project.

For the purposes of the "Biopattern" Project, we implemented a method for the detection of transient events in spontaneous electroencephalogram (EEG) [3]. After a computation of signal frequency components, performed according to conventional frequency band assignments, an event was defined as a transient increase in the activity amplitude in any frequency band. Another kind of events consisted in transient decreases in the correlation between pairs of EEG traces. The detected events, together with their main attributes, were stored as records in a database. The statistical analyses were carried out on the results of queries to this database.

For the purposes of the "Heartfaid" Project we implemented a method for image segmentation [4] that can also applied to neuroimages. The segmentation procedure consisted of two stages: the application of local thresholding based on criteria mimicking visual analysis, and the construction of a contour through an active contour method consisting in the minimization of a suitable "Energy Function". The evolving contour was represented as the zero-level of a time-varying function defined over the image.

In addition to one-dimensional data and images, we proposed methods for processing textual data of cognitive interest [5]. In particular, a method for the recognition of possible links between the various memory sources of sleep mentation was applied to dream reports and to associations with the various dream items provided by subjects woken from the Rapid-Eye-Movement stage.

1. Bekinschtein, T.A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., Naccache, L. *Proceedings of the National Academy of Sciences* 106, 1672-1677 (2009).
2. Raichle, M.E. *Journal of Neuroscience* 29, 12729-12734 (2009).
3 Righi, M., Barcaro, U., Starita, A., Karakonstantaki, E., Micheloyannis, S. *Brain Topography*, 21, 43-52 (2008).
4. Barcaro, U., Moroni, D., Salvetti, O. *Pattern Recognition and Image Analysis* 18, 351-358 (2008).
5. Barcaro, U., Rizzi, P. *Dreaming* 18, 139-157 (2008).

Mobilome inference in yeast genomes

Giovanni Battaglia, Roberto Grossi, Roberto Marangoni, Nadia Pisanti,

joint work with Giulia Menconi

Context. Mobile genomic elements (collectively *mobilome*) are found in large number in eukaryotic genomes: they represent between 15% and 20% of the total human DNA. Their relationships with the resident genome can be viewed as a competition of different species in an ecosystem. Most of the mobilome is constituted by transposons, which are sequences able to replicate and jump over the genome. Transposons can cause phenotypic variations between individuals and between cells in the same individual. Indeed, some complex pathologies whose molecular mechanisms and global inheritance are hard to explain by common inheritance laws, turned out to be correlated to transposons' translocations. A contemporary challenge in comparative genomics aims at understanding the dynamics of the mobilome, so as to design a model able to describe (and even forecast) the logic followed by mobile elements to decide when and where to transpose. To this aim, it is necessary to study different lineages of organisms, to identify and locate all the mobile elements on the whole genome, and to compare the obtained results. This task is computationally challenging since the classical alignment has two main drawbacks: a) alignment algorithms do not perform efficiently on large repositories of whole chromosomes in practice; b) most of sequenced strains have unresolved regions exactly in correspondence with transposable elements. In this work, we face these drawbacks using fast algorithmic techniques that we have experimented for the analysis of different yeast strains' genomes made publicly available [1].

Methods. The yeast genome contains 16 chromosomes. Our preliminary hypothesis is that the major chromosomal differences are caused by transposons movements, since the chromosomal mutations between different strains of the same specie occur mostly for these reasons. The high similarity in the available data allows us to compare the same chromosome in two different strains by searching for L-grams shared by the two chromosomes, say, S and T, thus creating a map of these correspondences. Then, we join the L-grams located within a given distance threshold into larger runs. This final map allows us to detect insertions or deletions existing between S and T. We employ hashing based on cyclic polynomials in order to search for all the L-grams of T; this turns out to be very effective on our datasets in practice. It processes a whole chromosome in just 6.5s with the longest sequence (IV, 1.5Mb) on a standard PC.

Results. We compared all the chromosomes of the yeast RefSeq@SGD to the corresponding chromosomes of two yeast strains, Y55 and YPS128. Our approach is able to identify chromosomal mutations: Figure 1 shows transposons deletions for chromosomic regions in RefSeq@SGD and the strains Y55 and

YPS128. Here, the correspondences between L-grams are represented by green straight lines; deletions with respect to RefSeq@SGD appear as downward white triangles, and insertions as upward white triangles. Black rectangles on the bottom sequences indicate runs of unresolved bases (N). A detailed count of the number of chromosomal mutations for all the 16 chromosomes in the two strains is provided. Our results fully justify the initial hypothesis: almost all the detected mutations are indeed related to mobile elements annotated in RefSeq@SGD. The few indels apparently not related to the mobilome can be attributed to genomic rearrangements or to un-annotated mobile elements. Moreover, our results highlight that the genomic segments left unassigned after the genome sequencing are almost always represented by mobile elements. This is caused by two reasons: (a) like all repeats, mobile elements are hard to be exactly located during sequencing; (b) since RefSeq is used as a reference when assembling, a failure may occur for transposons which are not annotated in RefSeq.

Future. We plan to apply our method to pairwise comparison of all strains of yeast presented in [1]. As a second step, we can automatically classify mobilome events and locate them in the hypothetical phylogeny reconstruction given in [1]. This could also possibly lead top detect errors or inconsistencies in such tree and in such case an alternative guess could be done based on our classification. Also, a systematic analysis could allow us to estimate the jumping frequency of transposons. We view this study on Yeast strains as a proof-of-concept, but actually our final goal is to design and apply a similar method to the Human genome.

References

1. Gianni Liti et al. Population genomics of domestic and wild yeasts. *Nature* **458**, 337–341, 2009.

Data Mining Meets Switching Theory

Anna Bernasconi, Fabrizio Luccio, and Linda Pagli

joint work with Valentina Ciriani

Studying associations among the items occurring jointly in a set of transactions is of paramount importance for conducting business of many kinds. Data mining is the main area in which such studies have been developed, where knowledge of the phenomena involved and related computational methods have reached maturity through a wealth of significant publications, e.g. see [1, 10, 7, 5], or the comprehensive bibliography of [3]. A point that has probably to be examined in more depth is the joint study of the items present *and absent* in a transaction, as brought to the general attention in [8].

We propose an innovative way at describing and processing transactions, whose origin comes from the theory of digital circuits. The two *products*:

$$\overline{a}\, b\, \overline{c}\, \overline{d}\, \overline{e}\, \overline{f}\, g\, \overline{h}\, \overline{i}\, \overline{j}\,, \qquad \overline{a}\, b\, \overline{c}\, \overline{d}\, \overline{e}\, \overline{f}\, g\, \overline{h}\, \overline{i}\, j$$

will indicate the two transactions consisting of items b, g and b, g, j, respectively. The product (same as before, with a missing variable j):

$$\overline{a}\, b\, \overline{c}\, \overline{d}\, \overline{e}\, \overline{f}\, g\, \overline{h}\, \overline{i}$$

is obtained as the *sum* of the two products above and accounts for the pair. I.e., products of variables will represent transactions or sets of transactions, where the absence of items, corresponding to *complemented* variables, appears as a byproduct of the notation. Two points, however, have to be clarified immediately. As the possible items are generally many more than the ones contained in a transaction, an efficient notation to avoid long chains of complemented variables in a product must be adopted. Second, the same transaction, or set of transactions, may appear more than once. In this case an integer coefficient will be appended to the corresponding product, thus requiring an extension of switching theory from Boolean to integer algebra. This will also lead to the introduction of a new problem in circuit design, and of a computational method for its solution.

References

1. A. Agrawal, H. Mannila, S. Srikant, H.‘ Toivonen, and A.I. Verkamo, Fast Discovery of Association Rules, *Advances in Knowledge Discovery and Data Mining*, MIT Press (1998) 307-328.
2. A. Bernasconi, V. Ciriani, F. Luccio, and L. Pagli, A New Heuristic for DSOP Minimization, Proc. 8th International Workshop on Boolean Problems (2008).
3. J. Han and M. Kamber, *Data Mining: Concept and Techniques* (2nd Edition), Bibliographic Notes for Chapter 5, Morgan Kauffman Publ. (2006).
4. J. Han, J. Pei, and Y. Yin, Mining Frequent Patterns without Candidate Generation, SIGMOD 2000, (2000) 1-12.

5. A. Kirsch, M. Mitzenmacher, G. Pucci, A. Pietracaprina, E. Upfal, and F. Vandin, An Efficient Rigorous Approach for Identifying Statistically Significant Frequent Itemsets *Proc. 27th ACM Symposium on Principles of Database Systems (PODS)* (2009) 117-126
6. S. Minato and H. Arimura, Frequent Closed Item Set Mining Based on Zero-suppressed BDDs, Information and Media Technologies 2(1) (2007) 309-316.
7. A. Pietracaprina and F. Vandin, Efficient Incremental Mining of Top-K Frequent Closed Itemsets, In: V. Corrube et al., (Eds): *DS2007, LNCS (LNAI)* 4755 (2007) 275-280.
8. C. Silverstein, S. Brin, and R. Motwani, Beyond Market Baskets: Generalizing Association Rules to Dependence Rules, *J. Data Mining and Knowledge Discovery* 2 (1) (1998) 36-68.
9. T. Uno, T. Asai, Y. Uchida, and H. Arimura, LCM: An Efficient Algorithm for Enumerating Frequent Closed Item Sets, In *Proc. IEEE ICDM03 Workshop FIMI03* (2003).
10. T. Uno, T. Asai, Y. Uchida, and H. Arimura, An Efficient Algorithm for Enumerating Closed Patterns in Transaction Databases, In: E. Suzuki and S. Arikawa, (Eds): *DS2004, LNCS (LNAI)* 3245 (2004) 16-31.
11. T. Uno, M. Kiyomi, and H. Arimura, LCM ver. 2: Efficient Mining Algorithms for Frequent/Closed/Maximal Itemsets, In *Proc. IEEE ICDM'04 Workshop FIMI'04* (2004).
12. T. Uno, M. Kiyomi, and H. Arimura, LCM ver.3: Collaboration of Array, Bitmap and Prefix Tree for Frequent Itemset Mining, In *Proc. of the 1st International Workshop on Open Source Data Mining OSDM'05* (2005), 77-86.
13. J. Wang, J. Han, Y. Lu, and P. Tzvetkov, TFP: An Efficient Algorithm for Mining Top-K Frequent Closed Itemsets, *IEEE Transactions on Knowledge and Data Engineering* 17 (2005) 652-664.
14. G. Yang The Complexity of Mining Maximal Frequent Itemsets and Maximal Frequent Patterns. *Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2004) 344-353.

Parallel programming issues, achievements and trends in high-performance and adaptive computing

Marco Danelutto, Marco Vanneschi, C. Bertolli, D. Buono, A. Mencagli, A. Pascucci

joint work with: M. Aldinucci, P. Kilpatrick, M. Meneghin, R. Ravazzolo, M. Torquati

The area of parallel high-performance programming is currently re ceiving a growing attention, owing to the technological (r)evolution of computer components and CPUs based on single-chip multi/many-core architectures. This phenomenon will imply radical changes in IT community and industry attitude and products, as the sequential programming model assumption is put into a rather difficult position and, in some way, it could be replaced by the parallel programming model assumption [7].

Currently, it is difficult to foresee the implications of this (r)evolution on programming methodologies, tools and frameworks on the software technology and applications of the next future. What we can assert is that future applications must be able to effectively exploit the parallel structure of emerging high-performance architectures, while the existing applications should be able at least to exploit their scalability features.

In the area of parallel programming we have assisted to several approaches with different goals and characteristics: from low-level and low-productivity approaches based on communication libraries (MPI, OpenMP), to high-level approaches specialized towards some classes of application structures (HPF, HPC). The most notable attempts to achieve a good trade-off between programmability, software productivity, portability and performance belong to what we call the *Structured Programming Model* class, like algorithmic skeletons, parallel paradigms, and design patterns [9, 14]. Our research group has produced many recognized results in this area, notably P3L/SkIE [8], ASSIST [15], Lithium [1], Muskel [3], a first adaptive (w.r.t. parallelism degree) version ASSIST [6], as well as contribution to high-performance software component technology with the Behavioural Skeleton concept [2], for several kinds of architectural platforms (multiprocessors, homogeneous and heterogenous clusters, high-performance grids).

Recently, some new structured programming frameworks are emerging, e.g. Google MapReduce and Apache Hadoop, Intel TBB (Thread Building Blocks library), as well as the recent Microsoft TPL (Task Parallel Library), though characterized by some limitations as far as concerns expressive power and software productivity in complex application development.

The main feature of the *Structured Parallel Programming Model* lies in its ability to define a parallel computation as the composition of parametric parallel paradigms with known implementation models and cost models. This feature has been extended in the ASSIST programming model to any computation

structure described by generic graphs whose nodes apply the parametric parallel paradigms, thus allowing the programmer to express arbitrarily complex and irregular compositions of parallel structured components. It also been exploited in the Behavioural Skeletons to support composition of autonomic management of non functional features (e.g. performance, security, fault tolerance and power management) in addition to functional composition of "business logic" code.

The experience with Structured Parallel Programming has shown that a static optimization approach based on analytical cost models is effective only on some, though notable application classes. There are cases in which the dynamic or irregular structure of the computation, as well as the heterogeneous nature of the underlying platform, prevent to achieve good optimizations according to a static approach. Interesting examples are currently investigated in the FIRB In.Sy.Eme project [11] in the context of heterogeneous platforms for emergency management, based on a variety of fixed and mobile nodes and networks.

These classes of applications are characterized by the adaptive capability of dynamically restructuring the parallel applications in order to respect given QoS requirements which, in turn, can change dynamically during the computation and according to contextual situations. The autonomic view of systems and applications [13] corresponds to this approach towards adaptive and context-aware parallel programming. Some notable results have been achieved in the European projects CoreGrid [10] and GridComp [12]. In these projects. and in the follow-up research activities, we demonstrated that rule based autonomic managers can be associated to component based structured parallelism exploitation patterns. These managers allow to ensure, best effort, single, user provided, non functional contracts (e.g. throughput or secure computation requirements) in hierarchical composition of parallel patterns [4], as well as coordinated ensuring of multiple non functional contracts, each one related to a different non functional concern [5].

In this presentation we highlight the relationship between Structured Parallel Programming and adaptive/autonomic computing, report on the current status and experience, and discuss some interesting trends in this area, including methodologies, tools and approaches for dynamically adaptive model of high-performance computations.

References

1. M. Aldinucci, M. Danelutto, and P. Teti. An advanced environment supporting structured parallel programming in java. *Future Gener. Comput. Syst.*, 19(5):611–626, 2003.
2. Marco Aldinucci, Sonia Campa, Marco Danelutto, Patrizio Dazzi, Peter Kilpatrick, Domenico Laforenza, and Nicola Tonellotto. Behavioural skeletons for component autonomic management on grids. In Marco Danelutto, Paraskevi Frangopoulou, and Vladimir Getov, editors, *Making Grids Work*, CoreGRID. Springer Verlag, May 2008.
3. Marco Aldinucci, M. Danelutto, and Patrizio Dazzi. Muskel: an expandable skeleton environment. *Scalable Computing: Practice and Experience*, 8(4):325–341, December 2007.

4. Marco Aldinucci, Marco Danelutto, and Peter Kilpatrick. Autonomic management of non-functional concerns in distributed & parallel application programming. In *IPDPS '09: Proceedings of the 2009 IEEE International Symposium on Parallel&Distributed Processing*, pages 1–12, Washington, DC, USA, 2009. IEEE Computer Society.
5. Marco Aldinucci, Marco Danelutto, Peter Kilpatrick, and Vamis Xhagjika. LIBERO: a LIghtweight BEhaviouRal skeletOn framework. Technical Report TR-10-07, Dept. Computer Science, Univ. of Pisa, Italy, April 2010. available at http://compass2.di.unipi.it/TR/Files/TR-10-07.pdf.gz.
6. Marco Aldinucci, Marco Danelutto, and Marco Vanneschi. Autonomic qos in assist grid-aware components. In *PDP '06: Proceedings of the 14th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*, pages 221–230, Washington, DC, USA, 2006. IEEE Computer Society.
7. Krste Asanovic, Rastislav Bodik, James Demmel, Tony Keaveny, Kurt Keutzer, John Kubiatowicz, Nelson Morgan, David Patterson, Koushik Sen, John Wawrzynek, David Wessel, and Katherine Yelick. A view of the parallel computing landscape. *Commun. ACM*, 52(10):56–67, 2009.
8. Bruno Bacci, Marco Danelutto, Salvatore Orlando, Susanna Pelagatti, and Marco Vanneschi. P^3L: a structured high level programming language and its structured support. *Concurrency Practice and Experience*, 7(3):225–255, May 1995.
9. Murray Cole. Bringing skeletons out of the closet: A pragmatic manifesto for skeletal parallel programming. *Parallel Computing*, 30(3):389–406, 2004.
10. Coregrid home page, 2008. http://www.coregrid.net.
11. Romano Fantacci, Marco Vanneschi, Carlo Bertolli, Gabriele Mencagli, and Daniele Tarchi. Next generation grids and wireless communication networks: towards a novel integrated approach. *Wirel. Commun. Mob. Comput.*, 9(4):445–467, 2009.
12. Gridcomp home page, 2009. http://gridcomp.ercim.org, autonomic part at http://gridcomp.ercim.eu/content/view/26/34/.
13. Markus C. Huebscher and Julie A. McCann. A survey of autonomic computing—degrees, models, and applications. *ACM Comput. Surv.*, 40(3):1–28, 2008.
14. Timothy Mattson, Beverly Sanders, and Berna Massingill. *Patterns for parallel programming*. Addison-Wesley Professional, 2004.
15. Marco Vanneschi. The programming model of assist, an environment for parallel and distributed portable applications. *Parallel Comput.*, 28(12):1709–1732, 2002.

Policy-aware Service Composition

Chiara Bodei, Pierpaolo Degano, and Gian Luigi Ferrari

Dipartimento di Informatica, Università di Pisa, Pisa, Italy

The *Software-as-a-Service* metaphor enables an evolutionary computational style, where applications are built by "gluing" together suitable software units called *services*. Web services are the best illustrative example of this computational paradigm, whose key issue is given by its compositional nature. Services are pluggable entities and composite services are obtained by combining existing elementary or complex services. A challenging problem is that of properly selecting and configuring services so to guarantee that their coordinated behaviour enjoys desirable properties. These properties may involve *functional* aspects, and also *non-functional* aspects, like e.g. security, availability, performance, transactionality, quality of service, etc.. This poses significant theoretical and technical challenges. In particular, the ability of coordinating service interactions while guaranteeing certain properties requires a novel view of the interplay between local and global views of services. For instance, the component services may enjoy locally a property (e.g. a security property) while their coordination globally does not enforce it.

Our research aims at developing software engineering techniques and tools that support the creation of more reliable and secure services. In order to accomplish this goal, we are developing a semantics-based framework tackling the problem of service coordination in the presence of both functional and non-functional constraints. The starting point of our approach is λ_{req} [1, 2], a core calculus for securely orchestrating services. The λ_{req} calculus extends the λ-calculus with primitive constructs for describing and invoking services in a *call-by-contract* fashion. Services are modelled as functions with side effects. A service is not a *pure* function as its execution yields a side effect reflecting changes of the service state. For instance, let us consider a simple storage service that gets a string query from the user and accordingly queries the database. The interface `Table fun Q(Query q): Effect e` characterizes the service above. The invocation of the service `Q` with a query value will yield a table as result. The side effect `e` provides the abstract representation of the modifications to the database.

Side effects are sequences of resource accesses called *histories*, that abstractly describe the possible run-time traces of resource accesses A run-time security monitor may inspect histories, and forbids those executions that would violate the prescribed policies.

Unlike standard discovery mechanisms that match syntactic signatures only, our approach suggests and implements a matchmaking algorithm based on service behaviour. The algorithm exploits static analysis techniques to resolve the call-by-contract involved in a service coordination. Operationally, the service registry is searched for a service with a functional type (the service signature) matching the request type; also, the semantic abstraction must respect the non-

functional constraints imposed by the request. Our orchestration machinery constructs a *plan* for the execution of services, e.g. a binding between requests and service locations, guaranteeing that the properties on demand are always satisfied. Finding viable plans is not a trivial task, because the effect of selecting a given service for a request is not always confined to the execution of that service. Since each service selection may affect the *whole* execution, we cannot simply devise a viable plan by selecting services that satisfy the constraints imposed by the requests, only. The identification of viable plans is a result of a suitable model-checking activity.

Our machineries can be applied to handle a variety of service properties provided that they can be expressed as safety properties of execution histories. In [4, 3] we handled the case of (local) resource usage contracts for services, whereas [5] Java run-time has been extended to support our security policies. Finally, the exploitation of powerful static techniques [6, 7] allows us to strengthen the expressiveness of our framework.

The call-by-contract mechanism makes standard testing practice even more effective, e.g. one can perform a request with a given policy and observe the resulting plans. The system must then consider *all* the services that satisfy the policy, and the observed effect is similar to running a *class* of tests. For instance, a designer of an online bookshop can specify a policy such as "order a book without paying" and then inspect the generated plans: the presence of viable plans could point out an unwanted behaviour, e.g. due to an unpredicted interaction between different special offers. As a matter of facts, standard testing techniques are yet not sophisticated enough to spot such kind of bugs.

Our planning construction properly formalizes the *business logic* of service coordination, i.e. the flow of service invocations. More problematic is the management of the *operation logic* of service coordination, namely the actual management of the operations acting over the available resources. For instance, in a storage service that provides facilities to store and retrieve data over the Internet, suitable operations are required to scale up and down the computing resources automatically, depending on the actual needs. Similar considerations on resource provisioning, monitoring and evolution apply when cloud applications are considered.

We advocate the usage of a declarative model based on functions with side effect to abstractly represent services acting over resources, where static analysis techniques allow to integrate the loosely coupled nature of services with a coarser view of resource usages. We plan to extend our techniques to include resource logic and quantitative information on resource usages. The novelty of our approach derives from the combination of static analysis techniques (types) and model checking effects. For instance, to add quantitative information it suffices to extend the notion of effect. In this perspective, the side effects of services will take the form of *stochastic processes* that include the rate of an exponentially distributed random variables that characterize the duration of the operations over resources.

References

1. M. Bartoletti, P. Degano, G. Ferrari, and R. Zunino. Semantics-based design for secure web services. *IEEE Trans. Software Eng.*, 34(1):33–49, 2008.
2. M. Bartoletti, P. Degano, and G. Ferrari. Planning and verifying service composition. *Journal of Computer Security*, 17 (5), 2009.
3. , M. Bartoletti, P. Degano, G. Ferrari, and R. Zunino. Model Checking Usage Policies Trustworthy Global Computing, Lecture Notes in Computer Science, 5474, 2009.
4. M. Bartoletti, P. Degano, G. Ferrari, and R. Zunino. Local policies for resource usage analysis, *ACM Trans. Program. Lang. Syst.*,31 (6), 2009.
5. M. Bartoletti, G. Costa, P. Degano, F. Martinelli, and R. Zunino. Securing Java with Local Policies, Journal of Object Technology, 8(4) 2009.
6. C. Bodei, L. Brodo, R. Bruni. Static Detection of Logic Flaws in Service-Oriented Applications. Proceedings of the International Workshop on Issues in the Theory of Security (WITS'09), LNCS 5511, pp. 70-87, 2009.
7. C. Bodei, L. Brodo, P. Degano, H. Gao. Detecting and preventing type flaws at static time. To appear in Journal of Computer Security, 2010

Teaching Computer Science at School: some ideas

Chiara Bodei, Roberto Grossi, and Maria Rita Laganà

joint work with
Marco Righi

Abstract. As a young discipline, Computer Science does not rely on longly tested didactic procedures. This allows the experimentation of innovative teaching methods at schools, especially in early childhood education. Our approach is based on the idea that abstracts notions should be gained as the final result of a learning path made of concrete and touchable steps. To illustrate our methodology, we present some of the teaching projects we proposed.

1 Our Didactic Projects

Robot [6, 4] Using and building robots is one of the most effective and consolidated activities to naturally initiate children to computer programming. It is sufficient to think how easily they understand that a robot without a program is a useless object and the program is the only thing able to "give life" to it. Programming concepts could be very difficult to teach without any reference to concrete things. Pupils are not aware that they are learning and implicitly using not trivial concepts like the one of algorithm. In particular, we played with the Lego programmable robots and we helped children to design them in order to simulate behaviour that recall "biological" ones, like: be frightened of the light, follow a clue, according to an imprinting phenomenon, find the way out of a labirinth, follow a light, as moths do, produce malicious look that invites to cuddle. In addition, robot construction offered a chance to cooperate, and realise how important is to work together as a team, like in a real team of industrial design. Since there are several required skills, pupils can find their own roles and feel involved and useful.

Danza dei bit [1, 3, 2] We have experimented a multidisciplinary approach to teach the binary representation of numbers and their arithmetic and also the basic notions on the computer architecture. We pretended that children had been turned in bits and lived in Computer City. As insiders, by acting in the imaginary town and by playing their roles, they could understand how numbers are formed, which are the parts a Computer is composed with and how it works. An operetta, called "Danza dei bit" (Bits' dance), concluded the experience. Children as actors have been taken to actively understand the target notions, with the help of dialogues, music and also the used sets.

AI-Game [5] Another didactic experimentation proposed to 9-10 year old pupils of the Primary School is based on a programmable LOGO-like environment,

called AI-Game, that we developed to smoothly introduce children to the art of programming. Without any programming experience, children can easily create drawings and animations. Four graphic characters can indeed move and act in a grid on the screen, following instructions, like: move forward, move backward, pick a ball, Children can compose programs, i.e. sequences of these instructions, in order to create simple games, consequently being for once behind and not in front of the video game console. Focussing on the construction of new games, children have mastered the basic programming notions quite naturally, because they represent just a mean to play. We only gave pupils the basic instructions and let them to ask us for new ones. It comes as a nice surprise that they also required something very similar to the standard conditional and iterative instructions. From their point of view, these composed instructions represented just a way to obtain more succinct programs.

Kara [7] Programming with finite state automata and data structures in the form of matrices is another way to develop and coordinate the abstraction capabilities of pupils: every automaton state represents a class of possible computations for the same configuration. Kara is a software tool, developed by W. Hartmann, J. Nievergelt and R. Reichert, that can be used to program the actions of a ladybird (Kara) on a bi-dimensional grid, by means of a finite state automaton. Movements can be towards the four directions; it is moreover possible to put tree logs as obstacles in crossing-points. The ladybird can put or get four-leaf clovers (bits), move mushrooms (like marks) and be aware of the objects it could be surrounded. This tool is simple, nevertheless it has the same expressive power of other computational models (it is Turing-complete). Its visual programming is easy to learn; as a consequence children can address not trivial computational problems, without mastering a particular programming language (that represents the next natural step, though). Once translated the documentation and the tool in Italian (from German), Kara has been presented to secondary school teachers, in order to propose it to their pupils and extend with them the set of case studies.

References

1. Chiara Bodei, Angela Giannetti, Maria Rita Laganà. La danza dei bit. In *Difficoltà di apprendimento, 14/1*, 2008 Erickson.
2. Chiara Bodei, Angela Giannetti, Roberto Grossi, and Maria Rita Laganà. Danzando coi bit. Da presentare allo Workshop GRIN@Kangourou: Informatica e Scuola, maggio 2010.
3. Chiara Bodei, Roberto Grossi, Maria Rita Laganà. La danza dei bit: un approccio multidisciplinare per l'apprendimento dell'informatica nella scuola primaria. *Atti del convegno DIDAMATICA 2009 (Informatica per la Didattica).*
4. Jasmine Del Vecchio, Michele Albano, Marco Righi, Maria Rita Laganà, Marco Maria Massai. Piccoli robot per riflettere insieme giocando. *Atti del convegno DIDAMATICA 2009 (Informatica per la Didattica).*
5. Michele Freschi. AI-Game: un ambiente didattico per la programmazione. Tesi di Laurea. Dipartimento di Informatica. Università di Pisa. [`http://etd.adm.unipi.it/theses/available/etd-06302009-112756/`]
6. Maria Rita Laganà, Marco Righi: Romeo and Juliet. An infrared search system. ROBOCOMM 2007: 38.
7. Kara. [`http://www.swisseduc.ch/compscience/karatojava/kara/`]

Ongoing Research in Wireless Sensor Networks

M. Bonuccelli, S. Chessa, S. Pelagatti

Dipartimento di Informatica, Università di Pisa,
Largo Pontecorvo 3, 56127 Pisa, Italy

1 Introduction

Wireless Sensor Networks (WSN) [1] are a recent technology suitable for unattended monitoring of a wide range of environments, spanning infrastructures such as factories or public buildings, as well as houses or even humans. In a WSN a set of low-power, inexpensive embedded devices (called sensors) spontaneously cooperate to construct a wireless network to support their monitoring activities. In such cases the sensors are active and typically comprise a processor, one or more sensing units (transducers), a radio transceiver and an embedded battery. A special sensor, called sink, acts as a gateway with the external networks, and it makes the sensed data available to external users. Prominent metrics for the evaluation of solutions for WSN are the energy consumption, the sensors' memory occupation, the code footprint, and the packet size.

In other applications the sensors are passive, they are not battery powered but they rely only on the energy induced by the electromagnetic waves emitted by a powered unit (the reader), and they have a limited storage functionality. Typically, passive sensors are implemented as RFID (Radio Frequency IDentification) tags.

Although in some applications active and passive sensors may coexist, in most cases their applicative scenarios are different as active sensors can self organize into a network and offer autonomous, unmanned, environmental monitoring service, while passive sensors do not construct an autonomous network.

2. Ongoing Research on WSN

In the early approaches, WSNs implemented an external storage scheme, for example using *Directed Diffusion* [2], where the user queries the network by injecting interests. Data matching an interest is then drawn along a path to the user, and, during this path, it can be aggregated and united with other data sensed by other sensors. This paradigm has evolved on different directions. One important evolution has been towards the use of query languages modeled on the SQL used in data bases to express more complex queries aimed at the specification of data acquisition, aggregation and filtering [3,4]. Another evolution has been towards the so-called *data centric storage* (DCS) models [5,6] where the sensed (and possibly aggregated) data is stored within the network in order to be retrieved later by the user.

In DCS it is assumed that the sensors deployed in the sensing area are aware of their position (for example they are equipped with GPS), and that each sensed datum *d* is

paired with a meta-datum k (also called key), denoted ($d:k$). DCS defines two primitives: put($d:k$) and get(k). The put($d:k$) primitive first computes the hash of the meta-datum k to obtain a pair of coordinates (x,y) within the sensing area, then it routes a packet containing the pair ($d:k$) towards (x,y) by means of the GPSR geographic routing protocol [7]. In general, GPSR identifies a set N_k of sensors forming the perimeter around (x,y), also called *home perimeter* (Fig. 1), and the put primitive requests each sensor in N_k to stores a copy of $d:k$. The data can be retrieved later by means of the get(k) primitive, which computes the hash of k to obtain the same pair of coordinates (x,y). Then, by means of GPSR, it sends a request for any data matching the key k towards (x,y). The request eventually reaches a sensor in N_k that, in turn, replies to the request by sending its stored data that match the request.
The seminal work in [5] adopts a simple model of the WSN in which the sensors are uniformly distributed in the WSN and do not attempt to balance the load of the sensors. In our work we are reconsidering this approach by analyzing scenarios in which the sensors are not uniformly distributed. Under this respect we have shown that this affects significantly the storage overhead balance, and that results in data loss. We have also proposed alternative methods to implement data redundancy provides gives guarantees on data availability, and we have considered the use of alternative data encodings based on erasure codes.
Future perspectives of this research are in the direction of using the techniques used for the management of non-uniformity in WSN to improve other protocols and results. Another development consists in the use of DCS mechanisms to construct new middleware for data management in WSN.

3. Ongoing Research on RFID

Radio Frequency IDentification (RFID) systems offer a promisingly affordable, cheap and flexible solution for object identification. An RFID system consists of radio frequency tags attached to the objects that need to be identified, and one (or more) networked electromagnetic reader. The reader is an entity with great computation power and memory, while tags have (very) limited computational resources. In these systems there is a single communication channel, and messages are exchanged between reader and tags, which are not able to communicate each other.
Typically, the reader broadcasts a message to the tags, which in case send back an answer. If many tags simultaneously answer, a collision occurs, i.e. their transmissions will merge in a meaningless message, and the reader is not able to identify any tag: it can only realize that more than one tag is communicating. Instead, if only one tag at a time transmits its ID, the reader can identify it. We proposed efficient protocols for identifying a set of tags in this setting [8,9].
In dense environments,, where there are many tags to be identified, usually multiple readers are employed. However, these readers may interfere each other, in two possible ways, known as reader-to-reader collision, and reader-to-tag collision. We are currently investigating these problems, looking for efficient protocols solving them.

References

1. Paolo Baronti, Prashant Pillai, Vince Chook, Stefano Chessa, Alberto Gotta, and Y. Fun Hu, “Wireless Sensor Networks: a Survey on the State of the Art and the 802.15.4 and ZigBee Standards”, Computer Communications, 30 (7): 1655-1695 (2007).
2. C. Intanagonwiwat, R. Govindan, and D. Estrin. Directed diffusion: a scalable and robust communication paradigm for sensor networks. In 6th annual ACM/IEEE international conference on mobile computing and networking, Boston, MA, USA, pages 56–67, 2000
3. S. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong. “TinyDB: an acquisitional query processing system for sensor networks”. ACM Trans. Database Syst., 30(1):122–173, 2005.
4. G. Amato, S. Chessa, C. Vairo. “MaD-WiSe: A Distributed Stream Management System for Wireless Sensor Networks”. To appear on Software - Practice and Experience (SPE).
5. Ratnasamy, S., Karp, B., Shenker, S., Estrin, D., Govindan, R., Yin, L., Yu, F., “Data-centric storage in sensornets with GHT, a geographic hash table. Mob. Netw. Appl. (MONET) 8(4) (2003) 427–442.
6. Michele Albano, Stefano Chessa, Francesco Nidito, and Susanna Pelagatti, “Q-NiGHT: Adding QoS to Data Centric Storage in Non-Uniform Sensor Networks”, IEEE Mobile Data Management (MDM ‘07), Mannheim, Germany, May 7 - 11, 2007, pp. 166-173
7. Karp, B., Kung, H.T.. “GPSR: Greedy Perimeter Stateless Routing forWireless Networks”. Proc. of MobiCom 2000, Boston, (2000) 243–254
8. M.A. Bonuccelli, F. Lonetti, F. Martelli "Instant Collision Resolution for Tag Identification in RFID Networks", Ad Hoc Networks (Elsevier), Vol. 5, n. 8 (November 2007), pp. 1220-1232.
9. M.A. Bonuccelli, F. Lonetti, F. Martelli: "Exploiting Signal Strength Detection and Collision cancellation for Tag Identification in RFID Systems", 14th IEEE Symposium on Computers and Communications (IEEE-ISCC2009), pp. 500-506, Sousse, Tunisia, July 5-8, 2009.

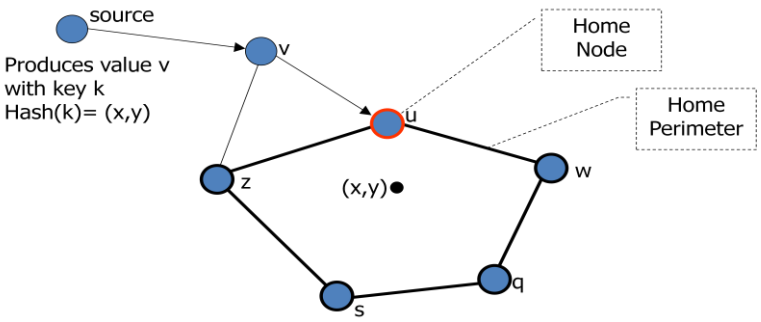


Fig. 1. The put protocol in Data Centric Storage

Formal methods for software integration: Achievements and challenges in a *discover-adapt-compose* journey

Antonio Brogi

Software integration continues to represent one of the major problems both for cross-enterprise and for intra-enterprise applications. The spread of component-based software development considerably contributed to reduce the problems of integrating heterogeneous software developed by different parties with different languages and/or platforms. The key idea of enabling component interactions via component-model dependent method signatures however introduced interoperability problems across different component platforms. The recently emerged service-oriented computing paradigm is simplifying further the reuse of existing applications by promoting the development of self-describing software services that can be published, discovered, and orchestrated by means of open standards over the network. However, while standardization solves the interoperability problems at the lower levels of the interaction stack (e.g., messaging), at the higher levels what have been standardised are the languages (e.g., WSDL and WS-BPEL) for defining business-level interfaces and protocols, not the specific interfaces or protocols. The result is that functionally equivalent services may present different interfaces and employ different interaction protocols. As a consequence, while service-oriented computing aims at reducing the cost and time needed to deploy loosely-coupled software compositions, various issues are still largely open, such as: How to discover the services that we *really* need? How to *engineer* the adaptation and the composition of services to match new business needs? How to guarantee that consumer-provider interactions will *really* work? How to *automate* service discovery, adaptation, and composition?

During the last years, our research activities have focused on exploiting formal methods to contribute setting a firm ground for the *discovery*, *adaptation*, and *composition* of software components and services. ***Service discovery*** is currently performed by querying UDDI registries that support only syntactic, keyword-based searches, and human intervention is required to choose among a set of candidate services. We have recently completed [1] an analysis of the potentially huge advantages of providing richer descriptions of services – e.g., including behaviour information and/or ontology annotations – in relation to the cost of generating such information and to the needed trade-off between expressiveness and cost and value of analysis. In this perspective, we have developed and prototyped various techniques [2,3,4] to enable beyond state-of-art ontology-aware and behaviour-aware discovery of services and of service compositions. Such techniques have also been extended and successfully implemented in the SMEPP middleware released at the end of the European project SMEPP ("Secure Middleware for Embedded Peer-to-Peer Systems", www.smepp.org). Software integration unavoidably leads to facing ***adaptation*** issues. Software components developed by different parties typically present different types of mismatches, ranging from signature to quality-of-service to behaviour. The need for adaptation is even more evident in the case of (Web) services, which are meant to support loosely-coupled interactions with generic consumers. Adaptation techniques are needed in a variety of situations such as to customise existing services to different types of consumers without re-implementing core application code of business application, to ensure backward compatibility of new versions of services, or to adapt legacy systems to meet new business demands. In this perspective, we have developed a full-fledged methodology [5,6] to support the automated adaptation of software components presenting behaviour mismatches among their interaction protocols. One of the most interesting features of the methodology – which has been implemented in the ITACA environment (itaca.gisum.uma.es) – is the ability to a priori

guarantee the correct interoperation of the adapted components. We have then also developed [7,8] a methodology for the automated generation of service adapters capable of solving behaviour mismatches among WS-BPEL processes. ***Service composition*** is the last – although clearly not least important - step of this *discover-adapt-compose* journey. Manual approaches to service composition are time-consuming and error-prone, and do not provide a priori guarantees on the properties of the obtained compositions. To contribute engineering the service composition process, we have developed and prototyped a methodology supporting workflow-based (adaptations and) compositions of services [9], we have analysed different languages for specifying service compositions [10], and we have developed a new compositional and decidable behaviour equivalence relation for services which can be fruitfully exploited to support sound service publication, enhanced discovery, and service replacement [11,12].

The above cited research activities have been developed in cooperation with various other researchers, some of them are listed as coauthors in the following list of selected publications. Space limitations do not allow us to properly discuss here the various open and challenging directions for future work. We just mention here some of them, on which we are open and willing to collaborate with other researchers interested in these themes. Such lines include for instance developing new adaptation techniques to cope with behaviour mismatches and with other dimensions of service adaptation (e.g., such as security and quality-of-service) as well as to address multi-layer adaptation issues, designing and experimenting light-weight service models for resource-bounded environments typical of peer-to-peer [13] and pervasive computing applications, and improving the performance and scalability of enhanced service discovery mechanisms.

Selected publications

1. A. Brogi. On the potential advantages of exploiting behavioural information for contract-based service discovery and composition. *Journal of Logic and Algebraic Programming*, 2010. (In press.)
2. A. Brogi, S. Corfini. Behaviour aware discovery of Web service compositions. *International Journal of Web Services Research*, 4(3): 1-25, 2007.
3. A. Brogi, S. Corfini, R. Popescu. Semantics-based Composition-oriented Discovery of Web Services. *ACM Transactions on Internet Technology*, 8(4): 1-39. 2008.
4. A. Brogi, S. Corfini. Ontology- and behaviour-aware discovery of Web service compositions. *Internat. Journal of Cooperative Information Systems*, 17(3):319-347, 2008.
5. A. Brogi, C. Canal, E. Pimentel. On the semantics of software adaptation. *Science of Computer Programming*, 61(2): 136-151, 2006.
6. A. Brogi, C. Canal, E. Pimentel. Component adaptation through flexible subservicing. *Science of Computer Programming*, 63(1):39-56, 2006.
7. A. Brogi and R. Popescu. Service Adaptation through Trace Inspection. *International Journal of Business Process Integration and Management*, 2(1): 1,9-16 2007.
8. A. Brogi, R. Popescu. Automated Generation of BPEL Adapters. In A. Dan and W. Lamersdorf, editors, *Proceedings of the 4th International Conference on Service Oriented Computing (ICSOC 06)*, LNCS vol. 4294, pages 27-39, 2006. (Best conference paper.)
9. A. Brogi, R. Popescu, M. Tanca. Design and Implementation of SATOR: a Web Service Aggregator. *ACM Transactions on Software Engineering and Methodology*, 19(3), 2010.
10. A. Roldan, E. Pimentel, A. Brogi. Software Composition with Linda. *Computer Languages Systems and Structures*, 36(4):395-405, 2009.
11. F. Bonchi, A. Brogi, S. Corfini, F. Gadducci. On the use of behavioural equivalences for Web services' development. *Fundamenta Informaticae*, 89(4): 479-510. 2008.
12. F. Bonchi, A. Brogi, S. Corfini, F. Gadducci. A net-based approach to Web services publication and replaceability. *Fundamenta Informaticae*, 94(3-4): 305-330. 2009.
13. A. Brogi and R. Popescu. Workflow-based semantics for peer-to-peer specifications. *Frontiers of Computer Science in China*, 2(4):398-412. 2008.

Negotiation, Commit, Execution: a three-phases approach to guaranteed dynamic assemblies

Roberto Bruni, Giorgio Ghelli, Ugo Montanari, Luca Pardini, and Matteo Sammartino,

joint work with Michele Bugliesi, Marzia Buscemi, Mariangiola Dezani-Ciancaglini, Anne Kersten, Davide Sangiorgi

Latest years are witnessing a shift of attention from the traditional concept of safe, overly controlled computation to open-ended, loosely coupled interactions, which permeate modern distributed systems, where separately developed, autonomous applications need to collaborate by dynamic assemblies.

Our department will coordinate the recently approved Italian MIUR (PRIN 2008) project "Interacting Processes in Open-ended Distributed Systems" (IPODS), in collaboration with the universities of Bologna, Torino and Venezia, aimed at developing robust mathematical foundations for highly dynamic collaborative scenarios, by extending and integrating static and dynamic analysis techniques. The abstract models for open and dynamic collaborative systems studied in IPODS are based on the key distinction of three different phases within an interaction among a set of so-called *participants* (i.e. the interacting, adaptive components):

1) Negotiation between different participants having different goals. Subsets of participants can be already involved in other ongoing interactions. The outcome of the negotiation can be either a failure and in this case no new interaction starts, or the stipulation of a *contract* that will bind participants' interaction, possibly joining existing ones.

2) Commit, where the participants accept to play the contract role formulated in the previous phase. In this phase it is important to check the conformance between the possible behaviors of the participants and the projections of the contract to their roles.

3) Execution of the interaction, according to the contract, but with the possibility of dynamic re-negotiations, e.g. in case of failures. Examples of failures are a participant abandoning the interaction or violating the contract. In absence of failures, the contract will bind participants to behave correctly.

The scheme given by the above phases is called NCE and it covers a wide range of situations like transaction processing (phases 1-2), type- and session-based interactions (phases 2-3), and proof-carrying code (phases 1 and 3).

An emblematic scenario is given by the service-oriented computing (SOC) paradigm, where service abstractions are published in public repositories to be discovered by and bound to other services: callers and callees are seen as participants. Different notions of contracts, sometimes called *choreographies*, have emerged to give suitable run-time guarantees by constraining the order in which

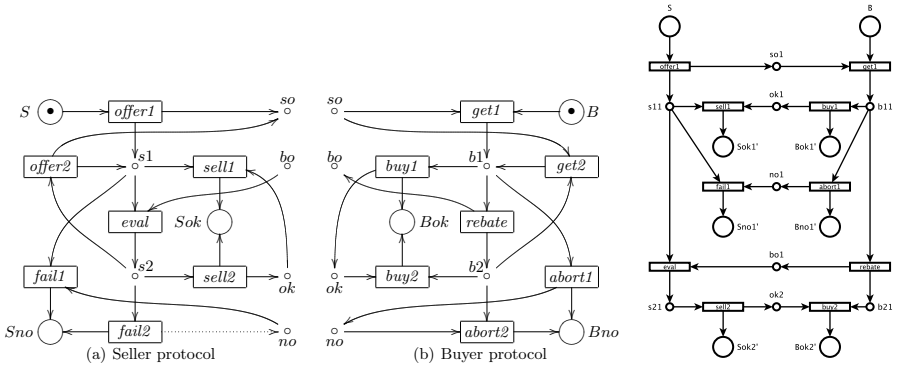


Fig. 1. Seller and Buyer example

interactions are carried out. Formally, this involves answering to two main questions: 1) Are the service abstractions compatible with the overall contract? 2) Is each participant consistent with its abstract description? Note that here we are often concerned with the operational behavior of services, not just with functional aspects. Moreover, the notions of compatibility and consistency should be such that whenever 1) and 2) are answered in the affirmative, then the overall interaction is guaranteed to be sound (e.g. communications are type-safe, no deadlock can arise).

Our research takes the theory of concurrency as a foundational ground, and focuses on observable aspects of interactions. In this regard, the IPODS approach is largely orthogonal to the typical game-theoretic approach, which mainly focuses on strategic aspects of interactions, for designing good individual or group strategies and for determining the system's stable configurations. On the contrary, IPODS focuses on the design of formal tools for system specification, and analysis/checking techniques and tools able to capture properties of the system that characterize the interactions from a strictly observational viewpoint. IPODS will exploit mathematical tools, models and results from concurrency theory (process calculi, Petri nets, graph rewriting, behavioral relations) and from type theory (session types, behavioral types, polymorphic types). For example, contracts can be described by means of types (session types are particularly suitable to this aim) and behaviors by terms of some process calculus, and the conformance check turns out to be some sort of dynamic type inference.

A concrete instance of the NCE scheme is shown in [4] in the context of safe multi-party interactions. Our formalization relies on Petri nets and it builds on classical results from concurrency theory and workflow management. In particular, it takes inspiration from Zero-Safe nets [5] and the unfolding construction [10]. Roughly, the procedure can be outlined as follows: each participant describes its admissible behaviour as a Zero-Safe net; the composition of all Zero-Safe nets is (partially) unfolded as a non-deterministic process; the unfolding is suitably pruned (e.g., by removing faulted situations) and presented as a

contract K (concluding phase 1 of NCE); if accepted by all participants, then their net tokens are tagged according to K (phase 2) and the firing of their transitions apply local consistency checks on such tags (phase 3). The main result establishes that any such execution cannot deadlock. We remark that the negotiated global contract K is a particular non-deterministic process that can be distilled automatically from the local protocol specifications of participants. This is better suited than, say, a deterministic process, because run-time choices are unavoidable in open transactions. It is also preferable to a soundness check over the conjunction of protocols, because the check would impose too strong compatibility requirements (deemed to fail in most cases) and would not guarantee proper termination.

For example, consider a simple two-party situation, with a seller and a buyer whose nets are in Fig. 1(a) and (b). Leaving aside all details related to the net formalism, intuitively their interaction begins with an offer from the seller. Both the buyer and the seller may accept the last offer proposed by the other, or make a different offer, or abort the negotiation. The transaction might not terminate if the strategy of both parties is to make offers repeatedly. Moreover, assume that the seller can abandon the transaction without informing the buyer (this is the case when the arc $(\mathit{fail2}, no)$ is actually missing). Then the interaction may lead to a state where the buyer is deadlock and the transaction cannot be completed. The stipulation of a contract may avoid divergence (e.g. by fixing a bound on the number of interactions) and may prune deadlock situations. An example of contract is the non-deterministic net-process in Fig. 1 (rightmost).

References

1. Bettini, L., Coppo, M., D'Antoni, L., Luca, M.D., Dezani-Ciancaglini, M., Yoshida, N.: Global progress in dynamically interleaved multiparty sessions. CONCUR'08. Vol. 5201 of LNCS, Springer (2008) 418–433
2. Bravetti, M., Zavattaro, G.: Contract-based discovery and composition of web services. SFM'09. Vol. 5569 of LNCS, Springer (2009) 261–295
3. Bruni, R., Mezzina, L.G.: Types and deadlock freedom in a calculus of services, sessions and pipelines. AMAST'08. Vol. 5140 of LNCS, Springer (2008) 100–115
4. R. Bruni and U. Montanari. Models for open transactions. FLACOS'09. Technical Report RR-385, Department of Informatics, University of Oslo (2009) 25–33.
5. Bruni, R., Montanari, U.: Zero-safe nets: Comparing the collective and individual token approaches. Inf. Comput. **156**(1-2) (2000) 46–89
6. Carbone, M., Honda, K., Yoshida, N.: Structured communication-centred programming for web services. ESOP'07. Vol. 4421 of LNCS, Springer (2007) 2–17
7. Castagna, G., Gesbert, N., Padovani, L.: A theory of contracts for web services. ACM Trans. Program. Lang. Syst. **31**(5) (2009)
8. Laneve, C., Padovani, L.: The pairing of contracts and session types. Concurrency, Graphs and Models. Vol. 5065 of LNCS, Springer (2008) 681–700
9. L. Mezzina. *Typing Services*. PhD thesis, IMT Alti Studi Lucca, 2009.
10. Nielsen, M., Plotkin, G.D., Winskel, G.: Petri nets, event structures and domains, part I. Theor. Comput. Sci. **13** (1981) 85–108
11. Vasconcelos, V.T.: Fundamentals of session types. SFM'09. Vol. 5569 of LNCS, Springer (2009) 158–186

User-friendly programming frameworks for parallel high-performance applications

Antonio Cisternino, Marco Danelutto, and Marco Vanneschi
joint work with Cristian Dittamo and Gabriele Cocco

Dipartimento di Informatica, University of Pisa

This presentation is focused on the feasibility of future user-oriented software tools and frameworks for parallel high-performance applications. In particular, we believe that a distinguishing feature of a user-oriented approach is that the parallel issues and paradigms of a high-performance application are fully compatible with, and indistinguishable from, standard and commercial tools and environments for sequential application development.

This topic is tackled jointly by two groups of the Department, working respectively on parallel programming models and methodologies and on advanced software technologies and infrastructures. Both competences and skills are considered absolutely necessary if we aim to take a significant step forward the much wider adoption of high-performance platforms by the IT community in accordance with the clear trends highlighted by both technology-push and technology-pull analyses [1].

As non-conventional architectures make their way into our devices, programming models appear increasingly more and more inadequate to exploit their capabilities. The presentation Parallel programming issues, achievements and trends in high-performance and adaptive computing [2] describes the state-of-the-art and trends in structured programming models and methodologies for parallel high-performance applications. The existing programming language infrastructure makes really hard to exploit these new abilities, in part because of the computational model behind their original design, and in part because the heterogeneous nature of this new scenario. Virtual machines have worsened, if possible, the situation by taking away from the programmer a direct control of the underlying hardware infrastructure. In principle the gap can be bridged by (automatic) tools that are responsible for compiling programs in a way that exploits the new computational resources. However, history taught us that compilers rarely manage to do magic, and the full intent of a programmer is not fully present in the source of the written program.

In [2] the features of a structured approach to parallel programming are discussed and some success histories of this approach are reported. Structured parallel programming tools and frameworks (e.g., ASSIST) are able to liberate the user from the knowledge of the underlying machine and allow him to utilize known and standard languages. However, these approaches require that the user has a deep conceptual and cultural knowledge of parallel paradigms (not of the hardware machine) from the point of view of their parallel semantics and cost model. One of the goal of the collaboration of the two research groups is to take a step forward, consisting in possible solutions that hides also this kind of

knowledge to the user, at least partially (of course, leaving the full knowledge of parallel paradigms and their implementation to the compiler and run-time support designer).

In this talk we present research activities aimed at finding ways that let the user expressing his will in terms of concurrent execution within a program, in a way that the underlying compilation infrastructure may benefit from these annotations. Both techniques based on usage of explicit parallel constructs and techniques more close to the software engineering concepts (based on AOP [4], intermediate code manipulation [3], etc.) will be discussed.

References

1. Asanovic, K., et al.: A view of the parallel computing landscape. Communications of the ACM, vol. 52, n. 10, October 2009.
2. Danelutto, M., and Vanneschi, M.: Parallel programming issues, achievements and trends in high-performance and adaptive computing. WiGoWin 2010.
3. Dittamo, C., Cisternino, A., and Danelutto, M.: Parallelization of C# Programs Through Annotations. Proc. of Practical Aspects of High-Level Parallel Programming Workshop (PAPP, co-located with ICCS 2007), pages 585–592, Springer Verlag, 2007
4. Aldinucci, M., Danelutto, M., Dazzi, P.: Muskel: an expandable skeleton environment. Scalable Computing: Practice and Experience, 8:4(325–341) 2007

Power-aware computing

Antonio Cisternino[1] and Paolo Ferragina[1]
joint work with Massimo Coppola[2] and Davide Morelli[1]

[1] Dipartimento di Informatica, University of Pisa, Italy
[2] ISTI "A.Faedo" National Research Council, Pisa, Italy

Power-aware computing is becoming popular as the effort for reducing energy consumption worldwide is growing. In fact computers affect the overall balance in a measurable way: studies indicate that PCs consumed in 1999 in the United States from 3

Significant power cuts have been made by hardware designers that affected mobile as well as traditional PCs: intelligent CPUs, Solid-State disks, and many other system and hardware tricks are contributing to achieve what Google's scientist Barroso called in 2007 [1] the need of "energy-proportional computin". A similar trend has been observed in software, though not on the same scale as hardware.

In order to circumvent those issues engineers are designing power-conscious mechanisms at system and hardware level but, as it occurred for performance in the last 40 years, it is well known that improvements achievable by means of proper structured computations can abundantly surpass any hardware improvements. And indeed, K.Kant of Intel recently [2] stated that"algorithmics may offer benefits that extend far beyond TCS into the design of systems". This is the reason why NSF promoted in April 2009 a Workshop titled "The Science of Power Management" in which scientists and engineers tried to set up a possible scenario for attacking in a principled way the problem of "power-saving computing" and posed many interesting questions ranging over all the software stack. Among the others we have been challenged by three issues: (1) the need for further scientific observations about the relation between power-consumption and computing, (2) the "design of good models" that explain this experimental observations and allow to understand the fundamental limitation and properties of energy vs performance issues; and (3) the development of a "set of canonical algorithmic techniques to make it easier for the new methods to filter into active use".

In this talk we present early results in these three issues, whose main goal will be to try to define a theory capable of relating computational complexity to power consumption: an experimental approach to computational complexity and a methodology for conducting measures which result independent of the underlying system running the algorithm/software to be tested. Early experimental results are presented and discussed, showing that our methodology is robust and can be used in many settings We also introduce the foundations of a theory for experimental algorithm complexity, which mimics what is predicted by the classic theory of computational complexity (big-O or Theta notations), except for some notable exceptions that we highlight and comment. This theory is val-

idated in many scenarios, by considering several architectures and algorithms. Because of the relation between time complexity and energy consumption, we may suggest that our work measures the "information work": namely, the energy required for performing information processing. These experimental figures will be the starting point for discussing some insights we have drawn about the design of power-aware algorithms, and about the introduction of a novel power-aware programming approach.

References

1. Barroso, A.: The Case for Energy Proportional Computing. IEEE Spectrum, December 2007
2. Kant, K.: Toward a Science of Power Management. IEEE Computer 42(9):99-101 2009.

This article was processed using the LaTeX macro package with LLNCS style

Collision avoidance using a wandering token in the PTP protocol

Augusto Ciuffoletti

Dept. of Computer Science – University of Pisa – Italy

Extended abstract

The Precision Time Protocol [1] (aka IEEE1588) is a clock synchronization protocol designed to allow sub micro-second clock accuracy using a packet network. It is another example of convergence of a problem usually solved using analog technologies towards packet networks.

The PTP protocol is in charge of computing data needed for the syntonization, intended as the tuning of clock frequency, and the synchronization, intended as the consensus about the clock value at a certain time, of the time reference internal to computing units. It is not in charge of interacting with the clock, but just of computing the data needed for the task of adjusting it.

The PTP is based on a regular flow of packets between a unit known as the *master* clock, and other units known as *slave* clocks, and is integrated by a master election algorithm, also based on packet exchange. To keep their clocks within the required accuracy bounds, the slaves need to send clock update requests (`Delay_Req` messages) to the master, and process the response (`Delay_Resp` messages).

The collision of requests coming from different slaves makes a problem for PTP, since serialization induces response jitter that degrades precision. The protocol provides a randomized solution that is adequate for networks of a limited size, but that does not scale. In large networks collisions may occur rather frequently

We introduce a solution to this problem that is based on a randomized token passing scheme: a token is passed from one slave to the other, according with a randomized scheme. The presence of the token enables the slave to synchronize with the master, thus introducing mutual exclusion. To make this solution of practical interest, we need to take into account a number of challenging issues: i) no additional traffic is introduced other than that strictly needed by PTP; ii) the PTP standard must not be broken; iii) each slave must be able to indicate another peer as the destination of the token; iv) each slave must be enabled to synchronize its clock every given time; v) the algorithm must not introduce single points of failure, other than those already introduced by PTP.

The protocol we propose meets these requirements in a system that we assume to be similar to an Ethernet, a unique collision domain where all packets are in sent broadcast. Environments that fulfill such requirement can be found in plant control applications, and in wireless ad-hoc networks.

In a nutshell, the token exchange protocol is embedded in the clock synchronization protocol: some space left *available for future use* inside the two packets of concern is used to convey the id of the next target. Therefore the execution of the protocol has no cost in terms of communication, and is compatible with the present version of the protocol.

Using the broadcast properties of the network each slave is able to build, simply observing the PTP traffic, a random neighborhood among which to select a target for the token. The neighborhood is updated dynamically, so that leave and joins are easely recorded. The way the random neighborhood is maintained is similar to that studied by [2]: the performance of the proposed solution is compared by simulation.

The master ensures that each slave receives the token with a timing that reflects the upper bound of the time between successive visits of the token, despite the random nature of the token circulation algorithm: before such upper bound is broken, the master *reroutes* a token to the starving slave.

In case of failure of the master, token circulation stops, but synchronization is equally unavailable. When the master recovers, the token circulation is immediately resumed: in case of a cold restart of the master, only a fraction of the units, those that are not present in any of the slave's neighborhoods, need to rejoin.

Future research As of today, the paper has been presented during the last IEEE1588 meeting (see [3]), a peer reviewed conference. However, there are extensions that might be addressed in future research.

One is targeting the extension to a specific application: the present version is agnostic with respect to the protocol supporting PTP, although a case study is shown for the Ethernet. However, the extension seems especially suited for wireless ad-hoc networks, as suggested by colleagues during the meeting referenced above.

Another is an analytic evaluation of the stochastic behavior, especially with respect to the frequency of master intervention to deroute packets, and for the evaluation of the efficacy of the shuffling rule for neighborhoods, especially when only a subset of the network is reachable by a generic slave.

References

1. Standard for a precision clock synchronization protocol for networked measurement and control systems. Technical Report IEEE Std 1588-2008, IEEE Instrumentation and Measurement Society, 3 Park Avenue, New York, NY 10016-5997, USA, July 2008.
2. Ziv Bar-Yossef, Roy Friedman, and Gabriel Kliot. RaWMS - random walk based lightweight membership service for wireless *ad-hoc* networks. *ACM Transactions on Computer Systems*, 26(2):66, June 2008.
3. Augusto Ciuffoletti. Collision avoidance for Delay_ Req messages in broadcast media. In *International IEEE Symposium on Precision Clock Synchronization for Measurement, Control and Communication*, 2009.

A framework for the verification of infinite-state graph transformation systems

Andrea Corradini

joint work with Paolo Baldan, Barbara König and Alberto Lluch Lafuente

Graph Transformation Systems (GTSs) [14] allow to describe in a rule-based way the evolution of systems whose states are represented as graphs, thus providing an explicit account of the logical or topological structure of the states. This can be relevant, for example, for object-oriented, concurrent, mobile, or distributed system, but it can also be exploited to specify algorithms acting on data structures with pointers.

The use of GTSs for the verification and analysis of such systems is still at an early stage, but there have been several proposals recently, either using existing model-checking tools [10, 15], or developing new techniques [12, 13], which are motivated by the fact that most GTSs of interest are infinite-state.

In this talk we will survey a technique for the analysis of infinite-state graph transformation systems, based on the construction of finite structures approximating their behaviour (see [2, 3, 6, 8, 9, 1, 7, 5, 4])

The *unfolding* of a GTSs, similarly to the more known unfolding of a Petri net, is a single partial-ordered structure representing all the possible computations of the system, including information about the causalities and the conflicts among the single transformation steps. Following a classical approach, one can construct a chain of finite under-approximations (k-truncations) of the unfolding of a GTS. More interestingly, also a chain of finite over-approximations (k-coverings) of the unfolding can be constructed: both chains converge (in a categorical sense) to the full unfolding.

The finite approximations of a GTS are Petri net-like structures, called *Petri graphs*, and they can be used to check properties of the system under consideration expressed in a suitable modal logic, which is a propositional μ-calculus where the propositions are formulæ of a monadic second-order logic over graphs. Once a finite approximation of the system is fixed, a formula belonging to a suitable fragment of this logic can be translated into a corresponding formula describing a property over the markings of the underlying net, and the verification of the latter formula implies the validity of the former.

Putting all this together, given a GTS $\mathcal{G}$ and a formula F in a suitable fragment of our logic, we can construct a Petri graph P which approximates $\mathcal{G}$. Then F can be translated into a Petri net formula $[F]$, such that if N_P is the Petri net underlying P, then $N_P \models [F]$ implies $\mathcal{G} \models F$. Therefore we reduce the verification of GTSs to the verification of Petri nets, for which we can exploit a rich theory [11] and efficient tools developed and tested along the years.

References

1. P. Baldan, A. Corradini, J. Esparza, T. Heindel, B. König, and V. Kozioura. Verifying Red-Black Trees. In D. Distefano, P. O'Hearn, and R. Iosif, editors, *Proceedings of the First international workshop on the verification of COncurrent Systems with dynaMIC Allocated Heaps (COSMICAH 2005)*, volume RR-05-04, pages 1–15. Queen Mary University, Dept. of Computer Science, 2005.
2. P. Baldan, A. Corradini, and B. König. A static analysis technique for graph transformation systems. In K. Larsen and M. Nielsen, editors, *Proceedings of CONCUR 2001*, volume 2154 of *Lecture Notes in Computer Science*, pages 381–395. Springer Verlag, 2001.
3. P. Baldan, A. Corradini, and B. König. Verifying finite-state graph grammars: an unfolding-based approach. In P. Gardner and N. Yoshida, editors, *Proceedings of CONCUR 2004*, volume 3170 of *Lecture Notes in Computer Science*, pages 83–98. Springer Verlag, 2004.
4. P. Baldan, A. Corradini, and B. König. A framework for the verification of infinite-state graph transformation systems. *Information and Computation*, 206:869–907, 2008.
5. P. Baldan, A. Corradini, and B. König. Unfolding graph transformation systems: Theory and applications to verification. In P. Degano, R. D. Nicola, and J. Meseguer, editors, *Concurrency, Graphs and Models*, volume 5065 of *Lecture Notes in Computer Science*, pages 16–36. Springer, 2008.
6. P. Baldan, A. Corradini, B. König, and B. König. Verifying a behavioural logic for graph transformation systems. In F. Honsell, M. Lenisa, and M. Miculan, editors, *Proceedings of the Workshop of the COMETA Project on Computational Metamodels*, volume 104 of *ENTCS*, pages 5–24. Elsevier/Forum, 2004.
7. P. Baldan, A. Corradini, B. König, and A. Lafuente. A temporal graph logic for verification of graph transformation systems. In *Recent Trends in Algebraic Development Techniques (WADT'06)*, volume 4409 of *Lecture Notes in Computer Science*, pages 1–20. Springer Verlag, 2007.
8. P. Baldan and B. König. Approximating the behaviour of graph transformation systems. In *Proc. of ICGT'02*, volume 2505 of *Lecture Notes in Computer Science*, pages 14–29. Springer, 2002.
9. P. Baldan, B. König, and B. König. A logic for analyzing abstractions of graph transformation systems. In *SAS'03*, volume 2694 of *Lecture Notes in Computer Science*, pages 255–272. Springer, 2003.
10. F. Dotti, L. Foss, L. Ribeiro, and O. Marchi Santos. Verification of distributed object-based systems. In *Proc. of FMOODS '03*, volume 2884 of *Lecture Notes in Computer Science*, pages 261–275. Springer, 2003.
11. J. Esparza and M. Nielsen. Decidability issues for Petri nets - a survey. *Journal of Information Processing and Cybernetic*, 30(3):143–160, 1994.
12. A. Rensink. Towards model checking graph grammars. In *Proc. of the* 3rd *Workshop on Automated Verification of Critical Systems*, Technical Report DSSE–TR–2003–2, pages 150–160. University of Southampton, 2003.
13. A. Rensink. Canonical graph shapes. In *Proc. of ESOP '04*, volume 2986 of *Lecture Notes in Computer Science*, pages 401–415. Springer, 2004.
14. G. Rozenberg, editor. *Handbook of Graph Grammars and Computing by Graph Transformation: Foundations*, volume 1. World Scientific, 1997.
15. D. Varró. Automated formal verification of visual modeling languages by model checking. *Software and System Modeling*, 3(2):85–113, 2004.

Reduction Systems: synthesis, refinement and verification of behavioural models

Fabio Gadducci
(with Filippo Bonchi, Andrea Corradini, Giacoma Monreale, Ugo Montanari)

The motivations inspiring the recently approved PRIN 2008 Project *Reduction Systems: synthesis, refinement and verification of behavioural models* (RedS) stem from recent advances in the theory of concurrent systems.

Reduction systems (RSs) offer a flexible methodology for specifying computational systems, defined by a set of states, which abstract some of their features, and a family of rules, which allow for the description of the evolutions of a system as a relation between states. Basically, a computation step amounts to an application of a rule to a state, through an adequate mechanism calculating its successor. When the state has a structure, the reduction relation can be calculated by means of an inductive closure of the rules under a family of "active contexts"; such a solution is the basis of many formalisms for specifying computational systems, from lambda-calculus to process algebras, where a term (a tree) represents the topology of links between subsystems and the reduction rules specify how the components can interact in order to produce new configurations.

A limitation of RSs is that the behaviour of a state is not described by those of its components, thus hindering the modularity and the chance of testing the system. Hence, in order to analyze a system, one cannot use the RS description, and should adopt, instead, a behavioural semantics capable of making explicit the interesting computational aspects such as the interaction between the components or their concurrency features. A sample archetype of computational models is given by labeled transition systems (LTSs), where the transitions of the various components are equipped with labels describing their interactions with the environment they live in. However, the definition of suitable behavioural semantics is often a difficult operation, prone to interminable revisions, due to the introduction of new behavioural concepts, of new variants of reduction semantics and of behavioural semantics, and of logics and other reasoning principles. For instance, to define LTSs characterizing bisimulations which are in turn congruences (i.e, such that state equivalence is closed with respect to context composition), thus allowing the compositional analysis of a system, is often very difficult.

Several categorical conditions were proposed in order to ensure the compositionality of the observational semantics, but these conditions are usually prescriptive: they provide no hint for the definition of the "adequate" LTS. An important result was achieved by Leifer and Milner [6] (and generalized by Sassone and Sobocinski [8]): given a RS, an LTS can be defined by means of the construction known as "relative pushout" (RPO), such that bisimilarity is a congruence. Such approach can be applied to a large class of RSs, as those defined by rewriting systems in adhesive categories [5], which provide an elegant characterization of the categories where many of the relevant properties about graph rewriting, such as the local confluence ("Church-Rosser") theorem, hold.

Starting from the above motivations, the goal of the project is to devise a general framework for the synthesis and the improvement of behavioural semantics and for their application to the development of verification and reasoning methods on computational systems. Thus, the project is not focused on the study of a concrete model, but rather on the definition of abstract and general methodologies for the construction of behavioural semantics, the improvement of such models and the definition of logics and verification techniques.

More precisely, this project will act towards three directions. First of all it aims at a deeper insight into the methodologies for the systematic derivation of behavioural semantics from reduction semantics, following the lines of the previous PRIN 2005 Project "ART". In particular, it aims at a rigorous study of the categories where the RPO construction can be easily carried out; a relevant role will be played by (hyper- and term) graphs [4] and bigraphs [7] and, obviously, by the rewriting systems on adhesive categories (ARSs).

It should be noted that the behavioural semantics thus distilled often are not satisfactory. Quite simply, they might present some undesired property, such as being of an unwieldy size. Hence, the project will study refinement methodologies of behavioural models, in order to better support verification and reasoning techniques. Among others, it will study the possibility of inductively characterizing the RPO-derived LTSs, for example in the so-called SOS style, according to the algebraic structure of states. The LTSs generated from by means of the RPO construction often have other problematic aspects: they do not correspond to the desired behavioural equivalences and they are infinite branching. We will investigate such problems with an insight into the theory of "semi-saturated" bisimulations [3, 2] and by developing pruning techniques in order to reduce the complexity of synthesized LTSs, while preserving the semantics of the system [1].

Finally, the project aims at defining abstract verification techniques subsuming specific proposals as instances. It will study "up-to" coinduction principles and Hennessy-Milner style logics for the characterization of weak bisimulations, verifying their applicability to LTSs yielded by RPOs. The use of observations accounting for the state structure suggests the development of suitable spatial logics, where the logic operators can explore and treat also the systems topology. This seems to be particularly indicated for the bi- and multi-graphical models and for the symbolic models for open systems (i.e., with unspecified components).

References

1. F. Bonchi, F. Gadducci, and B. König. Synthesising ccs bisimulation using graph rewriting. *Information and Computation*, 207(1):14–40, 2009.
2. F. Bonchi, F. Gadducci, and G. V. Monreale. Reactive systems, barbed semantics, and the mobile ambients. In L. de Alfaro, editor, *Foundations of Software Science and Computational Structures*, volume 5504 of *Lect. Notes in Comp. Sci.*, pages 272–287. Springer, 2009.
3. F. Bonchi, B. König, and U. Montanari. Saturated semantics for reactive systems. In *Logic in Computer Science*, pages 69–80. IEEE Computer Society, 2006.

4. Andrea Corradini and Fabio Gadducci. On term graphs as an adhesive category. In Maribel Fernández, editor, *Term Graph Rewriting*, volume 127(5) of *Electr. Notes in Theor. Comp. Sci.*, pages 43–56. Elsevier, 2005.
5. S. Lack and P. Sobociński. Adhesive and quasiadhesive categories. *Theoretical Informatics and Applications*, 39:511–545, 2005.
6. J.J. Leifer and R. Milner. Deriving bisimulation congruences for reactive systems. In C. Palamidessi, editor, *Concurrency Theory*, volume 1877 of *Lect. Notes in Comp. Sci.*, pages 243–258. Springer, 2000.
7. R. Milner. Pure bigraphs: Structure and dynamics. *Information and Computation*, 204(1):60–122, 2006.
8. V. Sassone and P. Sobociński. Deriving bisimulation congruences using 2-categories. *Nordic Journal of Computing*, 10:163–183, 2003.

It is Time to add Time!

Gianna M. Del Corso and Francesco Romani

Dipartimento di Informatica

It is everyone experience that the accessibility to internet has deeply changed with search engines. Search does not only concern web pages: engines for searching news, images or bibliographic references have become very useful to users. Almost fifteen years ago, when the first link-based search techniques were proposed, the ranking algorithms were based on spectral techniques, and the web was seen as a static set of pages. Search engines employing algorithms such as PageRank or Hits recompute the rank of the web pages reasonably often to allow the user to get up-to-date results. These pre-computed static measures do not reflect the timeliness or temporal evolution of pages and links. Even if time-aware ranking scheme have been proposed in the literature [1], it is only recently that we have seen attempt to incorporate time and freshness of web content into search engines (see Ask.com, OneRiot.com). Of course, in other scenario, where time awareness is more important, as for example, in news engines, the ranking is mostly based on the freshness of a story rather than authoritativeness of the news source.

In the last few years part of our research has focused on the application of numerical techniques to improve ranking algorithms, and more recently we investigated the possibility to add time into the ranking models. We first started analyzing PageRank [2, 3], proposing viable and alternative ways to compute the PageRank vector; then we addressed mathematically the problem of ranking news stories [4]. The proposed algorithm rates news information, seeking out the most authoritative sources and identifying the most current events in the particular category to which the news article belongs. The mathematical formulation of the ranking algorithm takes into account time, modeling the process as a stream of information whose importance changes over time. Since we have that generally important stories are posted many times by different press agencies, our algorithm incorporates an online process which measures the similarity of news articles using the cosine of the angle of two stories in the vector space, so that the (weighted) size of the cluster formed around a piece of news can be deemed a measure of its importance.

A related problem is the ranking of scientific publications [5–7], where the two aspects of citations (representing links between papers) and time of publications must to be combined to get a reasonable measure [8]. This is done introducing the concept of *aging of citations*, so that a paper highly cited in the past, but no longer cited in the present time, is penalized with respect to a recent paper that has received a lower number of citations but is highly cited at the present time.

We are currently applying similar ideas to Twitter in order to propose a ranking algorithm for tweets and users based on the link structure between

users-tweets-friends, and assuming a decadence over the time of the importance of tweets. This problem has differences with the ranking of news, since the link structure is more complex, but has also similarities since the ranking should take into account the freshness of tweets as well as the authoritativeness of the source. The friendship net is another aspect that should be taken into account to propose a satisfactory ranking scheme.

References

1. Berberich, K., Vazirgiannis, M., Weikum, G.: T-rank: Time-aware authority ranking. In Springer, ed.: Algorithms and Models for the Web-Graph. Volume 3243/2004 of Lecture Notes in Computer Science. (2004)
2. Del Corso, G.M., Gullí, A., Romani, F.: Fast pagerank computation via a sparse linear system. Internet Mathematics **3**(2) (2005) 251–273
3. Del Corso, G.M., Gullí, A., Romani, F.: Comparison of krylov subspace methods on the pagerank problem. Journal of Computational and Applied Mathematics **210**(1-2) (2007) 159–166
4. Del Corso, G.M., Gullí, A., Romani, F.: Ranking a stream of news. In: Proceedings of the 14th international conference on World Wide Web, ACM (2005) 97–106
5. Bini, D.A., Del Corso, G.M., Romani, F.: Evaluating scientific products by means of citation-based models: a first analysis and validatiuon. Electron. Trans. Numer. Anal. **22** (2008) 1–16
6. Bini, D.A., Del Corso, G.M., Romani, F.: A combined approach for evaluating papers, authors and scientific journals. Journal of Computational and Applied Mathematics (2010) to appear doi:10.1016/j.cam.2010.02.003.
7. Del Corso, G.M., Romani, F.: Versatile weighting strategies for a citation-based research evaluation model. Bullettin of the Belgian Math. Soc.-Sikmon Stevin **16**(4) (2009) 723–743
8. Del Corso, G.M., Romani, F.: A time-aware citation-based model for evaluating scientific products. In: SMCTools. (2009)

ESC: A Semantic-based Middleware for Service Oriented Computing

Gianluigi Ferrari[1], Roberto Guanciale[1],
Daniele Strollo[1], Emilio Tuosto[2]

[1] Dipartimento di Informatica,
Università degli Studi di Pisa, Italy
[2] University of Leicester, Computer Science Department
University Road, LE17RH, Leicester, UK

Service Oriented Computing envisages systems as combination of basic computational entities, called services. The main methodologies for composing services are *orchestration* and *choreography*. Services are orchestrated when their execution is described through an "external" process, the *orchestrator*. A *choreography*, instead, is a design that yields the architecture by specifying how services should be connected and interact to accomplish their tasks.

Within the SENSORIA project we designed and developed a programming middleware, called ESC (Event-based Service Coordination) for Service-Oriented applications. This research aims at putting real-world programming techniques on solid foundations to enable the construction of service oriented applications that are better understood and more robust. Indeed, we showed how foundational approaches could be fruitfully exploited and integrated into a lightweight platform for designing, deploying, and executing service-oriented applications. The ESC framework goes all the way from a foundational process calculus, the Signal Calculus, and its choreography model, Network Coordination Policy, over a Java middleware, JSCL. It supports a model driven development methodology devoted to manage Long Running Transactions.

The Signal Calculus [9,7] (*SC*) is an asynchronous process calculus acting as the conceptual counterpart of the middleware. Differently from other approaches, *SC* relies on the event notification paradigm: service interactions are specified by means of multicast notifications. Remarkably, the calculus does not assume any centralized mechanism for publishing, subscribing and notifying events. Being the formal counterpart of a programming middleware, *SC* just supplies a set of basic primitive constructs: higher level constructs can be automatically compiled over the basic primitives. The Network Coordination Policies [5,4] (*NCP*) equips the framework with a choreography model. Choreographies take the form of processes detailing the behavior as observed from a *global* point of view, namely by observing all the public interactions taking place on the network infrastructure. *SC* and *NCP* lay at two different levels of abstraction: the *NCP* specification declares *what* is expected from the service network infrastructure, the *SC* design specifies *how* to implement it. It has been proved that for each SC design, there exists an NCP choreography that reflects all the properties of the design [13,4]

At implementation level we find the Java Signal Core Layer (JSCL). JSCL is a set of Java API for programming distributed components interacting by notifying

multicast events. JSCL is the run-time support for *SC* networks of services. The ESC platform also provides a user-friendly interface supplying a graphical and a textual representation of *SC* networks. The graphical representation manages the choreography by considering the components and their interconnections, without detailing their internal logics. The textual notation, instead, allow programming service operation logic. One can easily pass from the graphical to the textual representation and to automatically generate the runnable JSCL code as well. The ESC platform supplies a set of model transformation tools that, starting from the high level specifications, automatically build the runnable code. The usefulness of the ESC framework has be illustrated by tackling the problem of designing, implementing and refining long running transactions [14,2,11,6]

The distinguished feature of our proposal is that any language and technique involved in the ESC middleware have formal foundation providing a clean implementation strategy, For instance, the formal foundations of long running transactions has driven the design and implementatiom of the corresponding run-time support [2].

One of our immediate concerns involves finding and developing techniques to strengthen the ESC platform. One possible improvement is the design of the debugging facilities. The loosely coupled nature of services complicated the activity of fixing programming errors. Indeed, SOA applications often require long-lasting computations that make practically impossible to reproduce erroneous executions. Also, traditional debugging may require high costs because service invocations may hide complex and expensive tasks. Finally, the computational burden on a complex service may be delegated to virtual machines (e.g. Amazon Elastic Cloud Service EC2). Virtual computing comes at a cost, therefore debugging such services could possible require repeated tests involving virtual machines. For instance, Amazon EC2 does not provide any debugging facility therefore each test involving a service invocation would be charged a cost. Some initial results in this direction can be found in [10,1]. In order to obtain a broader application of our platform, it is necessary to develop fully-automated methods for checking properties. In [3] we introduced a finite state technique to verify properties of Long Running Transactions. Scalability is still the main obstacle against the wide spread application of finite state verification techniques. We are exploring two research directions. From one hand, we plan to take advantage from software model ckecking techniques and related toolkits. From the other hand, we build upon our previous work where we developed efficient finite state verification techniques for mobile systems, based on a suitable class of automata called history-dependent (HD) automata [8,12].

References

1. C. Bodei and G. L. Ferrari. Choreography rehearsal. In *Web Services and Formal Methods, 6th International Workshop, WS-FM 2009*, volume to appear of *Lecture Notes in Computer Science*. Springer, 2010.
2. R. Bruni, G. L. Ferrari, H. C. Melgratti, U. Montanari, D. Strollo, and E. Tuosto. From theory to practice in transactional composition of web services. In

EPEW/WS-FM, volume 3670 of *Lecture Notes in Computer Science*, pages 272–286. Springer, 2005.
3. V. Ciancia, G. L. Ferrari, R. Guanciale, and D. Strollo. Checking correctness of transactional behaviors. In *Formal Techniques for Networked and Distributed Systems - FORTE 2008*, volume 5048 of *Lecture Notes in Computer Science*, pages 134–148. Springer, 2008.
4. V. Ciancia, G. L. Ferrari, R. Guanciale, and D. Strollo. Event based choregraphy. *Science of Computer Programming*, to appear, 2010.
5. V. Ciancia, G. L. Ferrari, R. Guanciale, and D. Strollo. Global coordination policies for services. *Electr. Notes Theor. Comput. Sci.*, 260:73–89, 2010.
6. V. Ciancia, G. L. Ferrari, R. Guanciale, D. Strollo, and E. Tuosto. Model-driven development of long running transactions. In *Rigorous Software Engineering for Service Oriented Systems - Results of the SENSORIA Project on Software Engineering for Service Oriented Computing (to appear)*, Lecture Notes in Computer Science. Springer, 2010.
7. G. Ferrari, R. Guanciale, D. Strollo, and E. Tuosto. Coordination via types in an event-based framework. In *Formal Techniques for Networked and Distributed Systems - FORTE 2007*, volume 4574 of *Lecture Notes in Computer Science*, pages 66–80. Springer, 2007.
8. G. L. Ferrari, S. Gnesi, U. Montanari, and M. Pistore. A model-checking verification environment for mobile processes. *ACM Trans. Softw. Eng. Methodol.*, 12(4), 2003.
9. G. L. Ferrari, R. Guanciale, and D. Strollo. Jscl: A middleware for service coordination. In *Formal Techniques for Networked and Distributed Systems - FORTE 2006*, volume 4229 of *Lecture Notes in Computer Science*, pages 46–60. Springer, 2006.
10. G. L. Ferrari, R. Guanciale, D. Strollo, and E. Tuosto. Debugging distributed systems with causal nets. *ECEASST*, 14, 2008.
11. G. L. Ferrari, R. Guanciale, D. Strollo, and E. Tuosto. Refactoring long running transactions. In *Web Services and Formal Methods, 5th International Workshop, WS-FM 2008*, volume 5387 of *Lecture Notes in Computer Science*, pages 127–142. Springer, 2009.
12. G. L. Ferrari, U. Montanari, and E. Tuosto. Coalgebraic minimization of hd-automata for the pi-calculus using polymorphic types. *Theor. Comput. Sci.*, 331(2-3), 2005.
13. R. Guanciale. *The Signal Calculus: Beyond Message-based Coordination for Service*. PhD thesis, PhD Thesis, Institute for Advanced Studies, IMT, Lucca, 2009.
14. D. Strollo. *Designing and Experimenting Coordination Primitives for Service Oriented Computing*. PhD thesis, IMT Institute for Advanced Studies, Lucca, 2009.

What May Be Next In Mathematical Modeling

Antonio Frangioni[1] and Luis Perez Sanchez[1]

Dipartimento di Informatica, Università di Pisa

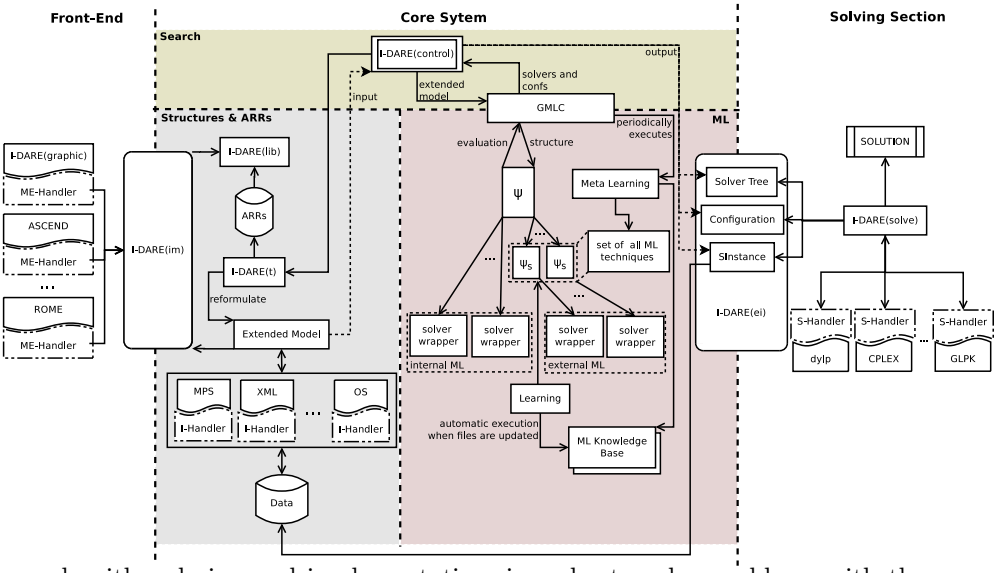
Abstract

We describe some of the common choke points which make applications of mathematical models less efficient than they could be. Based on this discussion we present our rather bold idea for a system aimed at easing those choke points, possibly making the development, deployment and improvement of systems using complex mathematical models more efficient. The idea is to provide tools which automate the discovery and re-use of information which is nominally "available", but in fact lost in the uncharted backwaters of the scientific literature and/or in prototypical solution algorithms which, despite holding great promise, are too specialized to be known and used outside small circles of hyper-specialists of the field. By doing so in an appropriate way, a positive feedback loop could be established where good performances of the system lure in more users and developers, which in turn provide it more data that further improve its performances. This may conceivably have a profound effect on the economics of development and deployment of software systems based on advanced mathematical models, as well as on more academic activities like testing new ideas, in some way like efficient web search completely revolutionized the use of Internet. While we could very well be getting wrong any and all of the details of the approach, and even the big picture, we believe that this is interesting in the "What's Going on and What's Next" context: by integrating many different aspects of Computer Science (algorithm design and implementation, logic programming and automatic deduction, machine learning, search in large structured spaces, performance evaluation of sequential and parallel programs, software development tools, graphical user interfaces, network science and social intelligence techniques ...) together into a multi-pronged attack one may conceivably produce a system with the potential of substantially improving the scientific and technological productivity worldwide.

The I-DARE System

Complex, hierarchical, multi-scale industrial and natural systems generate increasingly large mathematical models. Practitioners are usually able to formulate such models in their "natural" form; however, solving them often requires finding an appropriate *reformulation* to reveal *structures* in the model which make it possible to apply efficient, specialized approaches. The search for formulation of a given problem which allows the application of the solution algorithm that best exploits the available computational resources is a painstaking process which requires experts in solution algorithms for figuring out which algorithm is better used, considering the appropriate selection of the several obscure algorithmic parameters that each solution methods has.

We aim at improving the effectiveness and efficiency of this search in the (formulation, solver, configuration) space by developing a software system capable of automatizing its main steps. A possible sketch of the envisioned system is shown below, and a more detailed discussion of some of its main components can be found in [1, 2]. What is relevant here is that the system requires significant advances in, and *integration* among, several different areas of Computer Science:



- algorithm design and implementation, in order to solve problems with the countless many different *structures* that are required for applications;
- logic programming and automatic deduction, in order to allow the definition and efficient transversal of the space of (re)formulations;
- machine learning and performance evaluation of programs, in order to allow a-priori selection of the most appropriate solution algorithm, algorithmic parameters and available computational architecture for each formulation;
- techniques for efficient search in these large structured spaces, including population-based and biologically-inspired algorithms;
- software development tools and graphical user interfaces to allow rapid prototyping, development and deployment of actual applications of the system;
- parallel/grid/cloud programming techniques to allow for both quick responsiveness and solution of computationally intensive problems;
- network science and social intelligence techniques to establish, maintain and strengthen positive feedback loops in the system.

References

1. A. Frangioni, L. Perez Sanchez "Searching the Best (Formulation, Solver, Configuration) for Structured Problems" *Technical Report* **09-18**, Dipartimento di Informatica, Università di Pisa, 2009
2. A. Frangioni, L. Perez Sanchez "Artificial Intelligence Techniques for Automatic Reformulation of Complex Problems: the I-DARE Project" *Technical Report* **09-13**, Dipartimento di Informatica, Università di Pisa, 2009

Query Languages for Graph Databases

Giorgio Ghelli and Luca Pardini

Università di Pisa, Dipartimento di Informatica

1 Graph Databases

Graphs are mathematical objects used to represent collections equipped with a binary relation. Graphs arise in many application fields, such as social networks, transport networks, natural science. Graphs have trees and forests as important, and ubiquitous, special cases. Many typical graphs operations have been described and studied, such as connectivity analysis, analysis of distance, construction of spanning trees.

Graph databases are databases that represent such collections, possibly combined with data that has a non-graphical nature. In traditional database jargon, a graph is just a collection with a binary association; hence, any DBMS is able to manage a graph database. However, typical graph operations escape the expressive power of database languages, and in particular the expressive power of relational algebra or of first order logic.

Graph databases have been studied in the nineties [3, 6, 7, 5, 1, 2], in the early period of research on semi-structured databases, but the interest faded with the rise of XML and the move from graph-based to tree-based data models for semi-structured data. New interest in graph databases is now rising, partly because the old problems have not been solved yet, and partly because of new interest in the RDF data model [8, 4].

RDF (Resource Description Framework) is a formalism defined by the W3C to describe resources, and entities of any nature, through sets of triples 'subject-predicate-object'. RDF allows entities to be directly described at the semantic level, has happens with the Object-Oriented data model, with no need to translate into a more rigid structure, as happens with the relational data model. With respect to the Object-Oriented data model, RDF is more flexible, since data can be described without a schema. Moreover, entities are referenced through explicit URIs, while the object-oriented data model uses opaque 'object references', which only make sense inside a closed scope. Hence, the former reference schema is much better suited to the World Wide Web. W3C defined a query language for RDF, SPARQL [9], but this language is deeply relational, hence its expressive power is quite limited.

2 The Open Problems

Query languages for graph databases present a wide set of open problems. One should first understand whether graph data can be queried using exact languages

or whether they rather need languages with a ranking semantics. Exact languages can be used to solve questions like 'is X connected to Y?', but ranking query languages seem better suited to questions like: 'is there a shop of this kind which is near-enough to my current location?' Another basic choice is between languages that are designed for data with a schema and languages for semi-structured data.

Even assuming a conservative choice, based on traditional exact database algebras enriched with recursion, the problem of query processing is wide open. Graph algorithms have been deeply studied, and the optimization of relational algebra has been studied as well, but no satisfactory approach is known for the efficient execution of queries that combine the two aspects.

View update is another interesting problem in this context. View update is defined as the problem of transfering an update that is defined on the result of a query back to the data source. This classical problem has been recently faced through a new approach, called 'bidirectional programming', where the query is defined through a rich language that, for any query operation, specifies enough information to define the inverse direction. In this field, we are currently studying the extension of this approach to graph query languages.

References

1. A graph query language and its query processing. In *ICDE '99: Proceedings of the 15th International Conference on Data Engineering*, page 572, Washington, DC, USA, 1999. IEEE Computer Society.
2. Renzo Angles and Claudio Gutierrez. Survey of graph database models. *ACM Comput. Surv.*, 40(1):1–39, 2008.
3. Marc Gyssens, Jan Paredaens, and Dirk van Gucht. A graph-oriented object database model. In *PODS '90: Proceedings of the ninth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 417–424, New York, NY, USA, 1990. ACM.
4. G. Klyne and J. Carroll. Resource description framework (rdf) concepts and abstract syntax. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/, 2004.
5. J. Paredaens, P. Peelman, and L. Tanca. Merging graph-based and rule-based computation: The language g-log. *Data & Knowledge Engineering*, 25(3):267 – 300, 1998.
6. Jan Paredaens, Peter Peelman, and Letizia Tanca. G-log: A graph-based query language. *IEEE Transactions on Knowledge and Data Engineering*, 7:436–453, 1995.
7. Alexandra Poulovassilis and Mark Levene. A nested-graph model for the representation and manipulation of complex objects. *ACM Trans. Inf. Syst.*, 12(1):35–68, 1994.
8. W3C. Resource description framework. http://www.w3.org/RDF/.
9. W3C. Sparql query language 1.1. http://www.w3.org/TR/sparql11-query/.

The Rotation Distance of Binary Trees

Fabrizio Luccio and Linda Pagli

Dipartimento di Informatica, Università di Pisa

The *rotation distance* $d(S,T)$ between two binary trees S, T of n vertices is the minimum number of rotations to transform S into T (Culik and Wood, 1982). While it is known that $d(S,T) \leq 2n-6$, a conjecture based on hyperbolic geometry states that there are trees for which this bound is sharp for any value of $n \geq 11$ (Sleator, Tarjan, and Thurston,1988).

A vast literature on this subject has appeared henceforth, mainly aimed at studying good heuristics for transforming a tree into another, or determining particular families of trees for which upper and lower bounds can be more easily found. In fact it is not yet known if the problem of determining the rotation distance is NP-hard, or if the lower bound conjecture holds true for all n.

A satisfactory settlement of the above open problems is not likely to affect the future of mankind, still it may be nice to look into this matter occasionally. We revise the major findings published so far and look into possible improvements.

References

K. Culik and D. Wood,A note on some tree similarity measures,*Information Processing Letters* 15 (1982) 39-42.

P. Dehornoy, On the rotation distance between binary trees, (2009) http://arxiv.org/abs/0901.2557

E. Del Tessandoro, The upper bound on the rotation distance of binary tree: an experimental study, *Dipartimento di Informatica, University of Pisa*. Report on a Project for B.S. Degree in Informatics (2010).

F. Luccio and L. Pagli, On the upper bound on the rotation distance of binary trees, *Information Processing Letters* 31 (2) (1989) 57-60.

F. Luccio, A. Mesa Enriquez, and L. Pagli, Lower bounds on the rotation distance of binary trees, *Dipartimento di Informatica* Tech. Rep. TR-10-01, (2010).

J. Pallo, An efficient upper bound of the rotation distance of binary trees, *Information Processing Letters* 73 (3-4) (2000) 87-92.

D.D. Sleator, R.E. Tarjan, W.R. Thurston, Rotation distance, triangulations, and hyperbolic geometry, *J. Amer. Math. Soc.* 1 (3) (1988) 647-681.

Cheminformatics: emerging challenges in an interdisciplinary computational discipline

Alessio Micheli

The role of Computer Science (CS) in the frontier research of applied sciences is increasing. A specific emphasis is on the introduction of innovative methodologies to deal with complex systems in Natural and Life Sciences.

The term Cheminformatics is emerging to delineate the integration of studies in CS and Chemistry, with a further interdisciplinary nature crossing also the boundaries between Biochemistry, Pharmacology, Computational Toxicology, and Medicinal Chemistry fields of research.

In this talk we will introduce the Cheminformatics field, with a specific focus on the Machine Learning (ML[1]) approach. Cheminformatics is undergoing into rapid changing, involving model advancements and rigorosity and data issues, that will be briefly discussed. Through such topic we intend to look at paradigmatic instances of challenging problems belonging to the above-mentioned grand vision of the CS in the sciences.

Indeed, Cheminformatics offers open challenges of general interest for science and society, due to the potential impact for drug design, biomaterial design (including tissue engineering), environmental and health studies (including green chemistry and toxicology), up to possible new perspectives in the understanding of life processes. The relevance of the specific Cheminformatics perspective (complementary to Bionformatics) is understandable in the context of the increasing interest for the fundamental role played by *small molecules* in the living, and of the pressure to reduce costs and accelerate drug and new (bio)materials discovery cycles (including the need to reduce or ultimately to replace the use of test on animals) through the introduction of innovative computational approaches. To this aims two fundamental problems could be considered:

- the need of a systematic exploration of the "chemical space"[2]: the exploration of chemical space has so far been extremely limited, considering the enormous gap between known compounds (on the order of 10^7) and unknown small organic molecules that could be created (estimated to exceed 10^{60});
- the need of innovative and reliable methodologies for the prediction of molecular properties: this would allow the evaluation of the behavior (i.e of the measure of the biological activity and other physical-chemical properties) of the "virtual" compounds in advance to their expensive experimental essay or even to their synthesis.

In this respect, it is crucial to offer *flexible* and efficient computational tools to explore the "virtual" chemical space, posing a fairly well-defined challenge to CS.

Flexible computational approaches are especially useful when the micro-mechanism of interaction between compounds or, between active molecules and the bio-receptors, are not known in advance, and when coping with noisy experimental data.

[1] See the ref.s in *"What is Machine Learning?"*, `www.di.unipi.it/~micheli/DID`

[2] P. Kirkpatrick; C. Ellis; et al. *Nature* 432(7019), 823-865, 2004.

On the side of the CS methodologies, Machine Learning (ML) provides powerful approaches for the construction of hypotheses (functions) from sample data in order to achieve predictive models. ML is ideally suited to tackle complex problems where there is a lack of a (complete) theory explaining the underlying phenomena, and the presence of noisy data. These characteristics clearly match the flexibility needed in the Cheminformatics context, while posing new demands for reliable predictive systems.

A further peculiarity of the biochemical problems is that they are characterized by quite complex domains where the managing of relationships and structures is relevant to achieve suitable modeling of the solutions. In particular, the chemical space is naturally organized as a set of molecular graphs, traditionally represented by the atoms-bonds based chemical structure formulas. More in general, graphs are powerful tools in CS to model real data and to achieve effective solutions. The field of ML for structured domains (Structured Domain Learning - SDL) aims at dealing with structured data in the form of variable size labeled structures, such are sequences, trees and graphs.

In this respect SDL provides the suitable tools to *learn* a mapping between the structured domain representing molecules and the property/activity space in order to achieve prediction directly from structures. Major advantages of an adaptive and direct approach to the exploration of chemical structured domains are in the preservation of the graph information, in the abstraction with respect to the reduced need of prior knowledge, and in the generality of the affordable tasks. Moreover, SDL can contribute to address the new challenges emerging in the field of Life Science by the means of providing an uniform computational framework for the integration of the analysis of data from chemical, toxicological, genetical, and clinical sources.

The CIML group in Pisa has a research activity for new SDL methodologies, and has been among the pioneers in the introduction of adaptive processing of structures for chemical applications, developing successfully new approaches for the QSPR/QSAR (Quantitative Structure-Property/Activity Relationship) of small and macro molecules.

The CIML group. The Computational Intelligence & Machine Learning group at the Computer Science Department - University of Pisa (*http://www.di.unipi.it/groups/ciml*) has experience in the design and development of innovative Intelligent Systems methodologies, including Neural Networks, Probabilistic Learning, Structured Domain Learning, and Signal and Image processing techniques. More recent developments concern for instance adaptive models for structured information (sequence, tree and graph data), neural networks and kernel approaches for general classes of graphs, unsupervised learning, feature selection, and information fusion.
The group has participated to several EU and Italian funded projects and has a solid tradition of interdisciplinary research, that has led to the design of successful systems in different application domains, including Bioinformatics, Cheminformatics, Robotics, Image Processing and Web-Based Systems for Medicine. Such interdisciplinary vocation is confi rmed by several national and international collaborations that CIML has activated with research units in CS and in the areas of Medicine, Pharmacology, Chemistry, Biology, Robotics and Bioengineering.

Acknowledgments. The author would like to thank past and current collaborators, including researchers from pharmaceutical, chemical and computer science departments (see the CIML site). The author dedicates this work to the memory of Antonina Starita, who founded the CIML group and the Neurolab at the Computer Science department in Pisa.

Deconvolution with nonnegativity constraints

O. Menchi[1] with P. Favati[2], G. Lotti[3] and F. Romani[1]

[1] Department of Computer Science, University of Pisa
[2] IIT - CNR Via G. Moruzzi 1, Pisa
[3] Department of Mathematics, University of Parma

Inverse problems are frequently ill-posed problems. Take for instance the Fredholm integral equation of the first kind

$$g(s) = \int K(s,t) f(t)\, dt, \tag{1}$$

where the square integrable kernel $K(s,t)$ and the right-hand side $g(s)$ are given functions and $f(t)$ is the unknown solution. A typical 2D example of such a problem is the image deconvolution problem, where $f(t)$ and $g(s)$ represent a real object and its image respectively, and $K(s,t)$ represents the imaging system and is responsible for the blurring of the image. In many applications the blurred image $g(s)$ is not available, being replaced by a finite set $\boldsymbol{g}$ of measured quantities, and is degraded by the noise which affects the process of image recording. The problem of restoring $f(t)$ from $\boldsymbol{g}$ is an ill-posed problem and the linear system, which is obtained when equation (1) is discretized, inherits the ill-posedness: the resulting matrix is highly ill-conditioned, and regularization methods must be used to solve the system.

Another important feature of the problem is the nonnegativity of the functions involved in (1) and we expect the solution of the linear system to be nonnegative. Enforcing such a constraint is not an easy task. Iterative methods, often applied as regularization techniques, may give solutions with negative entries.

When the problem of deconvolution is formulated in a statistical frame, the data are seen as the realization of a random process, where the nature of the noise is taken into account. This formulation leads to the maximization of a likelihood function which depends on the statistical property assumed for the noise. In the deconvolution of astronomical and medical images a counting process is involved, of photons in the first case and of the rays emitted by body organs or by some injected substance in the second case. The noise is mainly due to the fluctuations in this counting process, which obeys to Poisson statistics. But there is also the readout noise, due to imperfections of the recording device, which obeys Gaussian statistics.

In our research we study, under the unifying statistical approach, the performances of some iterative methods coupled with suitable strategies for enforcing nonnegativity and other ones which instead naturally embed nonnegativity. Many of these methods can be seen as belonging to the framework of Scaled Gradient Projection methods (SGP). In general, the comparisons among methods take into account the relative error of the best reconstructed solution and the cost required to obtain it. We feel that the subject deserves a more systematic investigation, taking into consideration also other performance indicators, as the possibility to use a stopping rule based on the discrepancy principle (a standard technique for choosing the best regularization parameter) or the sensitivity to this choice.

An extensive experimentation has been conducted in simulated contexts using, besides the 2D image deconvolution problems, also 1D problems, arising in the computation of inverse transformations. These 1D problems are smaller than the 2D ones, but they are in general worst conditioned and can be used as a benchmark for detecting the best methods to tackle the much larger 2D problems. The major part of the computational cost of the methods we consider is due to the computation of matrix by vector products. In the 2D case the coefficient matrix of the system has frequently a 2-level Toeplitz or circulant structure and the computation can be performed by means of FFT. This feature restricts the search for efficient methods to iterative procedures requiring only matrix by vector products.

For this research, the set of Java libraries MAJA has been used. MAJA was developed to deal with structured matrices by means of natural storage techniques allowing an optimal use of the memory (main requirement with the huge sizes of the data involved in medical imaging). The behavior of "any" solving method applied to "any" problem can be implemented, and this versatility is obtained thanks to a high code modularity and a totally object-oriented approach. The complete portability of the code is guaranteed on all the platforms. The multithreaded structure of the language allows a simple parallelization of the execution.

Models and Languages for Service Component Ensembles

Ugo Montanari

Dipartimento di Informatica, University of Pisa, Italy
ugo@di.unipi.it

Ensambles are: (i) widely distributed, open-ended systems, (ii) with complex interactions and behaviors, but component-based, (iii) which adapt to changing environments and requirements, and (iv) where designers can control and engineer emergent behaviour with static and dynamic support from formal methods. Theory and practice of ensambles is the research area of the European FET IP project ASCENS, which will start soon and which includes eleven participants: München, Pisa, Firenze, Fraunhofer Berlin, Grenoble, Modena/Reggio Emilia, Dublin, Bruxelles, Lausanne, Volkswagen and Zimory. The project focuses on service component ensambles and includes workpackages covering foundational models, specification and programming languages and their logic, knowledge representation, adaptation, correctness, tool integration, and engineering/best practices. The case studies include robot swarms, e-vehicle mobility and cloud computing.

The participants for the University of Pisa are: Bruni, Corradini, Ferrari, Gadducci and Montanari. Their contribution is expected mainly in the areas of foundational models, languages/logic, and methods for correctness, even if the integration with the knowledge representation and adaptation parts is expected to require lots of work also from the more formal side, and to be extremely challenging. Some of our research themes are shortly described below.

Modeling Compositional Open Nets with Complex Behavior. In most process calculi, communication networks are not explicitly represented. Rather, either their structure is extremely simple, e.g. shared channels with handshaking and name passing in the π -calculus and tuple spaces with asynchronous communication in Linda, or it is left to the programmer to model the network as an ordinary process. We suggest to extend the option of being global/restricted and of exhibiting visible behavior also to network components and connectors. Rules combining possibly shared network fragments can determine e.g. permissible load and access rights of the resulting network.

Among the proposals of flexible connectors we can mention [8], where a chemical abstract machine with restriction is presented; [9], where suitable categorical diagrams (glues) represent connectors; [5], where connectors represent different synchronization policies; [2], where a combination of basic connectors, can model complex global constraints in *REO*; and [4], where Sifakis and coauthor base a general model on behavior, interaction, and priority. We are confident to reconcile all these approaches.

Resource-conscious models may focus on operations for resource creation and manipulation. Indexed labeled transition systems, as the nominal bialgebra models for the π-calculus, are likely to capture complex resources like memory or constraints. The deallocation scheme represented by *History Dependent Automata* [12] could be useful for making finite state model checking more applicable.

Negotiate, Commit, Execute. We find it useful to distinguish these three phases, which occur in every cooperation. In *negotiation* the prospective participants establish some guarantees in order to define a sort of contract. In *commit* participants can either accept or reject the contract. If they accept, the contract will bind their behaviours in the *execution* phase to guarantee globally correct actions. This NCE scheme allows part of the verification process to be carried on at run time. It is described in the contribution by Roberto Bruni to this workshop.

From Local Declarative Information to Global Knowledge and Back. Open endedness requirements and the autonomic approach have increased the need of declarative information. Typically, local declarative information is generated by processes or it is the result of physical measurements. Then some global mechanism computes pieces of global knowledge, e.g. the routing tables of a network. Its semantics can be often represented in the form of a closure operator (i.e. with extensiveness, idempotence and monotonicity properties), computable by chaotic iteration [1]. The resulting global information is then made available to local agents which can take it into account in their strategic decisions.

A good example of this approach is constraint programming, and its local propagation algorithms [1]. Even a better example is concurrent constraint programming [14, 7], where the closure operators are represented as rules of entailment operating on sets of tokens, which are produced by tell and read by ask operators. Another example is the well understood application of type systems to guarantee session properties [10, 6]. The typing rules represent the closure operators. In other cases the actors can have a dedicated structure for extracting such information, like the Autonomic Communication Elements of the CASCADAS project [11].

Feedback Models for Control/Adaptation and Competitive Behavior. Often the nondeterminism which corresponds to *internal* choices is caused by aspects of the system not fully represented in the model. We can take advantage of this kind of nondeterminism to establish feedback loops activated by local/global knowledge. Note that the feedback could involve low level control issues, e.g. maintaining the values of certain parameters constant; or could imply higher level capabilities, like synthesizing a new plan from facts and properties; or tuning the parameters of an algorithm according to the adequacy of its results. The optimization algorithms could sometimes be embedded into the closure operators needed to propagate the local information. For instance in [3] a Warshall-Floyd

shortest path algorithm for the multicriteria case is embedded in a soft constraint system based on a Hoare powerdomain for the min-plus semiring.

The optimization scenario is not the only possible. Often, when competing agents interact no global optimization strategy is realistic. Then the feedback loop could be conveniently implemented as a game. Typically, only Nash equilibria should be considered as possible system behaviors. We propose to combine the flexibility and compositionality of computer science semantic models with the expressiveness of models developed by economic theorists. A recent proposal is [13].

References

1. Krzysztof R. Apt. The essence of constraint propagation. *Theor. Comput. Sci.*, 221(1-2):179–210, 1999.
2. Farhad Arbab. Reo: a channel-based coordination model for component composition. *Mathematical Structures in Computer Science*, 14(3):329–366, 2004.
3. Stefano Bistarelli, Ugo Montanari, and Francesca Rossi. Soft constraint logic programming and generalized shortest path problems. *J. Heuristics*, 8(1):25–41, 2002.
4. Simon Bliudze and Joseph Sifakis. The algebra of connectors - structuring interaction in bip. *IEEE Trans. Computers*, 57(10):1315–1330, 2008.
5. Roberto Bruni, Ivan Lanese, and Ugo Montanari. A basic algebra of stateless connectors. *Theor. Comput. Sci.*, 366(1-2):98–120, 2006.
6. Roberto Bruni and Leonardo Gaetano Mezzina. Types and deadlock freedom in a calculus of services, sessions and pipelines. In José Meseguer and Grigore Rosu, editors, *AMAST*, volume 5140 of *Lecture Notes in Computer Science*, pages 100–115. Springer, 2008.
7. Maria Grazia Buscemi and Ugo Montanari. Cc-pi: A constraint-based language for specifying service level agreements. In Rocco De Nicola, editor, *ESOP*, volume 4421 of *Lecture Notes in Computer Science*, pages 18–32. Springer, 2007.
8. Andrea Corradini, Ugo Montanari, and Francesca Rossi. An abstract machine for concurrent modular systems: Charm. *Theor. Comput. Sci.*, 122(1&2):165–200, 1994.
9. José Luiz Fiadeiro, Antónia Lopes, and Michel Wermelinger. A mathematical semantics for architectural connectors. In Roland Carl Backhouse and Jeremy Gibbons, editors, *Generic Programming*, volume 2793 of *Lecture Notes in Computer Science*, pages 178–221. Springer, 2003.
10. Kohei Honda, Nobuko Yoshida, and Marco Carbone. Multiparty asynchronous session types. In George C. Necula and Philip Wadler, editors, *POPL*, pages 273–284. ACM, 2008.
11. Antonio Manzalini. Autonomic ecosystems: experience of the cascadas project. In Antonio Manzalini, editor, *Autonomics*, page 9, 2008.
12. Ugo Montanari and Marco Pistore. Structured coalgebras and minimal hd-automata for the π-calculus. *Theor. Comput. Sci.*, 340(3):539–576, 2005.
13. Dusko Pavlovic. A semantical approach to equilibria and rationality. In Alexander Kurz, Marina Lenisa, and Andrzej Tarlecki, editors, *CALCO*, volume 5728 of *Lecture Notes in Computer Science*, pages 317–334. Springer, 2009.
14. Vijay A. Saraswat and Martin C. Rinard. Concurrent constraint programming. In *POPL*, pages 232–245, 1990.

May policies change business processes?

Pole la politica permettisi di intervenire su un processo?

Carlo Montangero and Laura Semini

Context. The integration of Business Process Management (BPM) and Service Oriented Architecture (SOA) has been recognized as a promising approach to cope with the current need for *flexible* business support [4]. The problem originates from the pace at which enterprises change their business partners to follow evolving business goals, and, on a smaller scale, by the need to adapt business processes to the fluctuations in the available resources. Notably, these adaptations need to be enacted with minimal disruption of the supporting software operation. Hence, the trend to integrate BPM and SOA, which, however, still requires large efforts by highly skilled personnel (the business analysts) to transform the business rules introduced by business roles into directives to the programmers to update the actual workflows.

Previous work. Within the SENSORIA project we addressed the problem of simplifying the implementation of changes to the business process, introducing the Service-Targeted Policy-Oriented WorkfLow Approach (StPowla – to be read like "Saint Paula") [3,7]. The key idea is to integrate also *policies*, i.e., declarative rules that are deployed or removed dynamically, with BPM and SOA. More precisely, StPowla distinguishes between a *core* description of the process workflow and its *variations*. The integration of BPM and SOA occurs since each business *task* in the core workflow is ultimately carried out by a *service*, i.e., a computational entity characterized by two sets of parameters: the *invocation* ones (related to the service functionality), and the *Service Level* (SL) parameters, which relate to the resources the service exploits to perform. The policies control the choice of the service performing the task, determining the current required SL's. So, the workflow composes coarse–grain business tasks, and the policies control the fine–grain variations in the service level of each task. The integration occurs at the conceptual level and in the supporting environment, rather than at the linguistic level.

We have been working to an environment to model, deploy, and run StPowla business processes [7]. The relevant concepts (workflow, tasks, service level, etc.) to model the workflows in the standard framework of UML are defined in a specific part of the UML4SOA profile [2].

When several policies are composed or applied simultaneously, they might contradict each other: a phenomenon recognized as a problem and referred to as *policy conflict*. In the case of end-user policies the problem is significantly increased. To determine that the rules are free of conflicts, we filter the input statically, to detect those policies that, if entered in the policy engine, would originate conflicts. The advantages include that we can anticipate conflict detection—traditionally performed at run-time—at design-time. To this end, we explored two approaches: theorem proving [5,6] and model–checking [1].

Future. An important step is to augment the range of application of policies to the whole workflow, the project, or the enterprise. Indeed, besides altering the behavior of a single task, policies can express *reconfigurations*, or *generic instructions*. A workflow reconfiguration is the structural change of a workflow instance. Consider a supplier whose business process is to receive an order from a registered customer, and then to process that order. Under certain conditions (e.g. financial pressure), a guarantee may be required from all customers whose order is above a certain amount. We may add a policy to the process, to require a deposit, in the case the order value exceeds the given threshold. In this case the action is to insert a new task. Similarly, a policy can abort the current task and progress to the next task, or ask to wait until a predicate is true before continuing the current workflow instance. An example of generic instruction is not to use in the project services from provider X. The policy must be distributed across all the tasks: it can be expressed in StPowla as a SL of all service invocations.

Another future line of research addresses security. It may happen that a vulnerability is found when the system is in operation, even after an accurate specification, design, and implementation of security requirements. In most cases, it is not possible to update the system on the fly to resolve the problem, and policies can be a valuable means to patch it. For instance, consider the discovery of a new fraud in the ATM network. It is not feasible to update the whole system from one day to the next, while my bank can add a mitigating policy, e.g. that any attempt to withdraw money twice in a week is double checked with a text message to my mobile phone.

In general, the goal of StPowla is to raise the abstraction level at which the variations are specified, so that any stakeholders can adapt the core workflows directly.

References

1. M. ter Beek, S. Gnesi, C. Montangero, and L. Semini. Detecting policy conflicts by model checking UML state machines. In *Feature Interactions in Software and Communication System X*, pages 59–74. IOS Press, 2009.
2. H. Foster, L. Gonczy, N. Koch, P. Mayer, C. Montangero, and D. Varrò. *UML Extensions for Service-Oriented Systems.* LNCS. 2010.
3. S. Gorton, C. Montangero, S. Reiff-Marganiec, and L. Semini. StPowla: SOA, Policies and Workflows. In *Revised Selected Papers of Workshops, ICSOC'07*, volume 4907 of *LNCS*, pages 351–362. Springer, 2007.
4. F. Kamoun. A roadmap towards the convergence of business process management and service oriented architecture. *Ubiquity*, 8(14), 2007. ACM Press.
5. C. Montangero, S. Reiff-Marganiec, and L. Semini. Logic-based detection of conflicts in APPEL policies. In *Int. Symp. on Fundamentals of Software Engineering, FSEN 2007, Tehran, Iran*, volume 4767 of *LNCS*, pages 257–271. Springer, 2007.
6. C. Montangero, S. Reiff-Marganiec, and L. Semini. Logic-based conflict detection for distributed policies. *Fundamenta Informaticae*, 89(4):511–538, 2008.
7. C. Montangero, S. Reiff-Marganiec, and L. Semini. *Model-driven development of adaptable service-oriented business processes.* LNCS. 2010.

Privacy and Anti-Discrimination for a Fair Knowledge Society

Dino Pedreschi, Salvatore Ruggieri, and Franco Turini

Dipartimento di Informatica, Università di Pisa, Italy
{pedre,ruggieri,turini}@di.unipi.it

We live in times of unprecedented opportunities of sensing, storing and analyzing (micro-)data on human activities at extreme detail and resolution, at mass level. Wireless networks and mobile devices record the tracks of our movements. Search engines record the logs of our queries for finding information on the web. Automated payment systems record the tracks of our purchases. Social networking services record our connections to friends, colleagues, collaborators. Ultimately, this abundance of human activity data lies at the heart of the very idea of a *knowledge society*: a society where decisions - small or big, by business or policy makers - can be taken on the basis of reliable knowledge, distilled from the ubiquitous data generated as a side effect of our living. Increasingly sophisticated data analysis and data mining techniques support knowledge discovery from human activity data, enabling the extraction of models, patterns, profiles, and rules of human behavior - a steady supply of knowledge which is needed to support a knowledge-based society. The paradigm shift towards human knowledge discovery comes, therefore, with unprecedented opportunities and risks: *we the people are at the same time the donors of the data that fuel knowledge discovery and the beneficiaries - or the targets - of the resulting knowledge services.*

In order to fully understand the risks, we should consider that the knowledge life-cycle has two distinct phases: *knowledge discovery* and *knowledge deployment*. In the first step, knowledge is extracted from the data; in the second step, the discovered knowledge is used in support of decision making; the two steps may repeat over and over again, either in off-line or in real-time mode. For instance, knowledge discovery from patients health records may produce a model which predicts the insurgence of a disease given a patients demographics, conditions and clinical history; knowledge deployment may consist in the design of a focused prevention campaign for the predicted disease, based on profiles highlighted by the discovered model. Hence, people are both the data providers and the subjects of profiling. In our vision, the risks in each of the two steps in the knowledge life-cycle are:

– *Privacy violation:* during knowledge discovery, the risk is the uncontrolled intrusion into the personal data of the data subjects, namely, of the (possibly unaware) people whose data are being collected, analyzed and mined;

– *Discrimination:* during knowledge deployment, the risk is the unfair use of the discovered knowledge in making discriminatory decisions about the (possibly unaware) people who are classified, or profiled.

Continuing the example, individual patient records are needed to build a prediction model for the disease, but everyones right to privacy means that his/her health conditions shall not be revealed to anybody without his/her specific control and consent. Moreover, once the disease prediction model has been created, no one would like for such a model to be used to profile the applicant of a health insurance or a mortgage, without any transparency and control.

The concrete objective of our research is the definition of a theoretical, methodological and operational framework for *fair knowledge discovery in support of the knowledge society*, where fairness refers to privacy-preserving knowledge discovery and discrimination-aware knowledge deployment. The framework lays on legal and IT foundations, amalgamating data science, analytics, knowledge representation, ontologies, disclosure control, law and jurisprudence of data privacy and discrimination, and quantitative theories thereof. The proposed framework is being put in use as the enabling driver of two novel disruptive technologies:

– A new method for the construction of human knowledge discovery systems that, by design, offer native techno-juridical safeguards of data protection and discrimination freedom;
– A new generation of tools, powered by data mining and data analytics, to support legal protection and the fight against privacy violation and discrimination.

As a proof of concept of our approach, we will conduct large-scale experiments in a set of *playgrounds*, each endowed with real-life big datasets of personal data of specific forms: web activity, mobility/location, social networking and medical records, and aimed at realizing specific knowledge services. In each playground, we will instantiate the proposed framework, in order to demonstrate that a fair knowledge discovery technology is possible.

Our research intends to show that, in specific yet challenging contexts, it is possible to reach a win-win situation, where the opportunities of knowledge discovery and of high quality service provisioning coexist with the legitimate expectations of privacy and fair treatment. In other words, we are creating technologies for human knowledge discovery that, once specific data analysis objectives and minimum quality of service levels have been identified, allow to: 1) protect the privacy of the data subjects, by controlling the probability of re-identification of the data as well as data sensitivity, 2) preserve data utility with respect to the analytical goals, i.e., support the reproducibility of the knowledge discovery process with comparable results with respect to the original data, and 3) guarantee the fair character of the results, thereby ensuring that the deployed knowledge cannot be used for illegal discriminatory purposes. Alternatively, the new tools will provide novel means in support of institutions of legal protection, in their fight against privacy violation and discrimination.

Speeding up local multiple alignments

Nadia Pisanti

joint work with Pierre Peterlongo, Gustavo Akio Tominaga Sacomoto, Alair Pereira do Lago, Marie-France Sagot

Context. Repeats in genomes come under many forms, such as satellites that are approximate repeats of a pattern of up to a few hundred bases appearing in tandem (consecutively) along a genome, segmental duplications that are defined as the duplications of a DNA segment longer than 1 kb, and transposable elements that are sequences of DNA that can move to different positions within a genome in a process known as transposition, or retrotransposition if the element was first copied and the copy then moved. The last two are repeats dispersed along a genome. Most such repeats appear in intergenic regions and were for long believed to be junk DNA, that is DNA that has no specific function although the proportion of repeated segments in a genome can be huge. Transposable elements alone cover up to, for example, 45% of the human and 80% of the maize genomes. This view of repeats as junk is changing though, and there is an increasing interest in efficient algorithms for their detection.

The quantity of DNA in repeated sequences, the frequency of the repeat (that is, the number of times a given sequence is present per genome), and its conservation, show great variability across species. Frequencies from 100 to 1,000,000 have been observed, and the quantities of DNA involved range from 15 to 80 percent of a whole genome. Families of repeated sequences exhibit a degree of similarity among their members varying from perfect matching to matching of only two-thirds of the nucleotides. All these characteristics, plus the fact that in order to identify such repeats, it is necessary to work with whole genomes, makes the identification of repeated elements a very hard computational problem.

In this paper, we focus on the problem of eliminating from the input sequences as many regions as possible that are guaranteed not to contain any repeats of the type and characteristics specified.

Methods. We designed a very strong necessary condition for 2 strings to be at edit distance at most d and an efficient algorithm that checks it. Moreover, we also introduce a framework for detecting whether fragments of the input data fulfill the requirement with respect to at least $r-1$ other fragments belonging to distinct input strings, hence extending the use of filters to the case of *multiple* repetitions.

The necessary condition checked by TUIUIU is actually a series of three possible level of selectivity, resulting in as many versions of the filter. The first condition (already introduced in previous work) requires that two strings share at least a certain number of *q-grams* and that, moreover, in the virtual alignment matrix of the two strings, these result in matches that lay in a parallelogram shaped area.

The second (further) condition that TUIUIU imposes is very simple: for w and w' to be a $(L, 2, d)$-Erepeat, the q-grams that they must share have to occur in w at distinct positions. This apparently trivial condition actually resulted to give a substantial contribution to the strength of the filter in that TUIUIU can check it in negligible constant time and it does increase the selectivity.

The third and most strict condition additionally imposes that there is a set of p_2 shared q-grams that occur in w and w' in the same order. This third condition, involving Longest Common Subsequence (LCS) computations, checks the conservation in the order of the shared q-grams. It requires some extra time to be checked, but experiments showed evidence of the fact that, since this is done only for pairs of strings w and w' that already survived the previous condition, then the delay is limited. In practice, this most restrictive constraints proved was worth to be used in many interesting applications, as, for example, while using values of d larger than 10% of L.

In order to require that the repeat occurs in r distinct sequences, TUIUIU slides a window over all the input sequences. At each moment, it considers the window itself w and all the remaining sequences virtually divided into blocks that are candidate to contain w'. For the first position of the window, it builds an index of all its q-grams, and stores how many of them belong to each block. For every new position of the window, updating this information is very simple as w simply drops a q-gram and acquires a new one. It is thus also easy to check, for each block, whether it has enough shared q-grams. If for w there are enough blocks that satisfy the condition, then w is conserved, otherwise, w is discarded.

Results. We tested TUIUIU on random synthetic sequences with planted artificial repeats using a very wide range of parameters. We also tested it on three sets of real data, the bacterium Neisseria meningitidis strain MC58, the human chromosome 22, and the dataset known in the literature as CFTR (for Cystic Fibrosis Transmembrane conductance Regulator), adopting a similarly wide range of parameter sets. We found that our first additional filtration condition clearly leads to better results with negligible extra time, for all kinds of data and almost all parameter sets, with respect to the conditions previously used in the literature. Moreover, we also found that our second additional filtration condition considerably improves the selectiveness, with some time overhead, and becomes clearly advantageous mostly for large error rates. Our method may also be used to find anchors for global multiple aligners. We thus expect that our filter could serve as a preprocessing step to a local multiple alignment tool. To this purpose, TUIUIU was applied as a preprocessing step of a multiple alignment application, leading to an overall execution time (filter plus alignment) on average 63 and at best 530 times smaller than before (direct alignment) and also, in some cases, to a qualitative improvement of the alignment obtained.

Robust network design

Maria Grazia Scutellà[1]

Dipartimento di Informatica, Università di Pisa

Abstract

A crucial assumption in many network design problems is that of knowing the traffic demands in advance. Unfortunately, measuring and predicting traffic demands are difficult problems. Moreover, often communication patterns change over time, and therefore we are not given a *single* static traffic demand, but instead a set of *non-simultaneous* traffic demands. The network should be able to support any traffic demand that is from the given set.

Several methodologies have been proposed to address the traffic demand uncertainty in network design problems, such as *Robust Optimization*. Here we provide an overview of the main results which have been achieved in modelling and solving network design problems in the framework of robust optimization. We then display some promising avenues of research.

Robust network design: an overview

Let $\mathcal{G}$ be a communication network, and $\mathcal{K}$ be a set of users that wish to communicate, expressed in terms of origin-destination pairs. Usually the traffic demands associated with the origin-destination pairs are not known in advance, but can only be forecast or estimated. This situation fits perfectly into the framework of *Robust Optimization*, that entails modeling optimization problems with uncertain parameters to obtain a solution that is guaranteed to be 'good' for all possible realizations of the parameters in given *uncertainty sets*. Let $\mathcal{D}$ denote the set of the estimated uncertain demands, while c_{ij} be the non-negative cost of installing a unit of capacity along the link (i, j). The *Robust network design problem* (RND) then consists of determining a minimum cost capacity allocation for the links of $\mathcal{G}$ such that the network is able to route each demand in $\mathcal{D}$.

Several variants and generalizations of RND have been proposed in the literature in the last decade, with the aim of modelling and solving relevant aspects in practical applications. Concerning the routing constraints, each origin-destination pair may be required to communicate through a single path (*unsplittable* routing), or the traffic can be split among different paths (*splittable* routing). In addition, the routing can be *dynamic*, i.e., it can change as the demand varies in $\mathcal{D}$, or *static*, i.e., the same routing template must be used for each demand in $\mathcal{D}$ [1]. Static routing can be preferable in applications where migrating from one routing to another one is costly. In general, splittable routing leads to a cheaper solution than unsplittable routing, and dynamic routing

leads to a cheaper solution than static routing. From a time complexity perspective, the splittable static case is polynomially solvable [1]. On the other hand, in the unsplittable case RND is coNP-Hard. RND is also difficult in the splittable dynamic case, as it is coNP-Hard even for the so-called *Hose* model [3]. Thus, dynamic routing is, in general, substantially more difficult than static routing. This has motivated the study of "intermediate scenarios" such as the one where the demands in $\mathcal{D}$ can be served by *two alternative* routing templates [8] [9], which allows one to obtain cheaper solutions than static routing while being computationally tractable in some cases. Another possible approach is to study special cases of RND that are solvable in polynomial time due to the special structure of the demand polyhedron $\mathcal{D}$, such as the ones addressed in [6]. See [2] for a detailed survey on RND, its variants and its generalizations.

Some avenues of research

The research on robust network design has been essentially theoretical. In a few cases there have been computational studies on peculiar robust models, usually involving a comparison between robust and nominal approaches of the same kind. An interesting line of research is to compare robust models of different kinds, in order to assess their efficiency and the quality of the returned solutions. Some steps in this direction can be found in [7], which reports the results of a preliminary computational comparison of robust models of different kinds in a telecommunications setting. We aim to enlarge the computational analysis to additional robust models and benchmark instances. Furthermore, we plan to investigate whether the use of two (or a constant number of) alternative routing templates [8] may provide good results in practice. This aspects will be investigated in telecommunications and transportation settings.

References

1. D. Applegate, E. Cohen "Making intra-domain routing robust to changing and uncertain traffic demands: understanding fondamental tradeoffs" *Proc. of SIGCOMM*, 2003
2. C. Chekuri "Routing and network design with robustness to changing or uncertain traffic demands" *SIGACT News* **38(3)**, 106–129, ACM, New York, NY, USA, 2007
3. C. Chekuri, G. Oriolo, M.G. Scutellà, F.B. Shepherd "Hardness of Robust Network Design" *Networks* **50(1)**, 50–54, 2007
4. N.G. Duffield, P. Goyal, A. Greenberg, P. Mishra, K.K. Ramakrishnan, J. E. van der Merwe "A Flexible Model for Resource Management in Virtual Private Networks" *Proc. of SIGCOMM*, 1999
5. J. Andrew Fingerhut, S. Suri, J. Turner "Designing Least-Cost Nonblocking Broadband Networks" *Journal of Algorithms* **24(2)** , 287–309, 1997
6. A. Frangioni, F. Pascali, M.G. Scutellà "Static and dynamic routing under the single-source Hose model" *TR* **09-23**, Dip. di Informatica, Università di Pisa, 2009
7. F. Pascali "Chance constrained network design" *Ph.D. Dissertation*, Mathematics for Decision Sciences, Università di Pisa, 2009

8. M.G. Scutellà "On improving optimal oblivious routing" *Operations Research Letters* **37(3)**, 197–200, 2009
9. M.G. Scutellà "Hardness of some optimal oblivious routing generalizations" *TR* **10-05**, Dip. di Informatica, Università di Pisa, 2010