INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team Mostrare

Modeling Tree Structures, Machine Learning, and Information Extraction

Lille - Nord Europe

Theme : Knowledge and Data Representation and Management

Table of contents

MOSTRARE *is a joint project with the* LIFL - UMR CNRS 8022*, Lille 1 and Lille 3 universities*

1. Team

Research Scientists

Joachim Niehren [senior researcher (DR2), vice leader, HdR]
Gemma Garriga [junior researcher (CR1) since October 2010]

Faculty Members

Rémi Gilleron [professor, Team leader, HdR]
Iovka Boneva [assistant professor]
Anne-Cécile Caron [assistant professor]
Aurélien Lemay [assistant professor]
Yves Roos [assistant professor]
Sophie Tison [professor, HdR]
Marc Tommasi [professor, HdR]
Fabien Torre [assistant professor]
Sławek Staworko [assistant professor]

Technical Staff

Denis Debarbieux [INRIA, from December 2010 to November 2012]

PhD Students

Benoît Groz [AMN fellowship, since September 2008]
Édouard Gilbert [AMN fellowship, since November 2007]
Grégoire Laurence [MESR, since October 2008]
Jean-Baptiste Faddoul [CIFRE XEROX, since December 2008]
Jean Decoster [MESR, since October 2009]
Antoine M. Ndione [INRIA fellowship, since October 2010]
Tom Sebastian [CIFRE INNOVIMAX, since December 2010]

Post-Doctoral Fellows

Jérôme Champavère [ATER from October 2009 to August 2011]
Gillaume Bagan [INRIA, postdoc from September 2009 to August 2011]
Camille Vacher [ATER since September 2010]
Shunichi Amano [INRIA, postdoc from October 2010 to March 2011]

Administrative Assistant

Karine Lewandowski [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.

2.2. Highlights

- Lemay and Niehren [18] present at ACM PODS the first learning algorithm for top-down XML transformations. This is a breakthrough result on transducer learning since supporting copying and flipping of subtrees for the first time, while all previous proposals were either restricted to words or to relabelings on trees.
- A paper by Tison and Dauchet [29] received the 2010 IEEE LICS Test-of-Time Award. They settled innovative techniques on tree automata, for showing that the first-order theory of one-step rewriting is decidable.

3. Scientific Foundations

3.1. Modeling XML document transformations

Participants: Guillaume Bagan, Iovka Boneva, Anne-Cècile Caron, Benoît Groz, Joachim Niehren, Yves Roos, Sławek Staworko, Sophie Tison, Antoine M. Ndione, Camille Vacher, Shunichi Amano, 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 [36], Xtatic [34], [39], and CDuce [24], [25], [26]. 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 [38], [47]. 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 [42], [40].

The automata community usually approaches tree transformations by tree transducers [32], i.e., tree automata producing output structure. Macro tree transducers, for instance, have been proposed recently for defining XML transformations [40]. From the view point of logic, tree transducers have been studied for MSO definability [33].

3.2. Machine learning for XML document transformations

Participants: Jérôme Champavère, Jean Decoster, Jean-Baptiste Faddoul, Édouard Gilbert, Rémi Gilleron, Grégoire Laurence, Aurélien Lemay, Joachim Niehren, Sławek Staworko, Marc Tommasi, Fabien Torre, Gemma Garriga.

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 [27], [43]. 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 [35], [37]. 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 [31].

Probabilistic context free grammars (pCFGs) [41] are used in the context of PDF to XML conversion [28]. 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 [44]). 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 [46], [45]. 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 [30] 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. VOLATA

Participant: Fabien Torre [correspondent].

VOLATA *V*Otes of Least generAl generalizaTions in jAva

VOLATA is a bundle software containing several learning algorithms: learning algorithms for attribute-value datasets, grammatical inference algorithms and inductive logic programming algorithms. VOLATA has been applied to document classification tasks and information extraction tasks. The software is available at http://www.grappa.univ-lille3.fr/~torre/Recherche/Softwares/volata/.

5.2. PICCATA

Participants: Édouard Gilbert [correspondent], Feriel Lahlali, Marc Tommasi.

PICCATA: *Programming Interface for effiCient Computations and Approximation on multiplicity Tree Automata*.

Piccata is a programming interface for learning weighted and classical tree automata from examples. Piccata development started in 2009 with the former member Feriel Lahlali. Source code under Cecill licence is available on http://piccata.gforge.inria.fr.

5.3. EVOXS

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

EVOXS: *Earliest evaluation of XPath on streams*

This is an implementation of XPath query answering algorithms on `XML` streams following the algorithms developed in the PhD thesis of O. Gauwin in 2009 directed by J. Niehren and S. Tison. It consists of a compiler of a fragment of XPath to deterministic streaming tree automata and an earliest query answering algorithm for queries defined by deterministic streaming tree automata. The main developers of the first version are O. Gauwin and J. Niehren.

EVOXS will be the starting point for our cooperation with INNOVIMAX S.A.R.L in Paris, within the QuiXProc transfer project powered by D. Debarbieux, and the CIFRE PhD thesis of T. Sebastian, both starting in December 2010. The source code is available under the GNU public licence at https://gforge.inria.fr/projects/evoxs.

6. New Results

6.1. Modeling XML document transformations

Participants: Joachim Niehren, Sophie Tison, Sławek Staworko, Aurélien Lemay, Anne-Cécile Caron, Yves Roos, Shunichi Amano, Camille Vacher, Benoît Groz, Antoine M. Ndione, Tom Sebastian.

Query answering and control access. Groz, Boneva, Roos, Tison, Caron and Staworko [8] study the problem of update translation for views on XML documents. More precisely, given an XML view definition and a user defined view update program, the problem is to find a source update program that translates the view update without side effects on the view. Both views and update programs are represented by recognizable tree languages and different settings for the update problem are studied. The results of this line of research studied have been studied during 2010 and will be published in the 14th International Conference on Database Theory (ICDT'2011).

Answer enumeration. Bagan et. al [16] investigate efficient enumeration algorithms for conjunctive queries for databases over binary relations that satisfy the X-underbar property. In particular, their algorithm is able to enumerate answers of XPath queries with variables with linear delay and quadratic precomputation time, both in the size of the database.

Bagan and Niehren study in [21] a more efficient answer enumeration algorithm for a fragment of conditional XPath with variables, which is a first-order complete query language for unranked trees of bounded depth. Their algorithm requires only linear pre-computation time and constant delay for fixed queries, while depending linearly on the size of the query. It is based on a new enumeration algorithm for disjunctions of acyclic conjunctive queries on so called X-doublebar-structures they introduce, which are more restrictive previously known X-underbar structures.

Query answering on XML streams. Niehren started a transfer project QuiXProc with the industrial partner INNOVIMAX S.A.R.L. in December 2010, in which they plan to integrate the XPath query answering algorithms as developed by Mostrare [22] into the XML coordination language XProc of the W3C. In this line of research, Niehren and Tison (with their previous PhD student O. Gauwin) showed in [13] how to distinguish node selection queries on XML streams with bounded delay and concurrency. This is a journal version of a previous LATA 2009 conference paper.

Tree Automata. Tison et. al. revisit in the invited paper [12] their results on tree automaton with global equality and disequality constraints (TAGEDs), previously published in 2008 at the International Conference on Developments in Language Theory (DLT).

Vacher is starting his postdoctoral studies on tree automata with constraints under supervision of Niehren and Tison.

M. Ndione is starting his PhD thesis on probabilistic algorithms for tree automata and transducers under supervision of Lemay and Niehren.

Logic. Amano started his postdoc on XML data exchange under supervision of Niehren.

Staworko et. al. [15] introduce the framework of consistent query answers and repairs in order to alleviate the impact of inconsistency data on answers to a query. In particular, a repair is a minimally different consistent instance and an answer is consistent if it is present in every repair.

Programming languages Niehren at. al. [14] present a journal version of the attributed pi-calculus, a modeling language for systems biology. They add priorities compared to the CMSB 2008 conference version, while elaborating the analogy of priorities and stochastic rates in the pi-calculus.

6.2. Machine learning for XML document transformations

Participants: Gemma Garriga, Rémi Gilleron, Aurélien Lemay, Joachim Niehren, Sławek Staworko, Marc Tommasi, Fabien Torre, Jérôme Champavère, Jean Decoster, Jean-Baptiste Faddoul, Édouart Gilbert, Grégoire Laurence.

Learning tree transformations: transducer induction. Lemay, Niehren et. al. [18] present at ACM PODS the first learning algorithm for top-down XML transformations which allow to restructure trees by copying, flipping, and deleting of subtrees. This is a breakthrough result on transducer learning. Previous proposals were either restricted to transducers on words or to relabelings on trees. The results are obtained as a combination of a new top-down encoding of unranked into ranked trees guided by DTDs and a new learning algorithm for DTOPs. This learning result for DTOPs is derived from a new Myhill-Nerode theorem for DTOPs that the

paper establishes. This theorem also shows the existence of unique minimal DTOPs that are constistent with a DTD. An alternative minimization algorithm for DTOPs was obtained previously, but without the Myhill-Nerode theorem and any link to learning. This result was obtained in cooperation within our associated team TRANSDUCE with NICTA Sydney created in 2010, but actually started already in 2007.

Learning queries or schemas: automata induction. Champavère [11] defended his PhD thesis in September 2010 under the supervision of Niehren, Lemay and Gilleron. The thesis studies the incorporation of schema knowledge in XML query induction from annotated example trees. The results lift tree automata based induction algorithms for total monadic node selecting queries to partial queries whose domain is fixed by a known schema. Schema consistency is checked dynamically by testing language inclusion for tree automata. Another contribution of the thesis is to study query induction from pruned annotated examples where subtrees irrelevant for node selection may be cut away; in this line of research, the thesis presents a new learnability result for classes of queries that are stable under schema-guided pruning strategies.

Sequence classification Torre et. al. present in [19] a general framework for supervised classification; in particular, their goal is the classification of sequences by deciding whether a word belongs in some language or not. They integrate classical grammatical inference techniques into a general framework resulting from supervised classification: the hypotheses are therefore represented by automata or balls of strings, that are then combined by traditional boosting algorithms.

Induction of stochastic tree automata. Gilbert, Gilleron and Tommasi study in [23] the inference of probability distributions over sets of unranked trees (i.e. trees where a node can have an unbounded number of direct subtrees). The main objective here is to build probabilistic decision procedures able to classify (XML/HTML) trees from different sources, or to get concise representations of sets of trees. The problem is formalized as the more general problem of learning tree series. For the question of defining recognizable tree series of unranked trees, the authors specify weighted automata for unranked trees via the extension of previous hedge automata and weighted trees. The paper also considers binary representations of unranked trees and shows that recognizable tree series for unranked trees can be defined and studied from recognizable tree series of those binary representations. The paper also presents decidability results on probabilistic tree automata and algorithms for computing sums of convergent series.

Multitask learning. Faddoul, Torre and Gilleron with their partner from XEROX Grenoble study in [17] the problem of learning multiple related tasks from data simultaneously. For this problem, the authors propose a novel learning algorithm, called MT-Adaboos, which extends the traditional Adaboost algorithm to multitask setting by using simple decision stumps as weak classifiers. The practical and theoretical results of the paper show that the new algorithm learns the dependencies between tasks for different regions of the learning space.

Conditional random fields. Tommasi participated in the writing of a chapter of a book on conditional Markov fields for information extraction [20]. In this french book chapter, the authors review some machine learning methods for information extraction and we focus on Conditional Random Fields (CRF), for which some prototypes were developed last year. The book chapter also gives the connexions with with Hidden Markov Models and logistic regression.

Learning and mining in graphs Garriga was hired as researcher (CR1). She started a research project on learning and mining data and data streams in networks and she leaded a new working group inside Mostrare.

7. Contracts and Grants with Industry

7.1. Contracts and Grants with Industry

7.1.1. 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).

7.1.2. Cifre Innovimax (2010-2013)

Participants: Tom Sebastian, Joachim Niehren [correspondent].

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

7.1.3. QuiXProc: INRIA Transfer Project with Innovimax (2010-2012)

Participants: Denis Debarbieux, Joachim Niehren [correspondent].

Niehren and Debarbieux started 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.

8. Other Grants and Activities

8.1. National Actions

8.1.1. ANR Lampada (2009-2013)

Participants: Marc Tommasi [correspondent], Édouard Gilbert, 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/.

Lampada is a fundamental research project on machine learning and structured data. It focuses on scaling learning algorithms to handle large sets of complex data. The main challenges are 1) high dimension learning problems, 2) large sets of data and 3) dynamics of data. Complex data we consider are evolving and composed of parts in some relations. Representations of these data embed both structure and content information and are typically large sequences, trees and graphs. The main application domains are web2, social networks and biological data.

The project proposes to study formal representations of such data together with incremental or sequential machine learning methods and similarity learning methods.

The representation research topic includes condensed data representation, sampling, prototype selection and representation of streams of data. Machine learning methods include edit distance learning, reinforcement learning and incremental methods, density estimation of structured data and learning on streams.

8.1.2. 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/.

The Codex project seeks to push the frontier of XML technology in three interconnected directions. First, efficient algorithms and prototypes for massively distributed XML repositories are studied. Second, models are developed for describing, controlling, and reacting to the dynamic behavior of XML collections and XML schemas with time. Third, methods and prototypes are developed for composing XML programs for richer interactions, and XML schemas into rich, expressive, yet formally grounded type descriptions.

The main contributions of Mostrare are results on learning top-down XML transformations [18], on XPath query answering algorithms on XML streams [13], and on XML query learning [11]. In addition the Codex project has lead to the creation of an INRIA transfer project QuiXProc with Innovimax and of a CIFRE project with Innovimax.

8.1.3. ANR Blanc Enum (2007-2011)

Participants: Guillaume Bagan, Joachim Niehren [correspondent], Sophie Tison.

The Enum project on "Complexity and Algorithms for Answer Enumeration", is coordinated by A. Durand (Paris VII). The other partners are E. Grandjean (University of Caen), N. Creignou (University of Marseille). Public information on Enum can be found on http://enumeration.gforge.inria.fr.

Enum studies algorithmic and complexity questions of answers enumeration, the task of generating all solutions of a given problem. Answer enumeration requires innovative efficient algorithms that can quickly serve large numbers of answers on demand. The prime application is query answering in databases, where huge answer sets arise naturally.

Mostrare contributed in 2010 to new answer enumeration algorithms for XPath queries [16], [21].

8.1.4. ARC ACCESS (2010–2011)

Participants: Iovka Boneva [correspondent], Sophie Tison, Anne-Cécile Caron, Yves Roos, Benoît Groz, Sławek Staworko.

This is a collaboration on the subject Access Control Policies for XML: Verification, Enforcement and Collaborative Edition, supported by the INRIA Collaboration Program (Action de Recherche Collaborative). The other participants involved are from the INRIA teams DAHU (INRIA Saclay – Île de France), PAREO and CASSIS (INIRA Nancy – Grand Est). This project is concerned with the security and access control for Web data exchange, in the context of Web applications and Web services. We aim at defining automatic verification methods for checking properties of access control policies (ACP) for XML, like consistency or secrecy, and for the comparison ACPs. One of our goals is to apply formal tools from tree automata theory for this purpose. A second important goal is to design efficient methods for ACP enforcement for secure query evaluation. We will study several scenarios for solving different variants of this problem, based on the notion of secure user views. As a case study, we will apply our methods to an XML-based collaborative editing system.

8.2. International Cooperations

8.2.1. Transduce: INRIA Associated Team with NICTA Sydney.

Participants: Guillaume Bagan, Joachim Niehren [correspondent], Aurélien Lemay, Benoît Groz, Slawomir Staworko, Grégoire Laurence.

The leader of the our NICTA partner team is S. Maneth. Public information on Enum can be found on http://transduce.gforge.inria.fr/.

We keep cooperation on learning algorithms for XML to XML transformations and on XML query answering algorithms. The main result in 2010 was a learning algorithm for top-down XML transformations [18].

9. Dissemination

9.1. Animation of the scientific community

9.1.1. Program Committees

R. GILLERON was member of the program committee of CAP 2010 (Conférence Francophone sur l'Apprentissage Automatique).

J. NIEHREN is member of the steering committee of RTA (International Conference on Rewriting Techniques and Applications), of the editorial board of FUNDAMENTA INFORMATICAE. In 2010 he was member of the program committees of LPAR 2010 (International Conference on Logic for Programming, Artificial Intelligence and Reasoning) and ATANLP 2010 (ACL 2010 Workshop on Applications of Tree Automata in Natural Language Processing).

S. TISON was member of the program committee of RTA 2010 (the 21st International Conference on Rewriting Techniques and Applications) and STACS 2011 (Annual Symposium on Theoretical Aspects of Computer Science). She is member of of the editorial board of RAIRO - ITA and of the steering committee of STACS.

M. TOMMASI was member of the program committee of ECML 2010 (European conference on Machine Learning), ATANLP 2010 (ACL 2010 Workshop on Applications of Tree Automata in Natural Language Processing) and LATA 2010 (the 4th International Conference on Language and Automata Theory and Applications).

F. TORRE was member of the program committee of ECML 2010 (European conference on Machine Learning) and CAP 2010 (The Conférence Francophone sur l'Apprentissage Automatique).

9.1.2. French Scientific Responsibilities

A.C. CARON is member of the french national evaluation committee for computer science assistant professors (CNU 27). She was member of selection committee for assistant professor in Lille.

R. GILLERON was member of the scientific committee of the program Programme blanc SIMI2, ANR. He participated in the AERES evaluation committee of the computer science lab IRISA (Rennes). He was member of the selection committee in Nantes for a professor position, of the selection committee in Paris 6 for assistant professors, and of the selection committee in Lille for assistant professor.

S. TISON was co-head of the scientific committee of the program Programme blanc SIMI3, ANR. She is head of the computer science lab in Lille (LIFL). She chairs the scientific council of "Pôle de Compétitivité industries du Commerce". She was member of the national PES commission 27. She was invited member of the scientific council of ST2I (CNRS) until september 2010.

J. NIEHREN was president of the selection committee for postdocs and PhD students of the research center INRIA Lille Nord Europe, and member of the selection committee for 1 professor position at Ecole Polytechnique Lille.

M. TOMMASI was member of the Technological Development Committee of Inria Lille and of the scientific committee for selection of assitant professors at Rennes IFSIC and University of Marseille.

9.1.3. Miscellaneous

M.TOMMASI gave an invited talk on Conditional Random fields at the workshop ATALA (Association pour le traitement automatique des Langues).

9.2. Teaching

9.2.1. Teaching hours

Iovka BONEVA	192 hours	bachelor
Anne-Cécile CARON	192 hours	bachelor and masters
Jérôme CHAMPAVÈRE	96 hours	masters
Rémi GILLERON	140 hours	bachelor and masters
Aurélien LEMAY	49 hours	masters
Yves ROOS	192 hours	bachelor andf masters
Sławek STAWORKO	192 hours	bachelor and masters
Marc TOMMASI	192 hours	masters
Sophie TISON	96 hours	masters
Camille VACHER	90 hours	masters
Benoît GROZ	64 hours	masters

9.2.2. Master lectures at the University of Lille

- Affaire et Negociation Internationale: Base de Données, by A. LEMAY
- Traduction Spécialisé Multilingue: Création de Site Web, by A. LEMAY
- XML, by M. TOMMASI
- Networks, by M. TOMMASI
- Automatisation du traitement de l'information, by M. TOMMASI
- Supervised classification, by R. GILLERON
- Unsupervised classification, by R. GILLERON
- Information retrieval, by R. GILLERON
- Advanced algorithms and complexity, by B. GROZ, S. TISON
- Content Management Systems, by J. CHAMPAVÈRE
- Semantic Web, by J. CHAMPAVÈRE
- Web Programming by J. CHAMPAVÈRE
- Advanced databases, by A-C. CARON
- Semantic Web, by A-C. CARON

9.2.3. Master projects and internships

- Radu Ciucanu from IASI in Romania started an internship on implementing enumeration algorithms for XPath dialects with variables in Scala. Directed by G. BAGAN and J.NIEHREN.
- Surbi Maheshwari from from IIT Guwahati India started an internship on implementing learning algorithms for n-ary queries. Directed by A. LEMAY and J. NIEHREN.

9.2.4. PhD theses

- E. GILBERT, Learning weighted tree automata for information extraction from XML. Supervised by Tommasi and Gilleron
- J. CHAMPAVÈRE, Schema-guided query induction for information extraction. PhD defended in September 2010. Supervised by Niehren, Gilleron, and Lemay.
- G. LAURENCE, Learning XML transformations for data exchange on the web. Supervised by Tommasi, Niehren, Staworko and Lemay.
- B. GROZ, XML database security and access control. Supervised by Tison and Staworko.
- J.-B. FADDOUL, Machine learning and applications to social network analysis. Supervised by Gilleron and Chidlowskii from XEROX European Research Center (XRCE).

- J. DECOSTER, Statistical relational learning of XML transformations. Supervised by Tommasi and Torre.
- A. M. NDIONE, Probabilistic algorithms for tree automata and transducers. Supervised by Niehren and Lemay.
- T. SEBASTIAN, Streaming algorithms for XSLT. Supervised by Niehren.

9.2.5. PhD committees

- A. LEMAY belonged to the PhD committee of J.Champavère (Lille 1).
- J. NIEHREN reviewed the PhD thesis of M. John (Unversity of Rostock, Germany) and was member of the PhD committee of J.Champavère (Lille 1).
- M. TOMMASI belonged to the PhD committee of J. Champavère (Lille 1) and Nataliya Sokolovska (Telecom ParisTech).
- R. GILLERON was a reviewer of the PhD of A. Zidouni (Marseille).
- S. TISON belonged to the PhD committees of H. Sharif (Lille 1), H. Idabal (CRI Paris 1 Panthéon Sorbonne University), and C. Vacher (LSV, ENS Cachan).

9.2.6. Habilitation committees

- R. GILLERON participated in the habilitation committee of L. Ralaivola (Marseille, as reviewer), J.C. Janodet (Saint-Etienne, as reviewer) and A. Habrard (Marseille, as member).
- S. TISON belonged to the habilitation committee of F. Clautiaux (Lille 1).

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] 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.

[3] 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/.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] B. GROZ, I. BONEVA, Y. ROOS, S. TISON, A.-C. CARON, S. STAWORKO. *View update translation for XML*, in "ICDT 13th International Conference on Database Theory", 2011.

[9] 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.

[10] 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.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] J. CHAMPAVÈRE. *Induction de requêtes guidée par schéma*, Université des Sciences et Technologie de Lille - Lille I, Sep 2010, http://hal.inria.fr/tel-00517358.

Articles in International Peer-Reviewed Journal

[12] 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.

[13] 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.

[14] M. JOHN, C. LHOUSSAINE, J. NIEHREN, A. UHRMACHER. *The Attributed Pi Calculus with Priorities*, in "Transactions on Computational Systems Biology", Feb 2010, vol. XII, n^{o} 5945, p. 13-76, http://hal.inria.fr/inria-00422969.

[15] S. STAWORKO, J. CHOMICKI. *Consistent Query Answers in the Presence of Universal Constraints*, in "Information Systems", 2010, vol. 35, n^{o} 1, p. 1-22, http://hal.inria.fr/inria-00489298.

International Peer-Reviewed Conference/Proceedings

[16] 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.

[17] 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.

[18] 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.

[19] F. TANTINI, A. TERLUTTE, F. TORRE. *Sequences Classification by Least General Generalisations*, in "10th International Colloquium on Grammatical Inference", Espagne Valencia, Springer, Sep 2010, vol. 6339, p. 189-202, http://www.springerlink.com, http://hal.inria.fr/inria-00524707.

Scientific Books (or Scientific Book chapters)

[20] I. TELLIER, M. TOMMASI. *Champs Markoviens Conditionnels pour l'extraction d'information*, in "Modèles probabilistes pour l'accès à l'information textuelle", E. GAUSSIER, F. YVON (editors), Hermès, 2010, http://hal.inria.fr/inria-00514525.

Research Reports

[21] G. BAGAN, J. NIEHREN. *Efficient Answer Enumeration for XPath Dialects with Variables*, INRIA, Nov 2010, http://hal.inria.fr/inria-00533757.

[22] O. GAUWIN, J. NIEHREN. *Streamable Fragments of Forward XPath*, INRIA, 2010, http://hal.inria.fr/inria-00442250.

[23] É. GILBERT, R. GILLERON, M. TOMMASI. *Series, Weighted Automata, Probabilistic Automata and Probability Distributions for Unranked Trees.*, INRIA, Feb 2010, RR-7200, http://hal.inria.fr/inria-00455955.

References in notes

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] M. DAUCHET, S. TISON. *The theory of ground rewrite systems is decidable*, in "Logic in Computer Science, 1990. LICS '90, Proceedings., Fifth Annual IEEE Symposium on e", 1990, p. 242 -248.

[30] 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.

[31] 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.

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

[33] 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.

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

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

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

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.