INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team MOSTRARE

Modeling Tree Structures, Machine Learning, and Information Extraction

Lille - Nord Europe

THEME SYM

Table of contents

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, project leader, HdR]
Iovka Boneva [assistant professor, since september 2008]
Anne-Cécile Caron [assistant professor]
Aurélien Lemay [assistant professor]
Yves Roos [assistant professor]
Isabelle Tellier [assistant professor, professor in Orléans since september 2008, HdR]
Sophie Tison [professor, HdR]
Marc Tommasi [professor, HdR]
Fabien Torre [assistant professor]

Technical Staff

Matthieu Keith [INRIA young software engineer from November 2006 to November 2008]
Hanh-Missi Tran [ATASH project, software engineer from January 2007 to June 2008]
Feriel Lahlali [INRIA young software engineer since December 2007]

PhD Student

Jérôme Champavère [MESR fellowship, since October 2006]
Emmanuel Filiot [INRIA and Région Nord-Pas-de-Calais fellowship, from October 2005 to September 2008]
Olivier Gauwin [INRIA Cordi fellowship, since November 2006]
Benoît Groz [ENS Cachan since September 2008]
Edouard Gilbert [ENS Bretagne since November 2007]
Grégoire Laurence [MESR, since October 2008]
Damien Poirier [CIFRE FRANCE TELECOM since November 2007]
Jean-Baptiste Faddoul [CIFRE XEROX since December 2008]

Post-Doctoral Fellow

Mathias Samuelides [temporary assistant professor from September 2007 to June 2008]
Sławek Staworko [INRIA, from July 2007 to June 2009]
Ling-bo Kong [WEBCONTENT project, from April 2007 to November 2008]

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 summarise our two main research objectives:

Modeling tree structures for information extraction. We wish to extend studies of 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 generalise on node selection queries.

Machine learning for information extraction. We wish to extend our study of 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 of the year

- Filiot's thesis [11]: *Logics for n-ary queries in trees*.
- The Miele software: a Web service for information extraction was integrated in the WEBCONTENT platform.

3. Scientific Foundations

3.1. Modeling XML document transformations

Keywords: *automata*, *logic*, *queries*, *semi-structured documents*, *transformations*, *trees*.

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 [41], Xtatic [39], [44], and CDuce [28], [29], [31]. 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 or 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 [43], [53]. 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 [47], [45].

The automata community usually approaches tree transformations by tree transducers [37], i.e., tree automata producing output structure. Macro tree transducers, for instance, have been proposed recently for defining XML transformations [45], [48]. From the view point of logics, tree transducers have been studied for MSO definability [38].

3.2. Machine learning for XML document transformations

Keywords: *grammatical inference*, *statistical learning*, *tree annotations*, *tree transformations*, *wrapper induction*.

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 [32], [49]. 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 [40], [42]. 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 [36].

Probabilistic context free grammars (pCFGs) [46] are used in the context of PDF to XML conversion [33]. In a first step, a labeling procedure of leaves of the input document by labels of the output DTD is learned. In a 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 [50]). 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 [52], [51]. 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

Keywords: *Web intelligence*, *data integration*, *document processing*, *peer data management systems*, *semantic Web*, *semantic integration*.

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 [35] 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 organised w.r.t visual format (HTML, DOC, PDF) into documents organised 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.

5. Software

5.1. MIELE: a Web service for information extraction

Keywords: *Web data*, *Web service*, *table extraction*, *wrapper induction*.

Participants: Fabien Torre [correspondent], Matthieu Keith, Marc Tommasi, Aurélien Lemay, Missi Tran.

The MIELE project is in the final stage of development. The main goal of this project is to create an extensible Web Service framework for Web information extraction. It 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: SQUIRREL containing methods based on query induction using grammatical inference and PAF containing methods based on supervised classification algorithms. MIELE is integrated in the WEBCONTENT platform dedicated to the Semantic Web.

5.2. PICCATA

Keywords: *Multiplicity tree automata*, *unranked trees*.

Participants: Edouard 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 focus on real-valued automata. The model is the one introduced by [30]. 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. The DEES algorithm [34] is currently implemented.

6. New Results

6.1. Modeling XML document transformations

6.1.1. XML database queries, logic and automata

Keywords: *Querying*, *XPath*, *XQuery*, *node selection queries*, *streaming*, *tree transformations*.

Participants: Olivier Gauwin, Emmanuel Filiot, Mathias Samuelides, Sławek Staworko, Anne-Cécile Caron, Joachim Niehren, Yves Roos, Sophie Tison [correspondent].

Gauwin, Niehren, and Roos [12] introduce *streaming tree automata* (STAs), a new notion of tree automata for unranked trees. While being of interest for streaming XML processing, STAs can be shown to be equally expressive as both, Alur's (2007) nested word automata and Neumann and Seidl's (1998) pushdown forest automata. The advantage of streaming tree automata is that they directly operate on unranked trees, rather than nested words or forests.

Gauwin, Caron, Niehren, and Tison [19] apply STAs to streaming query answering. They investigate earliest query answering, as needed for query answering with optimal memory management. They propose a new algorithm for earliest query answering, which require only polynomial time combined complexity. It applies to n-ary node selection queries in unranked trees defined by deterministic STAs. This class is highly expressive in that captures all MSO definable n-ary queries (even though not modulo polynomial time). As a corollary, they obtain an earliest query answering algorithm for CoreXPath 2.0 with polynomial time data complexity. This seems close to optimal as they show, since deciding earliest selection is coNP-hard for XPath even when restricted to Forward XPath with descendant axis only. Without determinism, earliest selection becomes DEXPTIME-complete for n-ary queries defined by whatsoever kinds of tree automata.

Filiot and Tison [18] investigate the variable independence problem for n-ary queries in trees defined by MSO formulas with n free first-order variables. They show how to decide whether a regular query is equivalent to a union of cartesian products, independently of the input tree. They introduce variable independence w.r.t. a dependence forest between blocks of variables, which they prove to be decidable.

Filiot, Talbot[1], and Tison [17] study TAGEDs (*tree automaton with global equality and disequality constraints*). This kind of automaton on trees allows to test (dis)equalities between subtrees which may be arbitrarily faraway. In particular, it is equipped with an (dis)equality relation on states, so that whenever two subtrees t and t' evaluate (in an accepting run) to two states which are in the (dis)equality relation, they must be (dis)equal. They prove decidability of emptiness of several classes and give two applications of TAGEDs: decidability of an extension of Monadic Second Order Logic with tree isomorphism tests and of unification with membership constraints.

Staworko, Filiot and Chomicki [23] investigate the problem of querying (regular) sets of `XML` documents represented with tree automata and consider n-ary tree automata queries whose expressive power captures MSO on trees. Because finite automata can represent infinite sets of documents, they propose the notions of *universal* and *existential* query answers, answers that are present resp. in all and some documents. They study complexity of query answering and show that computing existential query answers is in PTIME under the assumption that the arity of the query is a fixed parameter. On the other hand, computing universal query answers is EXPTIME-complete, but they show that it is in PTIME if one assumes that the query is fixed (data complexity). Finally, the framework captures problems central to many novel `XML` applications like querying inconsistent `XML` documents. In particular, they demonstrate how to use this framework to compute consistent query answers in `XML` documents that do not satisfy the schema.

Niehren collaborated with Kuhlmann from Saarbrücken [21] on the monadic second-order logic (MSO) for totally ordered trees. Totally ordered trees are ground terms equipped with an additional total order on their nodes. They provide a formal model for data that comes with both a hierarchical and a sequential structure; one example for such data are streaming in natural language, another are natural language sentences, where a sequential structure is given by word order, and a hierarchical structure is given by grammatical relations between words. They show that the MSO satisfiability problem of unrestricted structures is undecidable, but give a decision procedure for practically relevant sub-classes, based on tree automata.

Roos, Terlutte and Latteux [13] define the notion of biRFSA which is a residual finate state automaton (RFSA) whose the reverse is also an RFSA. The languages recognized by such automata are called biRFSA languages. They prove that the canonical RFSA of a biRFSA language is a minimal NFA for this language and that each minimal NFA for this language is a sub-automaton of the canonical RFSA. This leads to a characterization of the family of biRFSA languages. They also define the family of biseparable automata and prove that every biseparable NFA is uniquely minimal among all NFAs recognizing a same language, improving the result of H. Tamm and E. Ukkonen for bideterministic automata.

Tison and Roos started the PhD project of Groz in September, jointly with Caron and André. They investigate XML database security, especially access control for `XML` documents.

6.1.2. Programming languages

Keywords: *Concurrency*, *rewriting*, *semantics*, *stochastic programming*, *system biology*.

[1] J.M. Talbot was a member of Mostrare until 2006 and is now professor in Marseille

Participants: Joachim Niehren [correspondent], Sophie Tison.

Niehren continues participating in the BioComputing activity, leaded by Lhoussaine at the LIFL. Together, they started a cooperation with John and Uhrmacher [20] from Rostock. This has resulted in *the attributed pi calculus ($\pi(L)$)*, an extension of the stochastic pi calculus with attributed processes and attribute dependent synchronization. A stochastic simulator for this modeling language for systems biology has been presented and implemented. This show that the extension by attributes can be handled with reasonable efficiency.

Niehren continues his cooperation with Schmidt-Schauß and Sabel from Frankfurt and Schwinghammer from Saarbrücken. In [22] they investigate methods and tools for analysing translations between programming languages with respect to observational semantics. The behaviour of programs is observed in terms of may-and-must convergence in arbitrary contexts, and adequacy of translations, i.e., the reflection of program equivalence, is taken to be the fundamental correctness condition. For compositional translations they propose a notion of convergence equivalence as a means for proving adequacy. This technique avoids explicit reasoning about contexts, and is able to deal with the subtle role of typing in implementations of language extension.

Tison continues her cooperation with Godoy from Barcelona and Maneth from Sidney [25]. They study the well known open problem of the decidability of regularity preservation by a homomorphism for regular tree languages. They consider two interesting subclasses. First, they prove that regularity preservation is decidable in polynomial time when the domain language is constructed over a monadic signature, i.e., over a signature where all symbols have arity 0 or 1. Second, they prove decidability for the case where non-linearity of the homomorphism is restricted to the root node (or nodes of bounded depth) of any input term. They also prove the decidability of this problem: *given a set of terms with regular constraints on the variables, is its set of ground instances regular?* This extends previous results where regular constraints where not considered.

6.2. Machine learning for XML document transformations

6.2.1. Grammatical Inference

Keywords: *node selection queries*, *tree automata*.

Participants: Jérôme Champavère, Rémi Gilleron, Aurélien Lemay [correspondent], Joachim Niehren, Grégoire Laurence, Marc Tommasi.

Champavère, Gilleron, Lemay, and Niehren [15] investigate the induction of monadic node selecting queries from partially annotated `XML`-trees. They show how incorporate the document schema information into existing algorithms for learning tree automata queries, like RPNI-based learning algorithm. None of the alternative approaches to wrapper induction has included schema information so far, most probably, since they cannot guide learning of queries when are represented stochastically (rather than by tree automata). In our case, monadic queries are represented by pruning node selecting tree transducers. Since target queries of the learning problem are subject to schema constraints, the idea is to avoid generalization errors in the learning process by taking the schema information into account. Compatible queries select answers only from documents that are consistent with the given schema. We have implemented the new learning algorithm with schema guidance. Experimental results for guidance by the schema of `HTML` are presented.

From the algorithmic perspective, the central problem of schema guided query induction is inclusion checking in deterministic tree automata or DTDs. Champavère, Gilleron, Lemay, and Niehren [14] present a new efficient algorithm for this inclusion problem. For testing language inclusion $L(A) \subseteq L(B)$ between tree automata, it operates in time $O(|A| * |B|)$ if B is deterministic. It can be applied to testing inclusion $L(A) \subseteq L(D)$ in deterministic DTDs D in time $O(|A| * |\Sigma| * |D|)$ where Σ is the signature. No previous algorithms with these complexities existed.

Tellier studies connections between Categorial Grammars and Recursive Automata in [24]. She exhibits connexions between learning strategies "by specialisation" implemented in both contexts. This leads to a new interpretation of previous works on learning categorial grammars and provides a better understanding of when this strategy can be effectively used. It is the case when some kind of "bounds" about the target is available, in the form of information derived from it by a morphism.

Tommasi, Lemay, Staworko and Niehren started the PhD project of Laurence on learning streaming tree transducers. The objective is to lift previous learning algorithms for node selection queries to transformations, in order to approach new applications in Web Data Exchange.

6.2.2. Statistical Inference

Keywords: *XML trees*, *conditional random fields*, *probabilistic automata*, *tree labeling*.

Participants: Edouard Gilbert, Florent Jousse, Lingbo Kong, Rémi Gilleron, Aurélien Lemay, Isabelle Tellier, Marc Tommasi [correspondent].

Gilbert, Gilleron, and Tommasi in collaboration with Denis, Habrard and Ouardi study probability distributions over free algebras of trees. and show that distributions can be defined by weighted tree automata or by tree series. They adapt definitions to handle the case of unranked trees and define learning algorithms for probability distributions in this case. In [16] they show that any representation of a rational stochastic tree language can be transformed in a reduced normalized representation that can be used to generate trees from the underlying distribution. They also study some properties of consistency for rational stochastic tree languages and discuss their implication for the inference.

Tommasi and Gilleron in collaboration with Senellart, Mittal and Muschick [26] propose an original approach to the automatic induction of wrappers for sources of the hidden Web that does not need any human supervision. This approach only needs domain knowledge expressed as a set of concept names and concept instances. There are two parts in extracting valuable data from hidden-Web sources: understanding the structure of a given `HTML` form and relating its fields to concepts of the domain, and understanding how resulting records are represented in an `HTML` result page. For the former problem, they use a combination of heuristics and of probing with domain instances; for the latter, they use a supervised machine learning technique adapted to tree-like information (`XCRFs`) on an automatic, imperfect, and imprecise, annotation using the domain knowledge. Some experiments demonstrate the validity and potential of the approach.

Laurence and Tellier have started experiments using `XCRFs` for natural language processing. In the context of the ANR MDCO "Crotal" (CRFs for TAL), Tellier and colleagues made experiments using `XCRFs` on linguistic treebanks. The corpus used was the French treebank produced by the Paris 7 team, a set of 10 000 sentences extracted from the French newspaper "Le monde", syntacticaly analyzed and tagged with functionnal labels (of the kind : "SUJ", "OBJ"...). The purpose was to evaluate whether these labels could be inferred from the syntactic structures alone by XCRFs.

Kong, Lemay and Gilleron proposed new algorithms for adapting keyword search to `XML` data. They proposed a framework of retrieving meaningful fragments in `XML` data. They defined new filtering mechanisms in order to improve the quality of answers.

7. Contracts and Grants with Industry

7.1. Contracts and Grants with Industry

7.1.1. RNTL ATASH

Participants: Rémi Gilleron [correspondent], Florent Jousse, 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 has begun in 2006. The TREECRF[2] library and the `R`^{2}`S`^{2} software were developed in the project.

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

7.1.2. RNTL Webcontent

Participants: Rémi Gilleron, Florent Jousse, 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. MOSTRARE is involved in the work packages "Content Extraction" and "Semantic Enrichment". The MIELE Web service for information extraction is a deliverable.

7.1.3. Others

I. TELLIER co-supervise with P. GALLINARI, LIP6 the PhD thesis (Cifre) of Damien POIRIER with the France Telecom company.

R. GILLERON supervise the PhD thesis (Cifre) of Jean-Baptiste FADDOUL with the Xerox european research center (XRCE).

R. GILLERON, M. TOMMASI and F. TORRE initiate a collaboration with the JOUVE company on automatic structuration of texts. the master project of S. ACCART was done inside this collaboration.

8. Other Grants and Activities

8.1. French Actions

8.1.1. ANR Defis Codex (2009-2011): Efficiency, Dynamicity and Composition for XML Models, Algorithms, and Systems

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

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.

Coordinated by MANOLESCU (GEMO, INRIA Saclay), with GENEVES (WAM, INRIA Grenoble), COLAZZO (LRI, Orsay), CASTAGNA (PPS, Paris 7), and HALFELD (Blois).

8.1.2. ANR Blanc Enumeration (2007-2010): Complexity and Algorithms for Answer Enumeration

Participants: Olivier Gauwin, Joachim Niehren [correspondent], Sophie Tison.

We propose to study 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.

Our partners are: Arnaud DURAND (coordinator - PARIS VII), Etienne GRANDJEAN (CAEN), Nadia CREIGNOU (MARSEILLE). 2008–2010. More information about the project can be found on http://enumeration.gforge.inria.fr.

8.1.3. ARA MDCO CROTAL (2008-2009): Conditional Random Fields for Natural Language Processing

Participants: Rémi Gilleron, Marc Tommasi, Isabelle Tellier [correspondent].

The CROTAL project 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 optimisation to be efficiently applied to large datasets, or to structured data. More precisely, 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.

The coordinator of the project is I. TELLIER. 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/.

8.1.4. ANR Jeune BioSpace (2008-2010): A Uniform Approach for Stochastic Modeling with Spatial Aspects in Systems Biology

Participant: Joachim Niehren [correspondent].

Stochastic modeling and simulation seeks to improve the understanding of genetic networks in systems biology. BioSpace proposes to develop, design, and implement a novel and generic stochastic modeling language that is able to cope with all kinds of spatial phenomena in molecular networks with space-dependent concurrent control. We hope to find a unifying framework accessible to biologists, that extends on existing rule based approaches while providing for compartments with variable volumes in particular. The development of our new language will be accompanied by its application to cellular biology. Our modeling studies will focus on spatial aspects in eukaryotic gene regulation: positioning of chromosomes in the nucleus, establishment and maintenance of nuclear compartments, and cross-talk between chromosomes. This project is leaded by Cédric LHOUSSAINE from the BioComputing activity at the LIFL in Lille.

8.1.5. ARA MDCO Marmota (2006-2008): Stochastic Tree Models and Stochastic Tree Transformations

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

We propose 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.

The coordinator of the project is M. TOMMASI. 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/.

9. Dissemination

9.1. Scientific animation

- **Program Committees:**

  R. GILLERON was PC member of ICML'2008 (International Conference on Machine Learning), ICGI'2008 (International Conference on Grammatical Inference), EGC'2008 (french conference on knowledge discovery).

  J. NIEHREN was PC member of CMSB'2008 (Conference on Computational Methods in Systems Biology) and UNIF'2008 (International Workshop on Unification).

  Y. ROOS was PC member of MFCS'2008 (International Symposium on Mathematical Foundations of Computer Science)

  I. TELLIER was PC member of ICGI'2008 (International Conference on Grammatical Inference), PC member of CORIA'2008 (french conference on information retrieval) and member of the Redaction Committee of the French journal TAL.

  S. TISON was member of the editorial board of RAIRO- THEORETICAL INFORMATICS AND APPLICATIONS, was PC member of

  M. TOMMASI was PC member of ECML'2008 (European Conference on Machine Learning).

  F. TORRE was PC member of CAP'2008. (french conference on Machine Learning).

- **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 the computer science department of Paris 6. He was member of the evaluation committee for research (PEDR). He was head of the selection committee for junior researchers in UR INRIA LILLE NORD-EUROPE. He was member of the steering committee for a spring school on machine learning EPIT'2008.

  J. NIEHREN was a member of the CR2 selection committee of INRIA Futurs in Bordeaux and the commission des spécialistes en section 26/27 à Lille 3.

  I.TELLIER was member of the CNU 27 (national committee for the evaluation of assistant professors and professors in computer science) until 2008. She was member of the evaluation committee of the computer science reseach department LIMSI of Paris 11.

  S. TISON is director of the LIFL (computer science department in Lille), head of the research group STC of the LIFL, 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 CSE (Commission de Spécialistes de l'Enseignement supérieur) of University of Valenciennes, is member of CNU 27 since 2008.

9.2. Teaching and scientific diffusion

- TEACHING

Anne-Cécile CARON	192 hours	bachelor and masters
Rémi GILLERON	192 hours	masters
Aurélien LEMAY	192 hours	bachelor and masters
Joachim NIEHREN	10 hours	masters
Yves ROOS	192 hours	bachelor and masters
Isabelle TELLIER	192 hours	bachelor and masters
Marc TOMMASI	192 hours	masters
Fabien TORRE	192 hours	bachelor and masters
Sophie TISON	96 hours	masters

- MASTER LECTURES PRESENTED AT THE UNIVERSITY OF LILLE
  - Logic and Modelisation: A.-C. CARON, J. NIEHREN, and S. TISON
  - Machine Learning for Information Extraction: F. TORRE
  - Supervised Classification: R. GILLERON
  - Classification: R. GILLERON
  - Advanced Algorithms and Complexity: SLAWEK STAWORKO and SOPHIE TISON
  - Frameworks for Web Programming and XML Publishing: M. TOMMASI
  - Advanced Databases: A.-C. CARON
  - Computational Linguistics: I. TELLIER
  - Information Retrival and the Semantic Web: I. TELLIER
- MASTER PROJECTS:
  - S. ACCART, automatic structuration for texts, supervised by R. GILLERON, M. TOMMASI, F. TORRE.
  - G. LAURENCE, learning recognizable relations by grammatical inference, supervised by A. LEMAY, J. NIEHREN, S. STAWORKO and M. TOMMASI.
- DIRECTION OF PHD THESIS SUBMITTED IN 2008:
  - E. FILIOT, logics for n-ary queries in trees, October 13th, supervised by S. TISON.
- HABILITATION THESIS IN 2008:

  S. TISON belonged to the committees of S. COTIN and J.-S. VARRÉ.
- PHD COMMITTEES:

  M. TOMMASI was member (reviewer) of the committee of A. BONDU (Angers and Orange telecom).

  R. GILLERON was member of the committees of C. NOGUERA, A. BONDU.

  J. NIEHREN was reviewer of the committees for S. RAEYMAEKERS, Louvain, Belgique, D. SABEL, Frankfurt Allemagne, and M. SAMUELIDES, LIAFA Paris 7.

  S. TISON belonged to the committees of J. TIERNY, M. ADDA, S. MONGY, C. COSTERMANS, E. FILIOT, B. FILA (Orléans, reviewer).

10. Bibliography

Major publications by the team in recent years

[1] Y. ANDRÉ, A.-C. CARON, D. DEBARBIEUX, Y. ROOS, S. TISON. *Path Constraints in Semi-Structured Data*, in "Theoretical Computer Science", vol. 385, n^{o} 1-3, 2007, p. 11-33.

[2] 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.

[3] 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.

[4] 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, https://hal.inria.fr/inria-00087226.

[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, https://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, Y. ROOS. *Streaming Tree Automata*, in "Information Processing Letters", vol. 109, 2008, p. 13-17, http://hal.inria.fr/inria-00288445/en/.

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

[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, https://hal.inria.fr/inria-00088406.

Year Publications

Doctoral Dissertations and Habilitation Theses

[11] E. FILIOT. *Logics for n-ary queries in trees.*, Ph. D. Thesis, Université des Sciences et Technologie de Lille - Lille I, 10 2008, http://tel.archives-ouvertes.fr/tel-00330524/en/.

Articles in International Peer-Reviewed Journal

[12] O. GAUWIN, J. NIEHREN, Y. ROOS. *Streaming Tree Automata*, in "Information Processing Letters", vol. 109, 2008, p. 13-17, http://hal.inria.fr/inria-00288445/en/.

[13] M. LATTEUX, Y. ROOS, A. TERLUTTE. *Minimal NFA and biRFSA Languages*, in "RAIRO - Theoretical Informatics and Applications", 2008, http://hal.inria.fr/inria-00296658/en/.

International Peer-Reviewed Conference/Proceedings

[14] J. CHAMPAVÈRE, R. GILLERON, A. LEMAY, J. NIEHREN. *Efficient Inclusion Checking for Deterministic Tree Automata and DTDs*, in "2nd International Conference on Language and Automata Theory and Applications Language and Automata Theory and Applications Lecture Notes in Computer Science, Espagne Tarragona", C. MARTIN-VIDE, F. OTTO, H. FERNAU (editors), Lecture Notes in Computer Science, vol. 5196, Springer, 2008, p. 184-195, http://hal.inria.fr/inria-00192329/en/.

[15] J. CHAMPAVÈRE, R. GILLERON, A. LEMAY, J. NIEHREN. *Schema-Guided Induction of Monadic Queries*, in "9th International Colloquium on Grammatical Inference Grammatical Inference: Algorithms and Applications Lecture Notes in Artificial Intelligence, France Saint-Malo", A. CLARK, F. COSTE, L. MICLET (editors), Lecture Notes in Artificial Intelligence, vol. 5278, Springer, 2008, p. 15-28, http://hal.inria.fr/inria-00309408/en/.

[16] F. DENIS, E. GILBERT, A. HABRARD, F. OUARDI, M. TOMMASI. *Relevant Representations for the Inference of Rational Stochastic Tree Languages*, in "International Colloquium on Grammatical Inference, France St Malo", F. COSTE, A. CLARK, L. MICLET (editors), vol. 5278, Springer Verlag, 2008, p. 57-70, http://hal.archives-ouvertes.fr/hal-00293511/en/.

[17] E. FILIOT, J.-M. TALBOT, S. TISON. *Tree Automata with Global Constraints*, in "12th International Conference on Developments in Language Theory (DLT), Japon Kyoto", 2008, p. 314-326, http://hal.inria.fr/inria-00292027/en/.

[18] E. FILIOT, S. TISON. *Regular n-ary Queries in Trees and Variable Independence*, in "5th IFIP International Conference on Theoretical Computer Science, Italie Milano", 2008, p. 429-443, http://hal.inria.fr/inria-00274648/en/.

[19] O. GAUWIN, A.-C. CARON, J. NIEHREN, S. TISON. *Complexity of Earliest Query Answering with Streaming Tree Automata*, in "ACM SIGPLAN Workshop on Programming Language Techniques for XML (PLAN-X), États-Unis d'Amérique San Francisco", 2008, http://hal.inria.fr/inria-00336169/en/.

[20] M. JOHN, C. LHOUSSAINE, J. NIEHREN, A. UHRMACHER. *The Attributed Pi Calculus*, in "Computational Methods in Systems Biology, 6th International Conference CMSB, Allemagne Rostock", Lecture Notes in Bioinformatics, Springer, 2008, p. 83-102, http://hal.inria.fr/inria-00308970/en/.

[21] M. KUHLMANN, J. NIEHREN. *Logics and Automata for Totally Ordered Trees*, in "19th International Conference on Rewriting Techniques and Applications Lecture Notes in Computer Science, Autriche Linz", Lecture Notes in Computer Science, vol. 5117, Springer Verlag, 2008, p. 217-231, http://hal.inria.fr/inria-00257278/en/.

[22] M. SCHMIDT-SCHAUSS, J. NIEHREN, J. SCHWINGHAMMER, D. SABEL. *Adequacy of compositional translations for observational semantics*, in "5th IFIP International Conference on Theoretical Computer Science IFIP, Italie Milano", IFIP, vol. 273, Springer Verlag, 2008, p. 521-535, http://hal.inria.fr/inria-00257279/en/.

[23] S. STAWORKO, E. FILIOT, J. CHOMICKI. *Querying Regular Sets of XML Documents*, in "International Workshop on Logic in Databases (LiD), Italie Rome", 2008, http://hal.inria.fr/inria-00275491/en/.

[24] I. TELLIER. *How to Split Recursive Automata*, in "ICGI LNAI, France St Malo", vol. 5278, Springer Verlag, Alexander Clark, François Coste, Laurent Miclet, 2008, p. 200-212, http://hal.inria.fr/inria-00341770/en/.

[25] S. TISON, G. GODOY, S. MANETH. *Classes of Tree Homomorphisms with Decidable Preservation of Regularity*, in "FOSSACS'08, Hongrie", vol. 4962, Springer Verlag, 2008, p. 127-141, http://hal.archives-ouvertes.fr/hal-00243123/en/.

[26] M. TOMMASI, R. GILLERON, P. SENELLART, A. MITTAL, D. MUSCHICK. *Automatic Wrapper Induction from Hidden-Web Sources with Domain Knowledge*, in "Web information and data managment, États-Unis d'Amérique Napa", ACM, 2008, p. 9-16, http://hal.inria.fr/inria-00337098/en/.

Research Reports

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

References in notes

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

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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, n° XI, Springer-Verlag, 18th International Conference, ALT 2007, Octobre 2007, p. 242–256.

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

[36] 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.

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

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

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

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

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

[48] H. MIYASHITA, M. MURATA. *Composable XML transformations with tree transducers*, 2005.

[49] 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.

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

[51] 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.

[52] 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.

[53] 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.