IN PARTNERSHIP WITH:
CNRS

Université Charles de Gaulle (Lille 3)

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2011

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

1. Members

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]
Mikaela Keller [assistant professor since January 2011]
Angela Bonifati [professor since September 2011]

Technical Staff

Denis Debarbieux [INRIA, since December 2010]

PhD Students

Benoît Groz [AMN fellowship, since September 2008]
Grégoire Laurence [MESR, since October 2008]
Jean-Baptiste Faddoul [CIFRE XEROX, since December 2008]
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]

Post-Doctoral Fellows

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

Administrative Assistant

Karine Lewandowski [shared by 3 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 [38], Xtatic [36], [41], and CDuce [27], [28], [29]. 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 [40], [49]. 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 [44], [42].

The automata community usually approaches tree transformations by tree transducers [34], i.e., tree automata producing output structure. Macro tree transducers, for instance, have been proposed recently for defining XML transformations [42]. From the view point of logic, tree transducers have been studied for MSO definability [35].

3.2. Machine learning for XML document transformations

Participants: Jean Decoster, 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 [30], [45]. 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 [37], [39]. 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 [33].

Probabilistic context free grammars (pCFGs) [43] are used in the context of PDF to XML conversion [31]. 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 [46]). 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 [48], [47]. 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 [32] 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 a compiler of the FXP library to nested word automata, efficient query answering algorithm for nested word automata on XML streams, and thus for FXP queries.
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.

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

- Version: 0-9-2011-03-25

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 Forward XPath with full support for the XML data model. The QuiXPath library provides a compiler from QuiXPath to FXP. Thereby, the efficient query answering algorithms for FXP are lifted to a fragment of Forword XPath. 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 owne by Innovimax.

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

- Version: QuixPath v1.0.0

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

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 Ndione, Tom Sebastian.

Query answering on XML streams. In [16], Gauwin and Niehren introduce the notion of finite streamability for query languages, and classify fragments of XPath that are finitely streamable or not. They show that if a query language is finitely streamable, then its satisfiability problem can be solved in polynomial time, which in turn is known to fail for mostly all fragments of XPath. They also show that FXP, the fragment of ForwardXPath with child and descendant axis, conjunction, and negation becomes finitely streamable if bounding the number of conjunctions. Since 3 conjunctions are enough in many practical applications, FXP is most relevant in practice. Without any bound, FXP is not finitely streamable, since its satisfiablity problem is DEXPTIME hard. The positive result for FXP with a bounded number of conjunctions is obtained by compilation of FXP to deterministic nested word automata. The compiler is in exponential time in the number of conjunctions, and thus polynomial if this parameter is bounded.

Answer enumeration for n-ary queries. Bagan, Filiot, Gauwin, and Niehren investigated answer enumeration algorithms for dialects of XPath with variables. The problem with n-ary queries is that answer sets may grow exponentially in $|t|^n$, so that algorithms depending polynomially on the size of the answer set might still be unfeasible. In such case, it might still be possible to enumerate elements of answer sets on need. The questions is then whether enumeration can be done efficiently without duplicates and failures, that is with constant delay between subsequent answers and polynomial time preprocessing in the size of the query and the tree. We obtained positive results on answer enumeration with constant delay enumeration for acyclic conjunctive queries over so called X-doublebar structures that we introduce [24]. These subsume tree structures with child, next-sibling and next-sibling* axis, but not the descendant axis. Our result can be lifted to a dialect of ConditionalXPath with variables, that is FO-complete on trees of bounded depth, so that the descendant axis is not needed.

Tree automata global constraints. TAGEDs are a new class of tree automata with constraints that currently receive much interest from top conferences on theoretical computer science. During its postdoc in Lille, Vacher improved complexity bounds for some fragments of decidability results [12].

Sequential tree-to-word transducers. Laurence, Lemay, Niehren, Staworko, Tommasi considered deterministic sequential top-down tree-to-word transducers (STWs), that capture the class of deterministic top-down nested-word to word transducers. While reordering and copying are not allowed, STWs are nevertheless very expressive because they allow concatenation of outputs, deletion of inner nodes and they can produce context free languages as output. Their expressiveness is incomparable with DTOPs (plus concatenation, but minus copying). While objecting for learning algorithms, they study normalization of STWs in a first step and then develop unique minimalization algorithms for normalized STWs in a second step in [19]. The idea of normalization is to produce the output in an earliest manner, when reading the input in document order. This works only on binary trees, but can be lifted to unranked trees modulo the binary top-down encoding. The normalization algorithm is by far nontrivial. The natural continuation of this approach will be toward learning algorithms for earliest STWs.

Access control for XML views. The PhD project of Groz, supervised by Staworko and Tison, is centered on access control for XML databases, and in particular on security of user views over XML documents. He obtained results on query rewriting for read-only queries, and translation for update queries. More precisely, given an XML view definition and a user defined query (resp. update program) q, the problem is to find a source query (resp. update program) that is equivalent to q on the view. Caron, Groz, Roos, Staworko and Tison study update programs and views represented by recognizable tree languages in [15], and devise algorithms for update translation in different settings, namely without or with constraints on the authorised source updates.

6.2. Machine learning for XML document transformations

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

Induction of tree automata. Champavère, Gilleron, Lemay and Niehren proposed to use schemas for improving induction algorithms for monadic queries represented by tree automata [26]. The idea is to use pruning strategies to eliminate useless parts of trees when learning from partially annotated trees such that only the structure of relevant fragments is learned. This allows to avoid generalization errors and to learn from fewer annotations. They define schema-guided pruning strategies. They define stable queries w.r.t. a pruning strategy and show that stable queries are learnable.

Further Results. In [21], Staworko proposed learning twig and path queries. Prioritized repairing and consistent query answering in relational databases was tackled in [13] and Bounded repairability for regular tree languages in [20].

Torre and Terlutte explored the combination of automata and words balls for sequences classification in [14].

Tommasi participated in the writing of a chapter of a book on conditional Markov fields for information extraction [23].

Garriga and collaborators from the Fraunhofer Institute in Bonn, studied fixed parameter tractable algorithms for the discovery of maximal order preserving submatrices in bioinformatic applications in [17].

We begun a new activity on *learning for social network and information network* supported by the arrivals of Gemma GARRIGA, Mikaela KELLER and Antonino FRENO. Freno, Garriga and Keller [22] proposed a model for predicting new links in a network which exploit both the current structure of the network and the content of its node.

7. Contracts and Grants with Industry

7.1. Contracts 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 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.

7.1.3. Cifre SAP (2011-2014)

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

Garriga and Gilleron continue supervising the PhD thesis (Cifre) of Thomas Ricatte together with TBA from SAP.

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

Participants: Denis Debarbieux, Joachim Niehren [correspondent].

Niehren and Debarbieux continue 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. Partnerships and Cooperations

8.1. National Actions

8.1.1. ANR Lampada (2009-2013)

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/.

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.

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.

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 Initiatives

8.2.1. INRIA Associate Teams

8.2.1.1. TRANSDUCE

Title: Automatic XML Data Conversion through Tree Transducers

INRIA principal investigator: Joachim Niehren

International Partner:

Institution: NICTA Sydney (Australia)

Laboratory: XML Query Technologies project

Duration: 2010 - 2012

See also: https://gforge.inria.fr/plugins/wiki/index.php?id=390&type=g

Data conversion between XML formats is a frequent, complex, and repetitive engineering task. It needs to be solved for data publishing, peer-to-peer data exchange, document processing, information extraction, and Web services. In this project, we propose to develop automatic methods generating conversion programs from examples, so that they can be used by non-expert users. Our approach is based on learning of tree transducers and XML queries.

8.2.2. Visits of International Scientists

Sebastian MANETH from NICTA Sydney visited us May-July 2011. He now moved to the University of Leipzig. In 2010-11, this cooperation was funded by the INRIA associate team program.

8.2.3. Participation In International Programs

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

9. Dissemination

9.1. Animation of the scientific community

9.1.1. Invited Talks

S. TISON was invited to RTA 2011 and TLCA 2011 to give talk on “Tree Automata, (Dis-)Equality Constraints and Term Rewriting: What’s New?”

9.1.2. Program Committees

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 CIKM 2011 (International Conference on Information and Knowledge Management), RTA 2011 (International Conference on Rewriting Techniques and Applications), LATA 2011 (International Conference on Language and Automata Theory and Applications), APWeb 2011 (Asian Pacific Web Conference), and LID 2011 (Logic in Databases).

S. TISON was member of the program committee of 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.

S. STAWORKO was member of the program committee of SUM 2011 (International Conference on Scalable Uncertainty Management) and co-chair of LID 2011 (International Workshop on Logic in Databases).

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 KDD 2011 (ACM SigKDD Conference on Knowledge Discovery and Data Mining), ICML 2011 (International Conference on Machine Learning), ICDM 2011 (IEEE International Conference on Data Mining), SDM 2011 (SIAM International Conference on Data Mining) and IDA 2011 (Intelligent Data Analysis conference).

A. BONIFATI was Chair for the program committee for the Semistructured Data, XML and Web Data Management track, ICDE 2011. She was also member of the program committees of SIGMOD 2011 (International Conference on Management of Data), VLDB 2011 (International Conference on Very Large Databases), EDBT 2011 (International Conference on Extending Data- base Technology).

9.1.3. 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 a member of the scientific committee of the program Programme blanc SIMI2, ANR and a member of the selection committees of assistant professors in Marseille and Lille and finally a member of the evaluation committees of the habilitation thesis of O. Teytaud (LRI, Paris 11) and R. Bailly (LIF, Marseille)

S. TISON is head of the computer science lab in Lille (LIFL). She was co-head of the scientific committee of the program Programme blanc SIMI2, ANR. She is member of the scientific committee of the program Chaires Industrielles, ANR. She chaired the AERES evaluation committee of the computer science lab I3S (Nice/Sophia). She chairs the scientific council of "Pôle de Compétitivité Industries du Commerce". She was member of the national PES commission 27. She was member of the selection committee in Lille for professor positions.

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 a member of the Technological Development Committee (CDT), a member of the Computers Users Committee (CUMI) of Inria Lille and a member of the scientific committee for selection of assitant professors in St Etienne. Expertises/rapport JEI/CIR

9.2. 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 G. GARRIGA

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

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

Licence: Propositional Logic, 18h, L2, Université Lille 3, France by S. STAWORKO

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

PhD & HdR

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

PhD in Progress: 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 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, TBA. 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.

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] 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

Articles in International Peer-Reviewed Journal

[11] O. GAUWIN, J. NIEHREN, S. TISON. *Queries on XML Streams with Bounded Delay and Concurrency*, in "Information and Computation", March 2011, vol. 209, n^{o} 3, p. 409-442 [*DOI:* 10.1016/J.IC.2010.08.003], http://hal.inria.fr/inria-00491495/en.

[12] F. JACQUEMARD, F. KLAY, C. VACHER. *Rigid Tree Automata and Applications*, in "Information and Computation", February 2011, vol. 209, n^{o} 3, p. 486-512 [*DOI :* 10.1016/J.IC.2010.11.015], http://hal.inria.fr/inria-00578820/en.

[13] 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/en.

Articles in National Peer-Reviewed Journal

[14] F. TANTINI, A. TERLUTTE, F. TORRE. *Combinaisons d'automates et de boules de mots pour la classification de séquences*, in "Revue d Intelligence Artificielle", June 2011, vol. 25, n^{o} 3, p. 411-434 [*DOI :* 10.3166/RIA.25.411-434], http://hal.inria.fr/hal-00643057/en.

International Conferences with Proceedings

[15] 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.

[16] O. GAUWIN, J. NIEHREN. *Streamable Fragments of Forward XPath*, in "16th International Conference on Implementation and Application of Automata", Blois, France, April 2011, http://hal.inria.fr/inria-00442250/en.

[17] J. HUMRICH, T. GAERTNER, G. GARRIGA. *A Fixed Parameter Tractable Integer Program for Finding the Maximum Order Preserving Submatrix*, in "Proceedings of the IEEE International Conference of Data Mining, ICDM 2011", Vancouver, Canada, 2011, http://hal.inria.fr/hal-00641896/en/.

[18] M. JOHN, C. LHOUSSAINE, J. NIEHREN, C. VERSARI. *Biochemical Reaction Rules with Constraints*, in "20th European Symposium on Programming Languages", Saarbrücken, Germany, LNCS, Springer, March 2011, vol. 6602, p. 338-357, http://hal.inria.fr/inria-00544387/en.

[19] G. LAURENCE, A. LEMAY, J. NIEHREN, S. STAWORKO, M. TOMMASI. *Normalization of Sequential Top-Down Tree-to-Word Transducers*, in "5th International Conference on Language Automata Theory and Appliciations", Tarragona, Spain, LNCS, Springer, 2011, http://hal.inria.fr/inria-00566291/en.

[20] G. PUPPIS, C. RIVEROS, S. STAWORKO. *Bounded repairability for regular tree languages*, in "International Conference on Database Theory (ICDT)", Berlin, Germany, March 2012, http://hal.inria.fr/hal-00643100/en.

[21] 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/en.

Conferences without Proceedings

[22] A. FRENO, G. GARRIGA, M. KELLER. *Learning to Recommend Links using Graph Structure and Node Content*, in "Neural Information Processing Systems Workshop on Choice Models and Preference Learning", Granada, Spain, 2011, http://hal.inria.fr/hal-00641419/en.

Scientific Books (or Scientific Book chapters)

[23] 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, 2011, http://hal.inria.fr/inria-00514525/en.

Research Reports

[24] G. BAGAN, J. NIEHREN. *Constant Delay Enumeration for Acyclic Conjunctive Queries over X-Doublebar Structures*, inria, 2011, http://hal.inria.fr/inria-00609719/en.

[25] B. GROZ, S. MANETH, S. STAWORKO. *Deterministic Regular Expressions in Linear Time*, inria, September 2011, http://hal.inria.fr/inria-00618451/en.

Other Publications

[26] J. CHAMPAVÈRE, R. GILLERON, A. LEMAY, J. NIEHREN. *Query Induction with Schema-Guided Pruning Strategies*, August 2011, journal submission, http://hal.inria.fr/inria-00607121/en.

References in notes

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

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

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

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

[44] 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.

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.