IN PARTNERSHIP WITH:
CNRS

Université Charles de Gaulle (Lille 3)

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2012

Project-Team MOSTRARE

Modeling Tree Structures, Machine Learning, and Information Extraction

IN COLLABORATION WITH: Laboratoire d'informatique fondamentale de Lille (LIFL)

RESEARCH CENTER
Lille - Nord Europe

THEME
Knowledge and Data Representation and Management

Table of contents

Project-Team MOSTRARE

Keywords: Machine Learning, Databases, Data, Web, Logics, Tree Automata

The MOSTRARE *project was created in 2004 and it will be closed at the end of 2012. Two start-up projects will follow:* LINKS *on linking dynamic data and* MAGNET *on learning for information networks.*

Creation of the Project-Team: April 01, 2004 .

1. Members

Research Scientists

Joachim Niehren [Senior Researcher (DR2), Team leader, HdR]
Gemma Garriga [Junior Researcher (CR1)]
Pascal Denis [Junior Researcher (CR1) since February 2012]

Faculty Members

Rémi Gilleron [Professor, HdR]
Iovka Boneva [Associate Professor]
Anne-Cécile Caron [Associate Professor]
Aurélien Lemay [Associate Professor]
Yves Roos [Associate Professor]
Sophie Tison [Professor, HdR]
Marc Tommasi [Professor, vice leader, HdR]
Fabien Torre [Associate Professor]
Sławek Staworko [Associate Professor]
Mikaela Keller [Associate Professor]
Angela Bonifati [Professor]

Engineer

Denis Debarbieux [INRIA, since December 2010]

PhD Students

Benoît Groz [AMN fellowship, until September 12, now postdoc in Tel Aviv]
Grégoire Laurence [MESR, since October 2008]
Jean-Baptiste Faddoul [CIFRE XEROX, until June 12, now engineer OWI Technologies]
Jean Decoster [MESR, since October 2009]
Antoine Ndione [INRIA fellowship, since October 2010]
Tom Sebastian [CIFRE INNOVIMAX, since December 2010]
Thomas Ricatte [CIFRE SAP RESEARCH, since May 2011]
Adrien Boiret [AMN fellowship, since September 2011]
David Chatel [INRIA and région NPdC fellowship, since September 2012]
Radu Ciucanu [MESR, since October 2012]

Post-Doctoral Fellows

Antonino Freno [postdoc since June 2011]
Guillaume Bagan [INRIA, postdoc until August 2012, now engineer in the project Music Story]

Administrative Assistant

Julie Jonas [shared by 2 projects]

2. Overall Objectives

2.1. Presentation

The objective of MOSTRARE is to develop adaptive document processing methods for XML-based information systems. Adaptiveness becomes important when documents evolve frequently such as on the Web. The particularity of MOSTRARE is that we develop semi-automatic or automatic information extraction approaches that can fully benefit from the available tree structure of XML documents.

Information extraction is an instance of document transformation. In order to exploit the tree structure of XML documents, our goal is to investigate specification languages for tree transformations. These are based on approaches from database theory (such as the W3C standards XQuery and XSLT), automata, logic, and programming languages. We wish to define stochastic models of tree transformations, and to develop automatic or semi-automatic procedures for inferring them. Once available, we want to integrate these learning algorithms into innovative information extraction systems, semantic Web platforms, and document processing engines.

The following two paragraphs summarize our two main research objectives:

Modeling tree structures for information extraction. We wish to continue our work on modeling languages for node selection queries in tree structured documents, that we contributed in the first phase of Mostrare. The new subject of interest of the second phase are XML document transformations and tree transformations that generalize on node selection queries.

Machine learning for information extraction. We wish to continue to study machine learning techniques for information extraction. One new goal is to develop learning algorithms that can induce XML document transformations, based on their tree structure. Another new goal is to explore stochastic machine learning techniques that can deal with uncertainty in document sources.

3. Scientific Foundations

3.1. Modeling XML document transformations

Participants: Guillaume Bagan, Adrien Boiret, Iovka Boneva, Angela Bonifati, Anne-Cécile Caron, Benoît Groz, Joachim Niehren, Yves Roos, Sławek Staworko, Sophie Tison, Antoine Ndione, Tom Sebastian.

XML document transformations can be defined in W3C standards languages XQuery or XSLT. Programming XML transformations in these languages is often difficult and error prone even if the schemata of input and output documents are known. Advanced programming experience and considerable programming time may be necessary, that are not available in Web services or similar scenarios.

Alternative programming language for defining XML transformations have been proposed by the programming language community, for instance XDuce [46], Xtatic [44], [49], and CDuce [35], [36], [37]. The type systems of these languages simplify the programming tasks considerably. But of course, they don't solve the general difficulty in programming XML transformations manually.

Languages for defining node selection queries arise as sub-language of all XML transformation languages. The W3C standards use XPath for defining monadic queries, while XDuce and CDuce rely on regular queries defined by regular pattern equivalent to tree automata. Indeed, it is natural to look at node selection as a simple form of tree transformation. Monadic node selection queries correspond to deterministic transformations that annotate all selected nodes positively and all others negatively. N-ary node selection queries become non-deterministic transformations, yielding trees annotated by Boolean vectors.

After extensive studies of node selection queries in trees (in XPath and many other languages) the XML community has started more recently to formally investigate XML tree transformations. The expressiveness and complexity of XQuery are studied in [48], [57]. Type preservation is another problem, i.e., whether all trees of the input type get transformed into the output type, or vice versa, whether the inverse image of the output type is contained in the input type [52], [50].

The automata community usually approaches tree transformations by tree transducers [42], i.e., tree automata producing output structure. Macro tree transducers, for instance, have been proposed recently for defining XML transformations [50]. From the view point of logic, tree transducers have been studied for MSO definability [43].

3.2. Machine learning for XML document transformations

Participants: Jean Decoster, Pascal Denis, Jean-Baptiste Faddoul, Rémi Gilleron, Grégoire Laurence, Aurélien Lemay, Joachim Niehren, Sławek Staworko, Marc Tommasi, Fabien Torre, Gemma Garriga, Antonino Freno, Thomas Ricatte, Mikaela Keller.

Automatic or semi-automatic tools for inferring tree transformations are needed for information extraction. Annotated examples may support the learning process. The learning target will be models of XML tree transformations specified in some of the languages discussed above.

Grammatical inference is commonly used to learn languages from examples and can be applied to learn transductions. Previous work on grammatical inference for transducers remains limited to the case of strings [38], [53]. For the tree case, so far only very basic tree transducers have been shown to be learnable, by previous work of the Mostrare project. These are node selecting tree transducer (NSTTs) which preserve the structure of trees while relabeling their nodes deterministically.

Statistical inference is most appropriate for dealing with uncertain or noisy data. It is generally useful for information extraction from textual data given that current text understanding tools are still very much limited. XML transformations with noisy input data typically arise in data integration tasks, as for instance when converting PDF into XML.

Stochastic tree transducers have been studied in the context of natural language processing [45], [47]. A set of pairs of input and output trees defines a relation that can be represented by a 2-tape automaton called a *stochastic finite-state transducer* (SFST). A major problem consists in estimating the parameters of such transducer. SFST training algorithms are lacking so far [41].

Probabilistic context free grammars (pCFGs) [51] are used in the context of PDF to XML conversion [39]. In the first step, a labeling procedure of leaves of the input document by labels of the output DTD is learned. In the second step, given a CFG as a generative model of output documents, probabilities are learned. Such two steps approaches are in competition with one step approaches estimating conditional probabilities directly.

A popular non generative model for information extraction is *conditional random fields* (CRF, see a survey [54]). One main advantage of CRF is to take into account long distance dependencies in the observed data. CRF have been defined for general graphs but have mainly been applied to sequences, thus CRF for XML trees should be investigated.

So called *structured output* has recently become a research topic in machine learning [56], [55]. It aims at extending the classical categorization task, which consists to associate one or some labels to each input example, in order to handle structured output labels such as trees. Applicability of structured output learning algorithms remains to be asserted for real tasks such as XML transformations.

4. Application Domains

4.1. Context

XML transformations are basic to data integration: HTML to XML transformations are useful for information extraction from the Web; XML to XML transformations are useful for data exchange between Web services or between peers or between databases. Doan and Halevy [40] survey novel integration tasks that appear with the Semantic Web and the usage of ontologies. Therefore, the semi-automatic generation of XML transformations is a challenge in the database community and in the semantic Web community.

Also, XML transformations are useful for document processing. For instance, there is need of designing transformations from documents organized w.r.t visual format (HTML, DOC, PDF) into documents organized w.r.t. semantic format (XML according to a DTD or a schema). The semi-automatic design of such transformations is obviously a very challenging objective.

Furthermore, quite some activities of Mostrare concern efficient evaluation of XPath queries on XML documents and XML streams. XPath is fundamental to all XML standards, in particular to XQuery, XSLT, and XProc.

5. Software

5.1. FXP

Participants: Joachim Niehren [correspondant], Denis Debarbieux, Tom Sebastian.

Software Self-Assessment: A-3, SO-4, SM-3, EM-3, SDL-4

The FXP language is a temporal logic for a fragment of Forward XPath that is suitable for querying XML streams. The FXP library of the Mostrare project of Inria Lille provides rewriting tool that removes backward axis, a compiler of the FXP library to nested word automata and an efficient query answering algorithm for nested word automata on XML streams.
FXP is developed in the Inria transfer project QuiXProc in cooperation with Innovimax. Both a professional and a free version are available. The owner is Inria. A new release was published in October 2012.

See also the web page http://fxp.lille.inria.fr/.

- Version: FXP v1.1.0

5.2. QuixPath

Participants: Joachim Niehren [correspondant], Denis Debarbieux, Tom Sebastian.

Software Self-Assessment: A-3, SO-4, SM-3, EM-3, SDL-4

The QuiXPath language is a large fragment of XPath with full support for the XML data model. The QuiXPath library provides a compiler from QuiXPath to FXP. QuiXPath is developed in the Inria transfer project QuiXProc in cooperation with Innovimax. Both, a free open source and a professional version are available. The ownership of QuiXPath is shared between Inria and Innovimax. The main application of QuiXPath is its usage in QuiXProc, an professional implementation of the W3C pipeline language XProc own by Innovimax. A new release was published in October 2012.

See also the web page http://code.google.com/p/quixpath/.

- Version: QuixPath v1.1.0

5.3. VOLATA

Participant: Fabien Torre [correspondant].

Software Self-Assessment: A-2, SO-4, SM-2, EM-2, SDL-2

VOLATA provides several machine learning algorithms for attribute-value inference, grammatical inference and inductive logic programming.

See also the web page http://www.grappa.univ-lille3.fr/~torre/Recherche/Softwares/volata/.

- ACM: I.2.6

5.4. JProGraM

Participant: Antonino Freno [correspondant].

Software Self-Assessment: A-3, SO-3-up, SM-2, EM-3, SDL-4.

JProGraM is a GPL-licensed Java library for machine learning and statistical analysis *over* graphs and *through* graphs. Supported models for vectorial data include e.g. Bayesian networks, Markov random fields, Gaussian mixtures, kernel density estimators, and neural networks, whereas random graph tools include small-world networks, preferential-attachment, exponential random graphs, and spectral models (as well as subgraph sampling algorithms). One strong point of the library is the extensive support for continuous random variables.

See also the webpage http://researchers.lille.inria.fr/~freno/JProGraM.html.

6. New Results

6.1. Modeling XML document transformations

Participants: Joachim Niehren, Angela Bonifati, Sophie Tison, Sławek Staworko, Aurélien Lemay, Anne-Cécile Caron, Yves Roos, Benoît Groz, Antoine Ndione, Tom Sebastian.

XML Schema Validation Groz, Staworko et. al. [26] present a new algorithm that tests determinism of regular expressions in linear time. All regular expressions used in DTDs and XML Schemas are required to be deterministic by the recommendation of the W3C. Whether this is the case can indeed been tested in linear time, as shown in this paper. The best known previous algorithm, which was based on the Glushkov automaton, required O(σ|e|) time, where σ is the number of distinct symbols in e. They also show that matching a word w against a deterministic regular expression e can be achieved in combined linear time O(|e| + |w|) for a wide range of cases.

Staworko et. al. studied bounded repairability for regular tree languages modulo the tree edit distance [28].

Ndione, Niehren, and Lemay [33] present a new probabilistic algorithm for approximate membership of words to regular languages modulo the edit distance on words. In the context of XML, this algorithm is relevant for sublinear DTD validity testing. The time complexity of the algorithm is independent of the size of the input word and polynomial in the size of the input automaton and the inverse error precision. All previous property testing algorithms for regular languages run in exponential time.

XML Query Answering Debarbieux, Niehren, Sebastian et. al. [32] present new algorithms for early XPath node selection on XML Streams. Early selection and rejection is crucial for efficiency, while earliest selection and rejection has high computational complexity in the general case. In contrast to all previous approaches, there algorithm does not rely on any expensive static analysis method. Instead, it is based on a compiler from XPath to nested word automata with selection and rejection states that they introduce. They cover a large fragment of downward XPath, with the main restriction that negation is forbidden above descendant axis and disjunctions. Non-determinism is used to deal with descendant axis and disjuctions. High run-time efficiency in practice is obtained by on-the-fly determinization for nested word automata, even in cases where static determinization produces automata of more than exponential size. Our experimental results confirm a very high efficiency in space and time. An implementation of our FXP/QuiXPath system is freely available and used for industrial transfer in the QuiXProc system.

Staworko et. al. tackled prioritized repairing and consistent query answering in relational databases in [20].

External Cooperations with other teams in Lille lead to the following publications [19], [31], [30].

6.2. Machine learning for XML document transformations

Participants: Adrien Boiret, Jean Decoster, Pascal Denis, Jean-Baptiste Faddoul, Antonino Freno, Gemma Garriga, Rémi Gilleron, Mikaela Keller, Grégoire Laurence, Aurélien Lemay, Joachim Niehren, Sławek Staworko, Marc Tommasi, Fabien Torre.

Learning XML Queries. Staworko et. al. [29] proposed learning twig and path queries.

Niehren, Champavère, Gilleron, and Lemay [34] propose new algorithm and learnability result for XML query induction based on schema-guided pruning strategies. Pruning strategies impose additional assumptions on node selection queries that are needed to compensate for small numbers of annotated examples. The class of regular queries that are stable under a given schema-guided pruning strategy was distinguished and shown to be learnable with polynomial time and data. The learning algorithm is obtained by adding pruning heuristics to the traditional learning algorithm for tree automata from positive and negative examples. While justified by a formal learning model, their learning algorithm for stable queries also performs very well in practice of XML information extraction.

Learning XML Transformations. Boiret, Lemay, and Niehren [21] solved the long open question of how to learn rational functions with polynomial time and data. Rational functions are transformations from words to words that can be defined by deterministic string transducers with lookahead. No previous learning results for classes of transducers with look-ahead existed, so this results is relevant for learning XML transformations defined by transducers with look-ahead, as with XSLT.

Multi-task Learning. We address the problem of multi-task learning with no label correspondence among tasks. In [22], Faddoul, Chidlovskii, Gilleron and Torre propose the multi-task Adaboost algorithm with Multi-Task Decision Trees as weak classifiers. They conduct experiments on multi-task datasets, including the Enron email set and Spam Filtering collection. Faddoul successfully defended his PhD thesis [16] in June 2012.

Probabilistic models for large graphs. We propose new approaches for the statistical analysis of large-scale undirected graphs. The guiding idea is to exploit the spectral decomposition of subgraph samples, and in particular their Fiedler eigenvalues, as basic features for density estimation and probabilistic inference. In [24], Freno, Keller, Garriga, and Tommasi develop a conditional random graph model for learning to predict links in information networks (such as scientific coauthorship and email communication). In [25], Freno, Keller, and Tommasi propose instead to estimate joint probability distributions through (non-linear) random fields, applying the resulting model to graph generation and link prediction.

Learning in Multiple graphs Ricatte, Garriga, Gilleron and Tommasi focus on learning from several sources of heterogeneous data. They represent each source as a graph of data and they propose to combine the multiple graphs with the help of small number of labeled nodes. They obtain a kernel that can be used as input to different graph-learning tasks such as node classification and clustering. The paper is under submission. Along a collaboration with physicians, Keller and Tommasi consider graphs that represents the structural connectivity of the brain (connectome). They develop a spatially constrained clustering method, combining heterogeous descriptions of the same objects through the graph of neighborhood on the cortex and the graph of connectivity. The paper is under submission.

Starting PhDs Boneva, Bonifati and Staworko started to supervise the PhD of R. Ciucanu on learning cross-model database mappings. Denis and Tommasi has begun to supervise the PhD of David Chatel on guided clustering for graphs (of texts).

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

7.1.1. QuiXProc: Inria Transfer Project with Innovimax (2010-2012)

Participants: Denis Debarbieux, Joachim Niehren [correspondent], Tom Sebastian.

QuiXProc is an Inria transfer project with Innovimax S.A.R.L in Paris, on the integration of XPath streaming algorithms into XProc, the XML coordination language of the W3C.

7.1.2. Music Story

Participants: Fabien Torre, Mikaela Keller [correspondent], Guillaume Bagan.

The MusicStory project is a transfer project with MusicStory, a company collecting musical metadata from heterogeneous sources. The project entails the design of automated data deduplication and field inference algorithms suited for MusicStory needs.

7.2. Bilateral Grants with Industry

7.2.1. Cifre Xerox (2009-2012)

Participants: Jean-Baptiste Faddoul, Rémi Gilleron, Fabien Torre [correspondent].

Gilleron and Torre continue supervising the PhD thesis (Cifre) of Jean-Baptiste Faddoul together with B. Chidlovski from the Xerox's European Research Center (XRCE).

7.2.2. Cifre Innovimax (2010-2013)

Participants: Tom Sebastian, Joachim Niehren [correspondent].

Niehren continue supervising the PhD thesis (Cifre) of Tom Sebastian on streaming algorithms for XSLT with M. Zergaoui from INNOVIMAX S.A.R.L. in Paris.

7.2.3. Cifre SAP (2011-2014)

Participants: Thomas Ricatte, Gemma Garriga, Rémi Gilleron [correspondent], Marc Tommasi.

Garriga, Gilleron and Tommasi supervise the PhD thesis (Cifre) of Thomas Ricatte together with Yannick Cras from SAP.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. Thèse Inria-Région NPdC (2012-2015)

Participants: David Chatel, Pascal Denis, Marc Tommasi [correspondent].

Denis and Tommasi supervise the PhD thesis of David Chatel on guided clustering. The PhD is funded by INRIA and the "région Nord - Pas de Calais".

8.2. National Initiatives

8.2.1. ANR

8.2.1.1. ANR Lampada (2009-2014)

Participants: Marc Tommasi [correspondent], Rémi Gilleron, Aurélien Lemay, Fabien Torre, Gemma Garriga.

The Lampada project on "Learning Algorithms, Models and sPArse representations for structured DAta" is coordinated by Tommasi from Mostrare. Our partners are the SEQUEL project of Inria Lille Nord Europe, the LIF (Marseille), the HUBERT CURIEN laboratory (Saint-Etienne), and LIP6 (Paris). More information on the project can be found on http://lampada.gforge.inria.fr/.

8.2.1.2. ANR Defis Codex (2009-2012)

Participants: Joachim Niehren [correspondent], Sławek Staworko, Aurélien Lemay, Sophie Tison, Anne-Cécile Caron, Jérôme Champavère.

The Codex project on "Efficiency, Dynamicity and Composition for XML Models, Algorithms, and Systems" and is coordinated by Manolescu (GEMO, Inria Saclay). The other partners of Mostrare there are Geneves (WAM, Inria Grenoble), COLAZZO (LRI, Orsay), Castagna (PPS, Paris 7), and Halfeld (Blois). Public information on Codex can be found on http://codex.saclay.inria.fr/.

8.2.2. Competitivity Clusters

8.2.2.1. FUI Hermes (2012-2015)

Joint project in collaboration with many companies (Auchan, KeyneSoft, Cylande, ...). The main objective is to develop a platform for contextual customer relation management. The project started in November 2012.

8.3. European Initiatives

8.3.1. Collaborations in European Programs, except FP7

MOSTRARE, in collaboration with SEQUEL and Rouen, is part of the Inria Lille - Nord Europe site for the European Network of Excellence in Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL2).

8.3.2. Collaborations with Major European Organizations

Publications [29] and [20] are results of collaborations with the University of Wroclaw and the University of Oxford respectively.

8.4. International Initiatives

8.4.1. Inria International Partners

The ongoing cooperation with our previous international partner at NICTA Sydney has lead to a publication at PODS'2012 [26].

8.5. International Research Visitors

8.5.1. Visits of International Scientists

Jan van den Bussche from the University of Hasselt and Werner Nutt from the University of Bolzano visited Bonifati and Niehren for a recent cooperation.

Fabien Suchanek from the Max-Planck Instutite in Saarbrücken visited Bonifati and Niehren and presented is work in the Mostrare seminar.

Yannis Valegrakis from the University of Trento visited Bonifati and presented his work in the Mostrare seminar.

George Fletcher and Toon Calders from the University of Eindhoven visited Bonifati and Staworko and presented their work in the Mostrare seminar.

8.5.1.1. Internships

Carles Creus from the University of Barcelona visited Boiret, Lemay, and Niehren for 4 months for working on tree transducers and compression.

Pavel Labath from the University of Bratislava visited Debarbieux, Sebastian, and Niehren for working on streaming algorithms for XSLT.

8.5.2. Visits to International Teams

Staworko visited Gabriele Pupis and Cristian Riverson at the Univerity of Oxford [28].

Niehren visited Mikael Benedikt, Georg Gottlob, and Marta Kwiatkowska at the University of Oxford.

Staworko visited Piotr Wieczorek at the University of Warclaw [29].

Groz left for postdoc to the database group of Tova Milo at the University of Haifa in Israel.

9. Dissemination

9.1. Scientific Animation

9.1.1. Invited Talks

Bonifati was invited to BDA 2012 (French Conference on databases) to give a tutorial on "Schema matching and mapping: from Usage to Evaluation". Niehren was invited to the workshop "INQUEST:INnovative QUErying of STreams at Oxford". The title of his presentation was "Querying XML Streams with Networks of Automata". Tommasi was invited to give a tutorial on (probabilistic) grammatical inference in "Journées SDA2 Systèmes Dynamiques, Automates et Algorithmique du pôle Algorithmique et Combinatoire du GDR IM (Informatique Mathématique) du CNRS", see http://jmc2012.colloques.univ-rouen.fr/. Tison gave a lecture on Tree Automata, Turing Machines and Term Rewriting at International School on Rewriting.

9.1.2. Program Committees

A. BONIFATI was member of the program committee of EDBT 2012, SIGMOD 2012, and VLDB 2012.

P. DENIS was member of the program committee of AAAI 2012, ACL 2012, COLING 2012, EACL 2012, ATALA Workshop on Discourse and NLP 2012, TALN 2012, CSE 2012; he was external Member on CS Assistant Professorship search committee at University of Marseille and expert Member on Research Engineer search committee at CNRS.

A. FRENO was member of the program committee of ANNPR 2012.

G. GARRIGA continues to be member of the editorial board of Machine Learning Journal and of the Intelligent Data Analysis Journal. She was member of the program committees of ECML 2012 (area chair), CIKM 2012, ASONAM 2012, ECAI 2012, ICDM 2012 PhD Forum.

R. GILLERON was member of the program committee of ECML 2012.

J. NIEHREN is member of the steering committee of RTA (International Conference on Rewriting Techniques and Applications), of the editorial board of FUNDAMENTA INFORMATICAE. He was in the program committees of CMSB 2012, RTA 2012, TTATT 2012, APWeb 2012, LATA 2012, NCMA 2012, ATANLP 2012.

S. TISON is member of the editorial board of RAIRO - ITA and of the steering committee of RTA and STACS. She was member of the program committee of TTATT 2012 , PODS 2012, LATA2012, SOFSEM 2012

M. TOMMASI was member of the program committee of ATANLP 2012

9.1.3. French Scientific Responsibilities

A. BONIFATI was member of the selection committee in Lille for professor positions.

R. GILLERON was a member of the scientific committee of the Program Programme Blanc SIMI2, ANR. He was member of the national PES commission 27. He was member of the selection committees of assistant professors in Saint Etienne.

S. TISON is head of the computer science lab in Lille (LIFL). She was member of the scientific committee of the program Chaires Industrielles, ANR. She chairs the scientific council of "Pôle de Compétitivité Industries du Commerce". She was member of the national PES commission 27. She is elected member of the “Comité National de la Recherche Scientifique (CoNRS)” since 2012. She was in the AERES evaluation committee of LIAFA and PPS.

M. TOMMASI is leader of the LAMPADA project.

F. TORRE is member of the french national evaluation committee for computer science assistant professors (CNU 27).

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Master (Mocad): Information extraction, 18h, M2, Université Lille 1, France by J. NIEHREN

Master (Mocad): Information extraction, 18h, M2, Université Lille 1, France by P. DENIS

Master (Mocad): Information extraction, 18h, M2, Université Lille 1, France by A. BONIFATI

Master (Mocad): Information extraction, 18h, M2, Université Lille 1, France by A. FRENO

Master: Supervised classification, 45h, M1, Université Lille 1, France by R. GILLERON

Master: Unsupervised classification, 30h, M1, Université Lille 3, France by R. GILLERON

Master: Information retrieval, 30h, M1, Université Lille 3, France by R. GILLERON

Master: Semantic Web, 36h, M2, Université Lille 1, France by R. GILLERON

Master: Advanced Databases, 36h, M1, Université Lille 3, France by S. STAWORKO

Licence: Statistical Learning, 63h, L3, Université Lille 3, France by M. KELLER

Master: Modelization XML, 24h, M1, Université Lille 3, France by S. STAWORKO

Licence: Artificial Intelligence and Logic, 63h, L3, Université Lille 3, France by S. STAWORKO

Master: Introduction to XML, 40h, M1, Université Lille 3, France by S. STAWORKO

Master : Networks, 25h, M2, Université Lille 3, France by M. TOMMASI

Master : XML, 25h, M2, Université Lille 3, France by M. TOMMASI

Master : Databases, XML, 25h, M2, Université Lille 3, France by M. TOMMASI

Master : Document Managment systems, 25h, M1, Université Lille 3, France by M. TOMMASI

Master : Introduction to algorithms, 25h, M1, Université Lille 3, France by M. TOMMASI

Licence : Databases, object oriented programming, 60h, L3, Université Lille 3, France by M. TOMMASI

Master : Advanced algorithms and complexity, M1, 57h, Université Lille 1, by S. TISON

Licence: Databases, 53h, L3, Université Lille 1, France, by A-C. CARON

Master: Advanced databases, 50h, M1, Université Lille 1, France, by A-C. CARON

Master: Semantic Web, 20h, M2, Université Lille 1, France, by A-C. CARON

Licence: XML Technologies, 50h, M1, Université Lille 1, France by Y. ROOS

Master: XML Modelization, 40h, M1, Université Lille 1, France by Y. ROOS

Master: XML Technologies, 16h, M2, Université Lille 3, France by A. LEMAY

9.2.2. Supervision

PhD June 2012: J.-B. FADDOUL, Machine learning and applications to social network analysis. Since Dec. 2008. Supervised by Gilleron and Chidlowskii from XEROX European Research Center (XRCE).

PhD September 2012: B. GROZ, XML database security and access control. Since Sept. 2008. Supervised by Tison and Staworko.

PhD in Progress: G. LAURENCE, Learning XML transformations for data exchange on the web. Since Sept. 2008. Supervised by Tommasi, Niehren, Staworko and Lemay.

PhD in Progress: J. DECOSTER, Statistical relational learning of XML transformations. Since Sept. 2009. Supervised by Tommasi and Torre.

PhD in Progress: A. M. NDIONE, Probabilistic algorithms for tree automata and transducers. Since Sept. 2010. Supervised by Niehren and Lemay.

PhD in Progress: T. SEBASTIAN, Streaming algorithms for XSLT. Since May 2011. Supervised by Niehren.

PhD in Progress: T. RICATTE, Graph Based Learning for Decisional Databases. May 2011. Supervised by Garriga and Gilleron.

PhD in Progress: A. BOIRET, Top-down tree transformations with look-ahead : foundations and learning. Since Sept. 2011. Supervised by Niehren and Lemay.

PhD in Progress: A. CHATEL, Guided Unsupervised Learning. Since Sept. 2012. Supervised by Denis and Tommasi.

PhD in Progress: R. CIUCANU, Towards learning cross-model database mappings. Since Sept. 2012. Supervised by Boneva, Bonifati and Staworko.

9.2.3. Committees

A. BONIFATI was member of the PhD committee of Asma Souihli, Telecom ParisTech (Paris, France) and Rashed Khalil Khalil Salem at Université Lumière Lyon II (Lyon, France); P. DENIS was member of the PhD jury of Emili Sapena at UPC Barcelona; R. GILLERON was member of the PhD jury of J.B. Faddoul (Lille), Y. Benabbas (Lille) and A. Bellet (Saint-Etienne); J. NIEHREN was a reviewer of the Habilitation committee of Pierre Senellart at Telecom Paris. S. STAWORKO was member of the PhD jury of B. Groz (Lille); M. TOMMASI was member of the PhD jury of M. Oita (Telecom Paris); F. TORRE was member of the PhD jury of J.B. Faddoul (Lille)

9.3. Popularization

QuiXProc A video (http://videotheque.inria.fr/videotheque/doc/787) was projected at WWW 2012 (at the booth of Inria); a demonstration “des flux financiers vers les flux XML” was done during the “Rencontres Inria - Industrie: Technologies du web et mobilité au service de l’innovation bancaire et de l’assurance”.

Artistic project Denis, Keller and Gilleron collaborate to the artistic project “This is Major Tom to Ground Control” by Véronique Béland (http://www.veroniquebeland.com/Veronique_Beland/This_is_Major_Tom_to_Ground_Control.html). They contribute to the definition of a probabilistic generator of texts.

10. Bibliography

Major publications by the team in recent years

[1] G. BAGAN, A. DURAND, E. FILIOT, O. GAUWIN. *Efficient Enumeration for Conjunctive Queries over X-underbar Structures*, in "19th EACSL Annual Conference on Computer Science Logic", Tchèque, République Brno, 2010, http://hal.inria.fr/hal-00489955.

[2] I. BONEVA, B. GROZ, S. TISON, A.-C. CARON, Y. ROOS, S. STAWORKO. *View update translation for XML*, in "14th International Conference on Database Theory (ICDT)", Uppsala, Sweden, March 2011, http://hal.inria.fr/inria-00534857/en.

[3] J. CARME, R. GILLERON, A. LEMAY, J. NIEHREN. *Interactive Learning of Node Selecting Tree Transducers*, in "Machine Learning", 2007, vol. 66, n^{o} 1, p. 33–67, http://hal.inria.fr/inria-00087226.

[4] J. CHAMPAVÈRE, R. GILLERON, A. LEMAY, J. NIEHREN. *Efficient Inclusion Checking for Deterministic Tree Automata and XML Schemas*, in "Information and Computation", 2009, vol. 207, n^{o} 11, p. 1181-1208, http://hal.inria.fr/inria-00366082/en/.

[5] J.-B. FADDOUL, B. CHIDLOVSKII, F. TORRE, R. GILLERON. *Boosting Multi-Task Weak Learners with Applications to Textual and Social Data*, in "The Ninth International Conference on Machine Learning and Applications (ICMLA 2010)", États-Unis Hayatt Regency Bethesda, Washington DC, IEEE, Dec 2010, http://hal.inria.fr/inria-00524718.

[6] E. FILIOT, J. NIEHREN, J.-M. TALBOT, S. TISON. *Polynomial Time Fragments of XPath with Variables*, in "26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems", ACM-Press, 2007, p. 205-214, http://hal.inria.fr/inria-00135678.

[7] E. FILIOT, J.-M. TALBOT, S. TISON. *Tree Automata With Global Constraints*, in "International Journal of Foundations of Computer Science", Aug 2010, vol. 21, n^{o} 4, p. 571-596, http://hal.inria.fr/hal-00526987.

[8] O. GAUWIN, J. NIEHREN, S. TISON. *Queries on XML Streams with Bounded Delay and Concurrency*, in "Information and Computation", 2010, http://hal.inria.fr/inria-00491495.

[9] R. GILLERON, F. JOUSSE, M. TOMMASI, I. TELLIER. *Conditional Random Fields for XML Applications*, Inria, 2008, RR-6738, http://hal.inria.fr/inria-00342279/en/.

[10] R. GILLERON, P. MARTY, M. TOMMASI, F. TORRE. *Interactive Tuples Extraction from Semi-Structured Data*, in "2006 IEEE / WIC / ACM International Conference on Web Intelligence", IEEE Comp. Soc. Press, 2006, vol. P2747, p. 997-1004.

[11] A. LEMAY, S. MANETH, J. NIEHREN. *A Learning Algorithm for Top-Down XML Transformations*, in "29th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems", États-Unis Indianapolis, ACM Press, 2010, http://hal.inria.fr/inria-00460489.

[12] W. MARTENS, J. NIEHREN. *On the Minimization of XML Schemas and Tree Automata for Unranked Trees*, in "Journal of Computer and System Science", 2007, vol. 73, n^{o} 4, p. 550-583, http://hal.inria.fr/inria-00088406.

[13] G. PUPPIS, C. RIVEROS, S. STAWORKO. *Bounded repairability for regular tree languages*, in "International Conference on Database Theory (ICDT)", Berlin, Germany, ACM, March 2012, p. 155-168 [*DOI :* 10.1145/2274576.2274593], http://hal.inria.fr/hal-00643100.

[14] S. STAWORKO, J. CHOMICKI, J. MARCINKOWSKI. *Prioritized Repairing and Consistent Query Answering in Relational Databases*, in "Annals of Mathematics and Artificial Intelligence", 2012, http://hal.inria.fr/hal-00643104.

[15] S. STAWORKO, P. WIECZOREK. *Learning Twig and Path Queries*, in "International Conference on Database Theory (ICDT)", Berlin, Germany, March 2012, http://hal.inria.fr/hal-00643097.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[16] J.-B. FADDOUL. *Modèles d'Ensembles pour l'Apprentissage Multi-Tache, avec des taches Hétérogènes et sans Restrictions*, Université Charles de Gaulle - Lille III, June 2012, http://tel.archives-ouvertes.fr/tel-00712710.

[17] B. GROZ. *Vues de sécurité XML: requêtes, mises à jour et schémas.*, Université des Sciences et Technologie de Lille - Lille I, October 2012, http://hal.inria.fr/tel-00745581.

Articles in International Peer-Reviewed Journals

[18] G. GARRIGA, R. KHARDON, L. DE RAEDT. *Mining Closed Patterns in Relational, Graph and Network Data*, in "Annals of Mathematics and Artificial Intelligence", November 2012, http://hal.inria.fr/hal-00754967.

[19] M. LATTEUX, Y. ROOS. *On One-Rule Grid Semi-Thue Systems*, in "Fundamenta Informaticae", 2012, vol. 116, n^{o} 1-4, p. 189-204 [*DOI :* 10.3233/FI-2012-678], http://hal.inria.fr/hal-00749289.

[20] S. STAWORKO, J. CHOMICKI, J. MARCINKOWSKI. *Prioritized Repairing and Consistent Query Answering in Relational Databases*, in "Annals of Mathematics and Artificial Intelligence", 2012, http://hal.inria.fr/hal-00643104.

International Conferences with Proceedings

[21] A. BOIRET, A. LEMAY, J. NIEHREN. *Learning Rational Functions*, in "16th International Conference on Developments of Language Theory", Taipee, Taiwan, Province Of China, August 2012, http://hal.inria.fr/hal-00692341.

[22] J.-B. FADDOUL, B. CHIDLOVSKII, R. GILLERON, F. TORRE. *Learning Multiple Tasks with Boosted Decision Trees*, in "ECML/PKDD - European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases - 2012", Bristol, United Kingdom, Lecture Note in Computer Science, 2012, http://hal.inria.fr/hal-00727749.

[23] A. FRENO. *Semiparametric Pseudo-Likelihood Estimation in Markov Random Fields*, in "AISTATS 2012 - Fifteenth International Conference on Artificial Intelligence and Statistics", La Palma, Canary Islands, Spain, 2012, http://hal.inria.fr/hal-00662933.

[24] A. FRENO, M. KELLER, G. GARRIGA, M. TOMMASI. *Spectral Estimation of Conditional Random Graph Models for Large-Scale Network data*, in "UAI 2012 - 28th Conference on Uncertainty in Artificial Intelligence", Avalon, United States, 2012, http://hal.inria.fr/hal-00714446.

[25] A. FRENO, M. KELLER, M. TOMMASI. *Fiedler Random Fields: A Large-Scale Spectral Approach to Statistical Network Modeling*, in "Neural Information Processing Systems (NIPS)", Lake Tahoe, United States, MIT Press, 2012, vol. 25, http://hal.inria.fr/hal-00750345.

[26] B. GROZ, S. MANETH, S. STAWORKO. *Deterministic Regular Expressions in Linear Time*, in "PODS-31th ACM Symposium on Principles of Database Systems", Scottsdale, United States, 2012, 12, http://hal.inria.fr/inria-00618451.

[27] M. JOHN, M. NEBUT, J. NIEHREN. *Knockout Prediction for Reaction Networks with Partial Kinetic Information*, in "Verification, Model Checking, and Abstract Interpretation", Rom, Italy, January 2013, http://hal.inria.fr/hal-00692499.

[28] G. PUPPIS, C. RIVEROS, S. STAWORKO. *Bounded repairability for regular tree languages*, in "International Conference on Database Theory (ICDT)", Berlin, Germany, ACM, March 2012, p. 155-168 [*DOI :* 10.1145/2274576.2274593], http://hal.inria.fr/hal-00643100.

[29] S. STAWORKO, P. WIECZOREK. *Learning Twig and Path Queries*, in "International Conference on Database Theory (ICDT)", Berlin, Germany, March 2012, http://hal.inria.fr/hal-00643097.

Conferences without Proceedings

[30] F. COUTTE, M. JOHN, M. BÉCHET, M. NEBUT, J. NIEHREN, V. LECLÈRE, P. JACQUES. *Synthetic Engineering of Bacillus subtilis to Overproduce Lipopeptide Biosurfactants*, in "9th European Symposium on Biochemical Engineering Science", Istanbul, Turkey, 2012, http://hal.inria.fr/hal-00717261.

[31] M. JOHN, F. COUTTE, M. NEBUT, P. JACQUES, J. NIEHREN. *Knockout Prediction for Reaction Networks with Partial Kinetic Information: Application to Surfactin Overproduction in Bacillus subtilis*, in "3rd International Symposium on Antimicrobial Peptides", Lille, France, June 2012, http://hal.inria.fr/hal-00702295.

Research Reports

[32] D. DEBARBIEUX, O. GAUWIN, J. NIEHREN, T. SEBASTIAN, M. ZERGAOUI. *Early XPath Node Selection on XML Streams*, Inria, March 2012, 12, http://hal.inria.fr/hal-00676178.

[33] A. NDIONE, J. NIEHREN, A. LEMAY. *Approximate Membership for Regular Languages modulo the Edit Distance*, Inria, February 2012, http://hal.inria.fr/hal-00666288.

Other Publications

[34] J. NIEHREN, J. CHAMPAVÈRE, R. GILLERON, A. LEMAY. *Query Induction with Schema-Guided Pruning Strategies*, July 2013, journal submission, http://hal.inria.fr/inria-00607121.

References in notes

[35] V. BENZAKEN, G. CASTAGNA, A. FRISCH. *CDuce: an XML-centric general-purpose language*, in "ACM SIGPLAN Notices", 2003, vol. 38, n^{o} 9, p. 51–63.

[36] V. BENZAKEN, G. CASTAGNA, C. MIACHON. *A Full Pattern-Based Paradigm for XML Query Processing.*, in "PADL", Lecture Notes in Computer Science, Springer Verlag, 2005, p. 235-252.

[37] G. CASTAGNA. *Patterns and Types for Querying XML*, in "10th International Symposium on Database Programming Languages", Lecture Notes in Computer Science, Springer Verlag, 2005, vol. 3774, p. 1 - 26.

[38] B. CHIDLOVSKII. *Wrapping Web Information Providers by Transducer Induction*, in "Proc. European Conference on Machine Learning", Lecture Notes in Artificial Intelligence, 2001, vol. 2167, p. 61 – 73.

[39] B. CHIDLOVSKII, J. FUSELIER. *A probabilistic learning method for XML annotation of documents*, in "Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05)", 2005, p. 1016-1021.

[40] A. DOAN, A. Y. HALEVY. *Semantic Integration Research in the Database Community: A Brief Survey*, in "AI magazine", 2005, vol. 26, n^{o} 1, p. 83-94.

[41] J. EISNER. *Parameter Estimation for Probabilistic Finite-State Transducers*, in "Proceedings of the Annual meeting of the association for computational linguistic", 2002, p. 1–8.

[42] J. ENGELFRIET. *Bottom-up and top-down tree transformations. A comparision*, in "Mathematical System Theory", 1975, vol. 9, p. 198–231.

[43] J. ENGELFRIET, S. MANETH. *Macro tree transducers, attribute grammars, and MSO definable tree translations*, in "Information and Computation", 1999, vol. 154, n^{o} 1, p. 34–91.

[44] V. GAPEYEV, B. PIERCE. *Regular Object Types*, in "European Conference on Object-Oriented Programming", 2003, http://www.cis.upenn.edu/~bcpierce/papers/regobj.pdf.

[45] J. GRAEHL, K. KNIGHT. *Training tree transducers*, in "NAACL-HLT", 2004, p. 105-112.

[46] H. HOSOYA, B. PIERCE. *Regular expression pattern matching for XML*, in "Journal of Functional Programming", 2003, vol. 6, n^{o} 13, p. 961-1004.

[47] K. KNIGHT, J. GRAEHL. *An overview of probabilistic tree transducers for natural language processing*, in "Sixth International Conference on Intelligent Text Processing", 2005, p. 1-24.

[48] C. KOCH. *On the complexity of nonrecursive XQuery and functional query languages on complex values*, in "24th SIGMOD-SIGACT-SIGART Symposium on Principles of Database systems", ACM-Press, 2005, p. 84–97.

[49] M. Y. LEVIN, B. PIERCE. *Type-based Optimization for Regular Patterns*, in "10th International Symposium on Database Programming Languages", Lecture Notes in Computer Science, 2005, vol. 3774.

[50] S. MANETH, A. BERLEA, T. PERST, H. SEIDL. *XML type checking with macro tree transducers*, in "24th ACM Symposium on Principles of Database Systems", 2005, p. 283–294.

[51] C. MANNING, H. SCHÜTZE. *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, 1999.

[52] W. MARTENS, F. NEVEN. *Typechecking Top-Down Uniform Unranked Tree Transducers*, in "9th International Conference on Database Theory", London, UK, Lecture Notes in Computer Science, Springer Verlag, 2003, vol. 2572, p. 64–78.

[53] J. ONCINA, P. GARCIA, E. VIDAL. *Learning Subsequential Transducers for Pattern Recognition and Interpretation Tasks*, in "IEEE Trans. Patt. Anal. and Mach. Intell.", 1993, vol. 15, p. 448-458.

[54] C. SUTTON, A. MCCALLUM. *An Introduction to Conditional Random Fields for Relational Learning*, in "Introduction to Statistical Relational Learning", MIT Press, 2006.

[55] B. TASKAR, V. CHATALBASHEV, D. KOLLER, C. GUESTRIN. *Learning Structured Prediction Models: A Large Margin Approach*, in "Proceedings of the Twenty Second International Conference on Machine Learning (ICML'05)", 2005, p. 896 – 903.

[56] I. TSOCHANTARIDIS, T. JOACHIMS, T. HOFMANN, Y. ALTUN. *Large Margin Methods for Structured and Interdependent Output Variables*, in "Journal of Machine Learning Research", 2005, vol. 6, p. 1453–1484.

[57] S. VANSUMMEREN. *Deciding Well-Definedness of XQuery Fragments*, in "Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems", 2005, p. 37–48.