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

1. Team **1**
2. Overall Objectives **1**
2.1. Presentation 1
2.2. Highlights 2
3. Scientific Foundations **2**
3.1. Modeling XML document transformations 2
3.2. Machine learning for XML document transformations 3
4. Application Domains **3**
5. Software **4**
6. New Results **4**
6.1. Modeling XML document transformations 4
6.1.1. XML database queries, logic and automata 4
6.1.2. Programming languages 5
6.2. Machine learning for XML document transformations 5
6.2.1. Grammatical Inference 5
6.2.2. Statistical Inference 6
7. Contracts and Grants with Industry **6**
7.1.1. RNTL ATASH (2006-2009) 6
7.1.2. RNTL Webcontent (2006-2009) 6
7.1.3. Cifre Xerox (2009-2011) 6
8. Other Grants and Activities **7**
8.1. National Actions 7
8.1.1. ANR Lampada (2009-2013) 7
8.1.2. ANR Defis Codex (2009-2011) 7
8.1.3. ANR Blanc Enum (2007-2010) 7
8.1.4. ARA MDCO Marmota (2006-2009) 8
8.1.5. ARA MDCO CROTAL (2008-2009) 8
8.2. International Cooperations 9
9. Dissemination **9**
9.1. Scientific Animation 9
9.1.1. Program Committees 9
9.1.2. French Scientific Responsibilities 9
9.1.3. Miscellaneous 10
9.2. Teaching and Scientific Diffusion 10
9.2.1. Teaching 10
9.2.2. Master lectures at the University of Lille 10
9.2.3. International Master lectures 10
9.2.4. Master projects 10
9.2.5. PhD theses 11
9.2.6. PhD committees 11
9.2.7. Habilitation committees 11
10. Bibliography **11**

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

1. Team

Research Scientist

Joachim Niehren [senior researcher (DR2), vice leader, HdR]

Faculty Member

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 since September 2009, INRIA from July 2007 to August 2009]

Technical Staff

Feriel Lahlali [INRIA young software engineer from December 2007 to November 2009]

PhD Student

Jérôme Champavère [MESR fellowship, since October 2006]
Olivier Gauwin [INRIA Cordi fellowship, from November 2006 to October 2009]
Benoît Groz [AMN followship, 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]

Post-Doctoral Fellow

Sławek Staworko [INRIA postdoc, from November 2009 to December 2009]
Guillaume Bagan [INRIA, from September 2009 to August 2010]

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 summarise 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

- Gauwin's PhD thesis [11] showed that even small syntactic fragments of XPath queries cannot be answered efficiently on XML streams, in contrast to queries defined by deterministic tree automata. This insight permitted to design new efficient query answering algorithms for XPath fragments with schema restrictions.
- The recruitment of Sławek Staworko permits to pursue recent approaches on tree transformations for XML control access and query adaptation to schema changes.

3. Scientific Foundations

3.1. Modeling XML document transformations

Participants: Guillaume Bagan, Iovka Boneva, Anne-Cécile Caron, Olivier Gauwin, Benoît Groz, Joachim Niehren, Yves Roos, Sławek Staworko, Sophie Tison.

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.

Alternatives programming language for defining XML transformations have been proposed by the programming language community, for instance XDuce [38], Xtatic [36], [41], and CDuce [25], [26], [28]. 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: Jérôme Champavère, Jean Decoster, Jean-Baptiste Faddoul, Édouard Gilbert, Rémi Gilleron, Grégoire Laurence, Aurélien Lemay, Joachim Niehren, Sławek Staworko, Marc Tommasi, Fabien Torre.

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 [29], [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 [30]. 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. 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 managing multiplicity tree automata, i.e. tree automata with weights. The current version focuses on real-valued automata. The model is the one introduced by [27]. Piccata takes advantage of the vector space structure using existing linear algebra library. The library allows to deal with ranked and unranked trees. Piccata is developed in collaboration with colleagues from the LIF in Marseille. The library will also include inference algorithms for weighted tree automata. Variants DEES algorithm [31] are currently implemented. A first version of the software has been released, and the software was tested with real XML data.

6. New Results

6.1. Modeling XML document transformations

6.1.1. XML database queries, logic and automata

Participants: Olivier Gauwin, Joachim Niehren, Sophie Tison, Sławek Staworko, Grégoire Laurence, Aurélien Lemay, Anne-Cécile Caron, Yves Roos, Benoît Groz.

Query answering on XML streams. Gauwin [11] introduces in his PhD thesis a hierarchy of streamability notions for query languages that restrict both space and time. He shows that query languages with hard satisfiability problems are not streamable even for the weakest notion of streamability. This result contradicts prominent streaming algorithms for small syntactic fragments of XPath in the database literature.

With his advisors Niehren and Tison [16], Gauwin shows that earliest query answering is feasible for queries defined by deterministic streaming tree automata. This permits to establish positive streamability results for n-ary queries defined by deterministic nested word automata, under the condition that the concurrency of queries and the depth of trees are bounded. It equally implies streamability results for fragments of XPath with schema restrictions. Gauwin, Niehren and Tison [15] show that bounded concurrency is decidable in polynomial time for queries defined by deterministic tree automata. The proof is by reduction to proving bounded valuedness of recognizable relations between ranked trees, which in turn can be reduced to bounded valuedness of bottom-up tree transducers.

Bagan started his Postdoc studies on efficient answer enumeration for queries in Conditional XPath with variables, in cooperation with Niehren. The topic was prepared by an internship of A. Venant from ENS Cachan-Bretagne supervised by Niehren.

Tree Transformations. The objective of PhD project of Laurence directed by Staworko, Lemay, and Niehren is to learn tree transformations. The novelty of their approach is to base such transformation tasks on deterministic nested word to word transducers [22]. Such transducers are useful for defining XML transformations such as used in XML style sheets. The paper shows that equivalence testing is in polynomial time for deterministic nested word to word transducers. This first result is obtained by reduction to Plandowski's result on polynomial time equivalence checking for morphism on context tree languages. The class of deterministic nested word to word transducers seems promising as starting point for automatic style sheet induction from examples.

Groz, Staworko, Caron, Roos and Tison study the view-based security framework for XML for an expressive class of security access specifications, the expressive RegXPath (regular XPath) query language, and without any restrictions on the DTD [17]. They are interested in query rewriting, allowing to answer user queries on a security view without materialization, and devise a quadratic algorithm for query rewriting. As second contribution they propose three approximations of the user DTD, each of which is indistinguishable from the real schema by a particular class of queries. Finally, they propose two methods for comparing security policies.

6.1.2. Programming languages

Participant: Joachim Niehren.

Niehren continues co-steering the newly created BioComputing project of the LIFL with C. Lhoussaine, who is the official leader. Jointly with L. Uhrmacher from the University of Rostock, they supervised the PhD project of M. John. In [13] they present the attribute pi-calculus with priorities, with generalizes on various pi-calculi used for stochastic simulation in systems biology. In [18], they enrich the attributed pi-calculus further, by an imperative store for global control. This allows to model dynamic compartments with mutable configurations, which are of interest for modeling biochemistry in live cells. An implementation of the imperative pi-calculus demonstrates high accuracy for modeling operations of compartment dissolution and merging, while keeping good efficiency.

Niehren continues his collaboration with M. Schmidt-Schauß and D. Sabel from Frankfurt and J. Schwinghammer from Saarbrücken. In [21], they show the equivalence of two synchronization primitives in Alice ML, an concurrent extension of Standard ML with handled futures and concurrent buffers. The proof uses a semantic-preserving translation, thus illustrating that observational semantics proof techniques allow to provide comprehensive proofs of the equivalence of concurrency primitives in realistic concurrent programming languages.

6.2. Machine learning for XML document transformations

6.2.1. Grammatical Inference

Participants: Jérôme Champavère, Jean Decoster, Rémi Gilleron, Grégoire Laurence, Aurélien Lemay, Joachim Niehren, Sławek Staworko, Marc Tommasi, Fabien Torre.

The PhD thesis of J. Champavère on schema guided query induction, directed by Niehren, Lemay, and Gilleron, will be submitted beginning of 2010. It presents query learning algorithms based on grammatical inference. Schema guidance is based on a new efficient inclusion algorithms for tree languages defined by deterministic tree automata or XML schemas [12]. In particular they show how to translate XML schemas defined by various classes of EDTDs to bottom-up or top-down deterministic tree automata, based on the Curried or firstchild-nextsibling encoding of unranked into ranked trees.

Kong (a Postdoc in 2008) with Lemay and Gilleron study keyword search for XML [19]. They designed an improvement of the MaxMatch algorithm, called Relaxed Tightest Fragment (RTF). RTF is a representation of the results of a keyword search, together with a query evaluation algorithm. Experiments showed that RTF is more precise than the MaxMatch algorithm, in the sense that it discards more irrelevant nodes while keeping more nodes relevant for the query.

Torre with Terlutte from Grappa reconsider classes of languages that learnable from positive examples alone [23]. They introduce so called rational languages with k-disjoint residuals, and show for every k that the family with k-disjoint residuals subsumes the family of k-reversible languages, known to be learnable from positive examples only. It is also shown that the union for all $k\mathbb{N}$ of the rationals with k-disjoint residuals is the set of all rational languages. Finally, for each k, the corresponding family can be identified in polynomial space and time from positive examples only, when represented by a DFA. In [24], they present general framework for supervised classification based on so called *most general generalizations*. They show that defining such generalizations offer without any further cost, the opportunity to apply algorithms for supervised learning. The authors show how this generic framework can be used for grammatical inference and for classification. The interest of the method is confirmed by experiments.

6.2.2. Statistical Inference

Participants: Jean Decoster, Marc Tommasi, Fabien Torre.

Faddoul, Gilleron and Torre, in collaboration with with Chidlovskii (Xerox Grenoble), study applications of machine learning to the task of labeling documents and authors in social networks. The aim is to propose new algorithms for this task, using relational representations for documents and networks, semi-supervised and multi-task techniques, and label propagation methods.

Gilleron and Torre started the PhD project of Decoster in October. They investigate the use of Inductive Logic Programming in order to automatically classify or transform XML trees. Inductive Logic Programming is a machine learning technique that aims to learn logic programs from examples.

Gilbert, Gilleron and Tommasi continue their work on Tree Series and Weighted Tree Automata as part of the PhD project of Gilbert. Their work has been focusing on problems of convergence of Tree Series and extension of the DEES learning algorithm to tagging tasks. Their collaboration with the LIF Marseille goes on, resulting in a new algorithm for Weighted Tree Automata inference based on Principal Component Analysis.

Laurence, Lemay, Niehren, Staworko and Tommasi continue their work on learning tree transducers, as part of the PhD project of Laurence. They have recently devised an algorithm for learning top-down deterministic tree-to-word transducers from examples. This should allow to infer transformations between documents in different XML schemata.

7. Contracts and Grants with Industry

7.1. Contracts and Grants with Industry

7.1.1. RNTL ATASH (2006-2009)

Participants: Rémi Gilleron [correspondent], Aurélien Lemay, Joachim Niehren, Marc Tommasi.

ATASH is a French industrial project supported by the "Agence Nationale de la Recherche (ANR)". It is a collaboration with the Xerox Research Center Europe XRCE in Grenoble and the LIP6 laboratory. The objective is the design of learning algorithms for tree transformations and their implementation for data integration of documents (PDF, html, doc) in XML databases according to a target DTD. The project began in 2006. The TREECRF[1] library and the R^{2}S^2 software were developed in the project. Mostrare has developped a library TreeCrf (http://treecrf.gforge.inria.fr) that implements conditional random fields for XML data. The library has been used in the R2S2 web application. RS2S is a machine learning application that builds tree transformations from HTML to RSS (an XML dialect). Its aim is to provide personalized feeds from web sites. TreeCrf has also been used to query hidden web in a collaboration with GEMO.

7.1.2. RNTL Webcontent (2006-2009)

Participants: Rémi Gilleron, Marc Tommasi, Fabien Torre [correspondent].

WEBCONTENT is a french industrial project supported by the "Agence Nationale de la Recherche (ANR)". It involves academic partners and companies. The objective is to develop a platform for Web document processing and semantic Web. The main goal of MOSTRARE was to create an extensible Web Service framework for Web information extraction. This software, called MIELE, and was integrated in the WEBCONTENT platform in january. MIELE mainly allows to create wrappers for table extraction from Web documents. The deliverable includes a set of user interface tools (WWW browser plugins) and implementation of wrapper inference algorithms developed in the MOSTRARE team: SQUIRREL containing methods based on query induction using grammatical inference and PAF containing methods based on supervised classification algorithms.

7.1.3. Cifre Xerox (2009-2011)

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

[1] http://treecrf.gforge.inria.fr/

R. GILLERON with F. TORRE started supervising the PhD thesis (Cifre) of Jean-Baptiste FADDOUL with B. CHIDLOVSKI from Xerox's European Research Center (XRCE).

8. Other Grants and Activities

8.1. National Actions

8.1.1. ANR Lampada (2009-2013)

Participants: Édouard Gilbert, Rémi Gilleron, Aurélien Lemay, Marc Tommasi [correspondent], Fabien Torre.

The Lampada project on "Learning Algorithms, Models and sPArse representations for structured DAta" is coordinated by M. 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 (2009-2011)

Participants: Joachim Niehren [correspondent], Sławek Staworko, Aurélien Lemay, Sophie Tison, Anne-Cécile Caron, Olivier Gauwin, 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).

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 PhD project of Gauwin (directed by Niehren and Tison) as described above, contributes to the Mostrare part of Codex on XML streaming. Mostrare also contributes to learning algorithms for XML transformations as needed for schema adaption. These projects are studied in the PhD projects of J. Champavère (directed by Lemay, Niehren, and Gilleron) and G. Laurence (directed by Tommasi, Staworko and Niehren). In this context, we succeeded to hire our previous postdoc S. Staworko as an assistent professor in 2009 by the University of Lille 3. See above for the contributions by these project members on this topic.

8.1.3. ANR Blanc Enum (2007-2010)

Participants: Guillaume Bagan, Olivier Gauwin, 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 (CAEN), N. CREIGNOU (MARSEILLE). 2008–2010. More information about the project 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 proposes to contribute answer enumeration algorithms for XML database queries. We want to distinguish classes of XQuery transformations that allow for efficient enumeration algorithms. We start from tractable fragments of XPath dialects with variables, and from n-ary queries defined by tree automata.

Gauwin is working on enumeration algorithms for conjunctive queries in cooperation with A. Durand (Paris 7) and F. Filiot (Brussels). In 2009, we succeeded to hire G.Bagan from Caen as postdoc, in order to work on efficient answer enumeration for queries defined in Conditional XPath with variables. The topic was prepared by a internship of A. Venant from ENS Cachan-Bretagne supervised by Niehren.

8.1.4. ARA MDCO Marmota (2006-2009)

Participants: Rémi Gilleron, Aurélien Lemay, Joachim Niehren, Marc Tommasi [correspondent].

The Marmota project on "Stochastic Tree Models and Stochastic Tree Transformations" is coordinated by M. TOMMASI from Mostrare. Our partners are: P. GALLINARI (LIP6), F. DENIS (LIF), and M. SEBBAN (SAINT ETIENNE). 2006–2008. More information about the project can be found on http://marmota.gforge.inria.fr/.

Marmota proposes to study computational issues at the intersection of three domains: formal tree languages, machine learning and probabilistic models. Our study is mainly motivated by XML data manipulation: data integration on the Internet from heterogeneous and distributed sources; XML annotation and transformation; XML document classification and clustering. However, fundamental intended results have an important impact in many application domains. For instance, in bioinformatics and music retrieval, it is actually relevant to model data by using probabilistic trees. Therefore, this project is also concerned with the specific problems of these two applications domains and we will use large data sets of these areas. We will consider generative models for tree structured data, non generative models for tree structured data, and models for probabilistic tree pattern matching and probabilistic tree transformations: tree pattern matching algorithms, learning pattern languages, induction of tree transformations.

8.1.5. ARA MDCO CROTAL (2008-2009)

Participants: Rémi Gilleron, Marc Tommasi.

The CROTAL project on "Conditional Random Fields for Natural Language Processing" is coordinated by I. TELLIER, a previous member of Mostrare (until 2008). Our partners are R. MARIN, A. BALVET (linguistics, Lille3), T. POIBEAU, A. ROZENKNOPF (Paris 13), F. YVON (Limsi-CNRS, Paris 11). 2008-2009. More information about the project can be found on http://crotal.gforge.inria.fr/pmwiki-2.1.27/.

Crotal aims at exploring and developing new techniques to access huge textual banks. The project will especially focus on an innovative technique : Conditional Random Fields (CRF), a family of graphical models developed for computational linguistic applications. CRFs allow to annotate data from examples of annotated data. They are at the state of the art level in many domains, including extracting and structuring knowledge. But they also require refinements and optimization to be efficiently applied to large datasets, or to structured data. Our aims are twofold: first, develop new algorithms to process large amount of data; second, apply these algorithms to texts and tree-banks, so that we are able to annotate, extract knowledge and fill knowledge banks from texts. The general purpose is to enrich textual data by learning to annotate them. We plan to work both on English and French corpora.

MOSTRARE proposes to use CRF for trees and to apply them to corpora by experienced teams in the field of Natural Language Processing.

8.2. International Cooperations

Sydney. S Maneth, a senior researcher from NICTA in Sydney (Australia), visited Mostrare for three weeks in September. We started a joint work on a new learning algorithm for top-down deterministic tree to tree transducers. Given that Maneth is also an expert on XML querying languages, the cooperation with his group yields an opportunity for the creation of an associated INRIA team.

Oxford. A cooperation with M. Benedikt and Gabriele Puppis from Oxford (England) on XML streaming has been started in 2009, around the PhD project of Gauwin.

Leuven and Hasselt. A cooperation of Niehren with J. van den Bussche from the University in Hasselt and M. Bruynooghe from the University of Leuven on minimization of tree automata is under way.

Girona. Niehren cooperates with M. Villaret form the University of Girona (Spain). He presented his master lecture on Foundations of XML there.

Frankfurt and Saarbrücken. Niehren continues his collaboration with M. Schmidt-Schauß and D. Sabel from Frankfurt and J. Schwinghammer from Saarbrücken on the semantics of concurrent programming languages.

Rostock and Bologna. Niehren co-supervised with L. Uhrmacher from the University of Rostock (Germany), the PhD project of M. John. The research will be pursued in collaboration with C. Versari from Bologna (Italy).

9. Dissemination

9.1. Scientific Animation

9.1.1. Program Committees

NIEHREN was member of the Steering committee of RTA (the International Conference on Rewriting Techniques and Applications). He participates in the editorial board of “Fundamenta Informaticae”. He is member of the program Committee of LATA’09 (3rd International Conference on Language and Automata Theory and Applications), and member of the Program committee of CIAA’09 (14th International Conference on Implementation and Application of Automata).

R. GILLERON was member of the program committees of ICML’09 (International conference of Machine Learning), EGC’09 (French conference on information extraction and management "Extraction et Gestion des Connaissances"), and CAp’09 (French conference on machine learning "Conference francophone sur l’apprentissage artificiel").

F. TORRE was member of the program committee of CAp’09.

9.1.2. French Scientific Responsibilities

R. GILLERON is head of the research group GRAPPA on machine learning in Lille. He was member of the evaluation committee of Paris 1 and of LIRMM laboratory in Montpellier. He was head of the recruitment jury of the INRIA Lille-Nord Europe center, member of the commission PEDR.

J. NIEHREN was a member of the selection committee for 2 assistant professor positions at Ecole Polytechnique Lille.

S. TISON is director of the LIFL (computer science department in Lille), member of the scientific council of Lille 1 university. She chairs the scientific council of "Pôle de Compétitivité industries du Commerce".

A.-C. CARON was member of the vivier de sélection 27 (selection committee) of University of Lille. She is member of CNU 27 since 2008.

M. TOMMASI was member of the vivier de sélection of Univ. Lille3, Univ. Lens, and Paris 6.

9.1.3. Miscellaneous

NIEHREN gave an invited talk at the FCT Workshop on Non-Classical Models of Automata and Applications (NCMA) in Wraclaw on Streaming Tree Automata and XPath.

9.2. Teaching and Scientific Diffusion

9.2.1. Teaching

Iovka BONEVA	192 hours	bachelor
Anne-Cécile CARON	192 hours	bachelor and masters
Jérôme CHAMPAVÈRE	48 hours	bachelor
Rémi GILLERON	192 hours	masters
Aurélien LEMAY	192 hours	bachelor and masters
Joachim NIEHREN	65 hours	bachelor and masters
Yves ROOS	192 hours	bachelor and masters
Sławek STAWORKO	96 hours	bachelor and masters
Marc TOMMASI	192 hours	masters
Sophie TISON	96 hours	masters

CARON is responsible of the master IPI-NT of Univ. Lille, since 2009.

GILLERON is responsible of the master MIASHS'.

LEMAY is coordinating computer science teaching at UFR LEA.

TOMMASI is responsible of the GIDE branch of the master ICD of Univ. Lille.

ROOS is responsible of the Miage Master of Univ. Lille.

9.2.2. Master lectures at the University of Lille

- Foundations of XML: A.-C. CARON and J. NIEHREN
- Bases de données avancées (BDA): J. NIEHREN
- Advanced course on databases : A.-C. CARON
- Project management (Gestion de projets informatiques) : S. STAWORKO
- Programming for XML : Y. ROOS
- M. TOMMASI is in charge of the following courses of the master "Computer Science and Documentation" : XSLT, Algorithmics and programming, Networks
- M. TOMMASI is in charge of the following courses of the master "Management of information in companies" : Numeric documents, Web technologies
- R. GILLERON is in charge of the following courses of the master MIASHS' : Supervised classification, Non supervised classification, Information retrieval.
- F. TORRE is in charge of the course on Machine Learning of the Master MOCAD.

9.2.3. International Master lectures

- Foundations of XML, University of Girona, Spain: J. NIEHREN

9.2.4. Master projects

J. DECOSTER, relational learning for classifying, transforming, and extracting information from XML documents. Supervised by GILLERON and TORRE.

9.2.5. PhD theses

- O. GAUWIN, Streaming Tree Automata and XPath. PhD defended in October 2009. Supervised by NIEHREN and TISON
- 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. 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, ROOS, CARON and STAWORKO
- J.-B. FADDOUL, Machine learning and applications to social network analysis. Supervised by GILLERON and SCHIDLOWSKII from Xerox european research center (XRCE).
- J. DECOSTER, statistical relational learning of XML transformations. Supervised by TOMMASI and TORRE.

9.2.6. PhD committees

- GILLERON belonged to the committees of G. STEMPFEL (Marseille, referee), F. MAES (Paris 6, referee), N. ZERIDA (Paris 8, member)
- NIEHREN belonged to the committees of G. BAGAN (Caen, examinateur) and O. GAUWIN (Univ. Lille, director)
- TISON belonged to the committees of E. TANTAR (Univ. Lille, president), and O. GAUWIN (Univ. Lille, co-director).
- TOMMASI belonged to the committee of F. TANTINI (Saint Étienne, president)

9.2.7. Habilitation committees

- TISON belonged to the committee of F. NAÏT (Univ. Lille, president)

10. Bibliography

Major publications by the team in recent years

[1] I. BONEVA, J.-M. TALBOT, S. TISON. *Expressiveness of a spatial logic for trees*, in "Proceedings of the 20th Annual IEEE Symposium on Logic in Computer Science (LICS'05)", IEEE Comp. Soc. Press, 2005, p. 280 - 289.

[2] L. CANDILLIER, I. TELLIER, F. TORRE, O. BOUSQUET. *Cascade Evaluation of Clustering Algorithms*, in "17th European Conference on Machine Learning (ECML'2006)", Lecture Notes in Artificial Intelligence, vol. 4212, Springer Verlag, 2006, p. 574–581.

[3] J. CARME, R. GILLERON, A. LEMAY, J. NIEHREN. *Interactive Learning of Node Selecting Tree Transducers*, in "Machine Learning", vol. 66, n^{o} 1, 2007, 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", vol. 207, n^{o} 11, 2009, p. 1181-1208, http://hal.inria.fr/inria-00366082/en/.

[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. *Satisfiability of a Spatial Logic with Tree Variables*, in "16th EACSL Annual Conference on Computer Science and Logic", Lecture Notes in Computer Science, vol. 4646, Springer Verlag, 2007, p. 130-145, http://hal.inria.fr/inria-00148462.

[7] O. GAUWIN, J. NIEHREN, S. TISON. *Bounded Delay and Concurrency for Earliest Query Answering*, in "3rd International Conference on Language and Automata Theory and Applications, Espagne Tarragona", A. H. DEDIU, A. M. IONESCU, C. MARTIN-VIDE (editors), vol. 5457, Springer, 2009, p. 350-361, http://hal.inria.fr/inria-00348463/fr.

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

[9] 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", vol. P2747, IEEE Comp. Soc. Press, 2006, p. 997-1004.

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

Year Publications

Doctoral Dissertations and Habilitation Theses

[11] O. GAUWIN. *Flux XML, Requêtes XPath et Automates*, Université des Sciences et Technologie de Lille - Lille I, 09 2009, http://tel.archives-ouvertes.fr/tel-00421911/en/, Ph. D. Thesis.

Articles in International Peer-Reviewed Journal

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

[13] M. JOHN, C. LHOUSSAINE, J. NIEHREN, A. UHRMACHER. *The Attributed Pi Calculus with Priorities*, in "Transactions on Computational Systems Biology", 2009, http://hal.inria.fr/inria-00422969/en/DE.

[14] M. LATTEUX, Y. ROOS, A. TERLUTTE. *Minimal NFA and biRFSA Languages*, in "RAIRO - Theoretical Informatics and Applications", vol. 43, n^{o} 2, 2009, p. 221-237, http://hal.inria.fr/inria-00296658/en/.

International Peer-Reviewed Conference/Proceedings

[15] O. GAUWIN, J. NIEHREN, S. TISON. *Bounded Delay and Concurrency for Earliest Query Answering*, in "3rd International Conference on Language and Automata Theory and Applications, Espagne Tarragona", A. H. DEDIU, A. M. IONESCU, C. MARTIN-VIDE (editors), vol. 5457, Springer, 2009, p. 350-361, http://hal.inria.fr/inria-00348463/en/.

[16] O. GAUWIN, J. NIEHREN, S. TISON. *Earliest Query Answering for Deterministic Nested Word Automata*, in "17th International Symposium on Fundamentals of Computer Theory, Pologne Wraclaw", vol. 5699, SV, 2009, p. 121-132, http://hal.inria.fr/inria-00390236/en/.

[17] B. GROZ, S. STAWORKO, A.-C. CARON, Y. ROOS, S. TISON. *XML Security Views Revisited*, in "International Symposium on Database Programming Languages, France LYON", F. G. PHILIPPA GARDNER (editor), vol. LNCS, Springer, 2009, p. pp 52-67, http://hal.archives-ouvertes.fr/hal-00396796/en/.

[18] M. JOHN, C. LHOUSSAINE, J. NIEHREN. *Dynamic Compartments in the Imperative Pi Calculus*, in "Dynamic Compartments in the Imperative Pi Calculus, Italie Bologna", vol. 5688, Spinger, 2009, p. 235-250, http://hal.inria.fr/inria-00422970/en/DE.

[19] K. LINGBO, R. GILLERON, A. LEMAY. *Retrieving Meaningful Relaxed Tightest Fragments for XML Keyword Search*, in "EDBT 2009, Russie Saint Petersburg", 2009, http://hal.inria.fr/inria-00433097/en/.

[20] E. MOREAU, I. TELLIER, A. BALVET, G. LAURENCE, A. ROZENKNOP, T. POIBEAU. *Annotation fonctionnelle de corpus arboré avec des Champs Aléatoires Conditionnels*, in "TALN 2009, France Senlis", 2009, ., http://hal.archives-ouvertes.fr/hal-00436330/en/.

[21] J. SCHWINGHAMMER, D. SABEL, M. SCHMIDT-SCHAUSS, J. NIEHREN. *Correctly Translating Concurrency Primitives*, in "The 2009 SIGPLAN Workshop on ML, Royaume-Uni Edinburgh", A. ROSSBERG (editor), 2009, p. 27-38, http://hal.inria.fr/inria-00429239/en/.

[22] S. STAWORKO, G. LAURENCE, A. LEMAY, J. NIEHREN. *Equivalence of Deterministic Word to Word Transducers*, in "Fundamuntals of Computation Theory, Pologne Wroclaw", M. GEBALA, M. KORZENIOWSKA, W. CHARATONIK (editors), vol. 5699, Springer, Maciej Gebala and Malgorzata Korzeniowska, 2009, p. 310-322, http://hal.inria.fr/inria-00423961/en/.

National Peer-Reviewed Conference/Proceedings

[23] A. TERLUTTE, F. TORRE. *Identification des langages rationnels à résiduels k-disjoints*, in "11e Conférence francophone sur l'Apprentissage automatique (CAp'2009), Tunisie Hammamet", Y. BENNANI, C. ROUVEIROL (editors), 2009, p. 21-32, http://hal.inria.fr/inria-00425073/en/.

[24] F. TORRE, A. TERLUTTE. *Méthodes d'ensemble en inférence grammaticale : une approche à base de moindres généralisés*, in "11e Conférence francophone sur l'Apprentissage automatique (CAp'2009), Tunisie Hammamet", Y. BENNANI, C. ROUVEIROL (editors), 2009, p. 33-48, http://hal.inria.fr/inria-00425072/en/.

References in notes

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

[26] 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.

[27] J. BERSTEL, C. REUTENAUER. *Recognizable formal power series on trees*, in "Theorical computer science", vol. 18, 1982, p. 115–148.

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

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

[30] 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.

[31] F. DENIS, A. HABRARD. *Learning rational stochastic tree languages*, in "Algorithmic learning theory", M. HUTTER, R. A. SERVEDIO, E. TAKIMOTO (editors), Lecture Notes in Artificial Intelligence, vol. 4754, Springer-Verlag, 18th International Conference, ALT 2007, Octobre 2007, p. 242–256.

[32] A. DOAN, A. Y. HALEVY. *Semantic Integration Research in the Database Community: A Brief Survey*, in "AI magazine", vol. 26, n^{o} 1, 2005, 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", vol. 9, 1975, p. 198–231.

[35] J. ENGELFRIET, S. MANETH. *Macro tree transducers, attribute grammars, and MSO definable tree translations*, in "Information and Computation", vol. 154, n^{o} 1, 1999, 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", vol. 6, n^{o} 13, 2003, 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, vol. 3774, 2005.

[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, vol. 2572, Springer Verlag, 2003, 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.", vol. 15, 1993, 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", vol. 6, 2005, 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.