INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team MOSTRARE

Modeling Tree Structures, Machine Learning, and Information Extraction

Futurs

THEME SYM

Table of contents

1. Team

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

Head of project team
Rémi Gilleron [professor, HdR]

Vice-head of project team
Joachim Niehren [senior researcher (DR2), UR Futurs, HdR]

Administrative assistant
Karine Lewandowski [shared by 3 projects]

Research scientist
Anne-Cécile Caron [assistant professor]
Aurélien Lemay [assistant professor]
Yves Roos [assistant professor]
Isabelle Tellier [assistant professor, HdR]
Sophie Tison [professor, HdR]
Marc Tommasi [assistant professor, HdR]
Fabien Torre [assistant professor]

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

Ph. D. student
Florent Jousse [INRIA and Région Nord-Pas-de-Calais fellowship, from October 2004 to October 2007]
Patrick Marty [INRIA and Région Nord-Pas-de-Calais fellowship, from October 2003 to November 2007]
Jérôme Champavère [MESR fellowship, since October 2006]
Emmanuel Filiot [INRIA and Région Nord-Pas-de-Calais fellowship, since October 2005]
Olivier Gauwin [INRIA Cordi fellowship, since November 2006]
Edouard Gilbert [ENS Cachan since November 2007]
Damien Poirier [CIFRE FRANCE TELECOM since November 2007]

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]

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

- Mostrare's first paper at the ACM Conference on Principles of Databases (PODS'07) by Filiot *et al* crowns our fundamental research activities on XML query languages.
- Jousse's thesis pioneers novel statistical learning techniques in XML applications. To the best of our knowledge, it is the first thesis on conditional random fields in France.
- The R^{2}S^2 software for automatic generation of personnalized RSS feeds is made available as a Web service. It combines two Mostrare tools (SQUIRREL and TREECRF) in a previously unexpected manner into an adaptative system for Web information extraction, a main goal the project.

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 [44], Xtatic [42], [47], and CDuce [31], [32], [35]. 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 [46], [56]. 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 [50], [48].

The automata community usually approaches tree transformations by tree transducers [40], i.e., tree automata producing output structure. Macro tree transducers, for instance, have been proposed recently for defining XML transformations [48], [51]. From the view point of logics, tree transducers have been studied for MSO definability [41].

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 [36], [52]. 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 [43], [45]. 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 [39].

Probabilistic context free grammars (pCFGs) [49] are used in the context of PDF to XML conversion [37]. 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 [53]). 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 [55], [54]. 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 [38] 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: Aurélien Lemay [correspondent], Matthieu Keith, Patrick Marty, Marc Tommasi, Fabien Torre, Missi Tran.

The MIELE project is in the final stage of development. The main goal of this project is to create an extensible Web Services 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 existing wrapper construction algorithms: SQUIRREL containing methods based on query induction using grammatical inference and PAF containing methods based on statistical classification. This project is a part of a platform for Semantic Web, a project of a broader scope developed together with other participants of the WEBCONTENT project.

5.2. TreeCRF: conditional random fields for trees

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

Participants: Florent Jousse [correspondent], Missi Tran, Marc Tommasi.

The TREECRF library is available at http://treecrf.gforge.inria.fr/. It allows to define `CRF` for `XML` trees providing automatic generation of features from pairs (`XML` input tree, its labeling) or allowing to enter user-defined features. Efficient implementations for inference and training algorithms are provided in the library. Once a `CRF` for `XML` trees is defined and trained, efficient algorithms for labeling `XML` trees are available. It should be noted that TREECRF has been used in the $\mathtt{R}^2\mathtt{S}^2$ system (see below) and in an automatic wrapper induction system from hidden-Web sources with domain knowledge (joint work with the GEMO project-team).

5.3. R2S2: automatic generation of RSS feeds

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

Participants: Florent Jousse [correspondent], Missi Tran, Marc Tommasi.

$\mathtt{R}^2\mathtt{S}^2$ (Really Really Simple Syndication), a tool for automatic generation of RSS feeds, has been developed using the TREECRF library. This tool allows to easily create a script generating RSS feeds from the contents of a website. The script is created using an annotation of a sample page from the website. Because the TREECRF algorithms assume the annotation to include all instances of the extracted information on the page, to simplify the annotation procedure for long pages SQUIRREL algorithms have been incorporated to automatically extend a partial annotation. $\mathtt{R}^2\mathtt{S}^2$ is equipped with an intuitive user interface (created in Google Web Toolkit) and is available for use online: http://r2s2.futurs.inria.fr.

6. New Results

6.1. Modeling XML document transformations

6.1.1. XML database queries, logic and automata

Keywords: *XPath*, *XQuery*, *n-ary node selection queries*, *tree transformations*.

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

Filiot, Niehren, Talbot[1], and Tison [21] distinguish a fragment of Core XPath 2.0 that we call the polynomial-time path language (PPL). XPath 2.0 is a recent recommendation of W3C that is a fragment of XQuery 1.0. The intension of XPath 2.0 is to enrich XPath 1.0 by constructs that turn it into a self contained query language with the same expressiveness as first-order logic. Variables in XPath 2.0 are fundamental for selecting n-tuples of nodes in trees. The navigational core of XPath 2.0 is known to capture first-order logic while being PSPACE complete with respect to model checking. Filiot et. al show that PPL remains first-order complete while enjoying polynomial time query answering (and thus model checking).

Gauwin, Caron, Niehren, Roos, and Tison have started a new PhD project on streaming query answering. The starting point is n-ary queries defined by tree automata as investigated in the previous Mostrare PhD projects of Planque and Filiot. As a first contribution, they introduce a new notion of tree automata that they call *streaming tree automata*. They relate it to a Alur's (2007) model of nested word automata, and to Neumann and Seidl's (1998) model of pushdown forest automata. All three models are shown to do the same in principle, but on different data structures (unranked trees, forests, nested words). This leads to subtle technical difference relevant to query formalisms. As second contribution [23] they study earliest query answering algorithms on basis of streaming tree automata. Earliest query answering is essential for efficient memory management. It is shown that the problem is DEXPTIME complete for queries defined by nondeterministic automata, while being in polynomial time in the deterministic case.

Gilleron, Tommasi, Tison et al.[2] have revised the book *Tree Automata Technique and Applications* [10]. This book has become THE international standard reference on tree automata; it is available for free on the Web since its first publication in 1997. The revised version contains a new chapter on tree automata for unranked trees as needed in XML database theory. This subsumes recent results by the Mostrare project on stepwise tree automata. Martens and Niehren [17] study minimization of XML Schema and tree automata for unranked trees. First, they investigate hedge automata that are most popular in the XML community. They show that minimal bottom-up and horizontally deterministic automata are not unique and that minimization is NP complete. Second, they study stepwise tree automata for unranked trees invented in the Mostrare project in 2005, and show that this yields unique deterministic automata, that can be computed efficiently.

Filiot, Talbot and Tison [22] adapt the spatial logic TQL[3] to the context of unranked trees over an infinite alphabet. Dealing with an infinite alphabet allows them to consider data-values (text content, attributes, etc...) which are usually ignored. This logic extends the tree-pattern matching language of CDuce. It allows for instance to test equalities of whole subtrees. They prove the satisfiability problem to be decidable for several expressive fragments of TQL. They reduce this problem to emptiness of a new class of tree automata which allows global equality (and disequality) test of subtrees. Finally, they relate these automata to Monadic Second Order logic extended with tree isomorphism tests.

Staworko started his postdoc project on adapting queries to the changes in the schema of the document. He is developing a new approach in which the changes are identified using transducers, and consequently the transducers are composed with with the automata queries to obtaine the adapted query. Since there may be more than one transducer representing the changes in the schemata, to capture them all alignment graphs for pairs of regular expressions are studied. Alignment graphs generalize the trace graphs which have been previously used to efficiently compute consistent answers to a query in a (possibly inconsistent) XML database. There an answer is a consistent if it is present in every repair of the database. Repairs are obtained by resolving inconsistencies, usually by removing all but one of the conflicting values. Staworko et. al. [4] [30] investige adding user priorities on how particular conflicts should be resolved in order to restrict the number of repairs considered when computing the consistent answers, and in consequence refine further the query evaluation.

[1] J.M. Talbot was a member of Mostrare until 2006 and is now professor in Marseille
[2] With co-authors from Aachen, (C. Löding) Paris (H.Comon, F. Jacquemart), Lille (M.Dauchet) and Marseille (D. Lugiez)
[3] TQL was proposed previously by Cardelli and Ghelli [34]
[4] Colaboration with J. Chomicki from the Buffalo University, Staworko's PhD supervisor

Erk and Niehren [16] study conjunctive queries in ranked trees with respect to satisfiablity. They show how to express dominance constraints in the once-only nesting fragment of stratified context unification, which therefore is NP-complete.

André, Caron, Debarbieux, Roos, and Tison [13] study the implication problem for path inclusion constraints in the context of semi-structured databases, an important problem in the context of query optimization and query approximation. The authors propose new improved algorithms for deciding the implication problem and introduce the notion of exact model of a set of path inclusion constraints $\mathcal{C}$: a model that satisfies $\mathcal{C}$ and also all constraints implied by $\mathcal{C}$. The authors also provide a decidable characterization of a class of sets of path inclusion constraints which have a finite exact model that can be effectively computed. Thus, deciding the implicaion for for this class of sets of constraints is reduced to evaluating queries over the exact finite model.

6.1.2. Programming languages

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

Participants: Joachim Niehren [correspondent], Sophie Tison.

Niehren et al.[5] [25] present an extension of the stochastic π-calculus by concurrent objects. The pure stochastic π-calculus was proposed by Regev and Shapiro (2000), Cardelli (2004), and Priami (2005) as a promising modeling language for systems biology. The implicit idea there is to map molecular actors in systems biology to communicating objects in the π-calculus. Kuttler, Lhoussaine and Niehren show in their new contribution, how to extend the stochastic π-calculus by an explicit notions that support static inheritance. They show that the resulting language supports many programming techniques, that are highly wishful for concrete modeling studies in systems biology. They provide a stochastic semantics, that defines the communication speed of concurrent objects with respect to real time. Finally, they show how to compile the object extension down to the basic stochastic π-calculus without any objects.

Niehren et al.[6] [26] present an observational semantics for a concurrent lambda calculus with reference cells and futures. This calculus, called lambda(fut), was introduced for modeling the operational semantics of the concurrent higher-order programming language Alice ML. It is a minimalist extension of the call-by-value lambda-calculus that is sufficiently expressive to define and combine a variety of standard concurrency abstractions, such as channels, semaphores, and ports. Niehren et al. prove for the first time, that call-by-value beta reduction is correct as a program transformation for concurrent languages such as lambda(fut), while general beta reduction is not.

Tison et. al.[7] [24] investigate the decidability of the properties of normalization and unique normalization for two sytactically restricted classes of term rewriting systems (TRS): (i) shallow right-linear TRS, and (ii) linear right-shallow TRS. They show that the normalisation property, i.e. whether all terms can reach a normal form, is decidable for both cases. The unique normalization property, i.e whether every term can reach at most one normal form, is shown to be undecidable for linear right-shallow TRS and remains unknown for shallow right-linear TRS. The authors show, however, that for shallow right-linear TRS if the normalization property is satisfied then, the unique normalization property becomes decidable too.

6.2. Machine learning for XML document transformations

6.2.1. Wrapper induction by grammatical inference

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

Participants: Jérôme Champavère, Rémi Gilleron, Aurélien Lemay [correspondent], Joachim Niehren, Isabelle Tellier.

[5] Cooperation with Céline Kuttler from the Microsoft Research and University of Trento Research Center on Computational Systems Biology, and Cédric Lhoussaine from the LIFL in Lille

[6] Cooperation with D. Sabel and M. Schmidt-Schauß from the Goethe University in Frankfurt and with J. Schwinghammer from the Saarland University in Saarbrücken

[7] Cooperation with G. Godoy from the Polytechnical University Barcelona

Champavère, Gilleron, Lemay, and Niehren [19] started a PhD project towards schema guided query induction. The aim is to incorporate the document schema information into existing algorithms for learning tree automata queries. 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. The most basic schema of interest is the DTD of XHTML. More precise schemas can be inferred by existing algorithms for schema learning from positive examples (for instance by Neven et al [33]).

Champavère et. al. [27] present an efficient algorithm for testing language inclusion for deterministic tree automata and DTDs. This is the most fundamental algorithmic problem for schema guided query induction. The algorithm has to check incrementally, whether the current hypothesis automaton is consistent in the given schema. The new algorithm for testing language inclusion $L(A) \subseteq L(B)$ between tree automata operates in time $O(|A| * |B|)$. It is extended to an inclusion test for tree automata for unranked trees A and deterministic DTDs D in time $O(|A| * |\Sigma| * |D|)$. No previous algorithms with these complexities existed.

Tellier et at.[8] [14] investigate learning of Pregroup Grammars by methods from grammatical inference. A boader survey on grammatical inference in the field of human language aquisition is presented in [18].

6.2.2. Probabilistic XML tree labeling

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

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

Gilleron, Jousse, Tellier, and Tommasi adapt conditional random fields for `XML` trees. They extend the `CRF` model by introducing constraints to reduce the complexity of training algorithms. They also define combination techniques based on domain knowledge such as schemas. Tree transformations can be defined by tree labelings with tree edit operations. `CRF` are used for different tasks such as Web information extraction and document transformation [28]. `CRF` for `XML` trees have been used to define a Web service for the automatic generation of personalised `RSS` feeds (see the software section).

Muschick (Erasmus student from Graz Universität, master project in Mostrare), Gilleron, and Tommasi adapt `CRF` for `XML` trees to learn in an unsupervised way [29]. They allow the system to learn from imperfect and imprecise annotations obtained from domain knowledge. They also design bootstrapping techniques. This system is integrated in an automatic wrapper induction system from hidden-Web sources with domain knowledge.[9]

Gilbert, Gilleron, and Tommasi begin to study probability distributions over free algebras of trees. Distributions can be defined by weighted tree automata or by tree series. They adapt definitions to handle the case of unranked trees. In close cooperation with H. DENIS and A. HABRARD from the LIF-MARSEILLE, they define learning algorithms for probability distributions for unranked trees. Preliminary results are proposed in [20].

Kong centers his research on approximate tree pattern matching. There are two major tasks: searching for and similarity measuring of tree-structured data, especially rooted labeled tree. His recent work focuses on the latter. Based on the lowest common ancestor (LCA) concept, he proposes XTreeRank method for generic unordered rooted labeled tree, and KCAM (Keyword Common Ancestor Matrix) method for tightest unordered rooted labeled tree, which means the root node is the only LCA for all its leaf nodes. The test data is XML, but the methods proposed can also be used in other related applications, such as BioInformatics, Image retrieval. Besides this, the other work is about combining similarity measuring with tree pattern searching together, which has not been studied enough in streamed XML processing.

[8] Cooperation with D. Béchet and A. Foret through the INRIA cooperative research action (ARC) Gracq (Lille, Nantes, Nancy, and Rennes)

[9] This is joint work with Mittal and Senellart from GEMO project-team of Inria Futurs, Saclay.

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 library and the R^{2}S^2 software are developed in the project.

7.1.2. RNTL Webcontent

Participants: Rémi Gilleron, Florent Jousse, Patrick Marty, 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

We make a research overview in the field of automatic maintenance of wrappers for the UpandNet company.

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

8. Other Grants and Activities

8.1. French Actions

8.1.1. ANR Blanc Enumeration: 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.

8.1.2. ARA MDCO CROTAL: 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://www.grappa.univ-lille3.fr/~tellier/crotal.html.

8.1.3. ARA MDCO Marmota: Stochastic Tree Models and Stochastic Tree Transformation

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://www.grappa.univ-lille3.fr/marmota.

8.1.4. ACI TraLaLA: Transformation Languages, Logic and Application

Participants: Anne-Cécile Caron [correspondent], Emmanuel Filiot, Joachim Niehren, Yves Roos, Sophie Tison.

We are involved in the French cooperation project “ACI masse de données – TraLaLA – XML Transformation Languages, Logic and Application” (2004–2007). We pay particular attention to the programming languages and query languages problems. We aim to cover in a uniform way a wide spectrum of different areas, namely: programming languages (expressiveness, typing, new programming primitives, query underlying logics, logical optimization), data access (streamed data, compression, access to secondary memory storages, persistency engines), implementation (pattern matching compiling, physical optimization, sub-typing verification, execution models for streamed data).

Ours partners are: Giuseppe CASTAGNA (coordinator - LIENS), Luc SÉGOUFIN (GEMO INRIA project), Silvano DAL ZILIO (LIF) and Véronique BENZAKEN (LRI). 2005-2007. More information about the project can be found on http://www.cduce.org/tralala.html.

9. Dissemination

9.1. Scientific animation

- **Program Committees:**

  R. GILLERON was PC member of EGC'2007 (french conference on knowledge discovery), PC member of ECML'2007 workshop Graph Labelling Workshop and Web Spam Challenge

  J. NIEHREN was PC member of FROCOS'2007 (International Symbosium on Frontiers in Combining Systems) and TDMM'2007 (ICDE Workshop on Text Data Mining and Management)

  S. TISON was member of the editorial board of RAIRO- THEORETICAL INFORMATICS AND APPLICATIONS, was PC member of PLANX'2007, RTA'2007, TALE'2007 and FOSSACS'2007.

  I. TELLIER was PC member of CORIA'2007 (french conference on information retrieval), was PC member of the EMNLP-CONLL'2007 conference, PC member of workshop TALN'2007 "higher order syntactic formalisms" and member of the Redaction Committee of the French journal TAL.

  F. TORRE was PC member of Conférence Francophone sur l'Apprentissage Automatique CAP'2007.

- **Workshop Organization**

  I. TELLIER organized a one-day workshop "natural language learning : observations and models" which occurred in Lille3, June 21. http://www.grappa.univ-lille3.fr/~tellier/journeeMSH.html

- **Invited talks**

  J. NIEHREN presented in

  – the Logikseminar of the Saarland University in Saarbrücken on *learning n-ary queries in trees for Web information extraction*, and in
  – the Informatikseminar of the Goethe University in Frankfurt. on *XPath dialect with variables*.

- **French Scientific Responsibilities**

  R. GILLERON is head of the research group GRAPPA on machine learning in Lille, member of the scientific council for the program ARA - MDCO de l'ANR.

  J. NIEHREN was a member of the CR2 selection committee of INRIA Futurs in Bordeaux.

  I.TELLIER was member of the CNU 27 (national committe for the evaluation of assistant professors and professors in computer science).

  S. TISON is vice-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 is member of the national evaluation committee (MSTP-DS9) for teaching and research.

  A.-C. CARON was member of CSE (Commission de Spécialistes de l'Enseignement supérieur) of University of Valenciennes, is member of CNU (Conseil National des Universités)

9.2. Teaching and scientific diffusion

- TEACHING

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

- MASTER LECTURES PRESENTED AT THE UNIVERSITY OF LILLE 1
  - Logic and Modelisation: A.-C. CARON, J. NIEHREN, and S. TISON
  - Machine Learning for Information Extraction: M. TOMMASI
  - Supervised Classification: M. TOMMASI
  - 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:
  - G. LAURENCE, natural language processing with XCRF, supervised by I. TELLIER.
  - D. MUSCHICK, unsupervised learning of XML tree annotations, supervised by R. GILLERON and M. TOMMASI.
- DIRECTION OF PHD THESIS SUBMITTED IN 2007:
  - F. JOUSSE, learning XML tree transformations with probabilistic models for labeling, October 31th, supervised by R. GILLERON, I. TELLIER and M. TOMMASI.
  - P. MARTY, wrapper induction for table extraction from Web documents, December 4th, supervised by R. GILLERON, M. TOMMASI and F. TORRE.
- HABILITATION THESIS IN 2007:

  S. TISON belonged to the committee of M. HALFELD FERRARI ALVES (Blois).
- PHD COMMITTEES:

  J. NIEHREN belonged to the committees of S. THATER (Saarbrücken, reviewer), M. SAMUELIDES (Paris VII, reviewer), and S. RAEYMAEKERS (Leuven, reviewer).

  R. GILLERON belonged to the committees of L. ORSEAU (Rennes) and T. URRUTY (Lille).

  S. TISON belonged to the committees of C. KUTTLER (Lille), R. DRIDI (Lille), N. IHADDADENE (Lille), J. GROSLAMBERT (Franche-Comté), and P. PILLOT (Orléans, reviewer).

  M. TOMMASI belonged to the committee of G. WISNIEWSKI (Paris VI, 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, http://dx.doi.org/10.1016/j.tcs.2007.05.010.

[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] F. DENIS, R. GILLERON, F. LETOUZEY. *Learning from Positive and Unlabeled Examples*, in "Theoretical Computer Science", vol. 348, n^{o} 1, 2005, p. 70-83.

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

[7] 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.

[8] 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.

[9] 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

Books and Monographs

[10] H. COMON, M. DAUCHET, R. GILLERON, C. LÖDING, F. JACQUEMARD, D. LUGIEZ, S. TISON, M. TOMMASI. *Tree Automata Techniques and Applications*, release October, 12th 2007. The first version appeared 1997, October 2007, http://www.grappa.univ-lille3.fr/tata.

Doctoral dissertations and Habilitation theses

[11] F. JOUSSE. *Transformations d'arbres XML avec des modèles probabilistes pour l'annotation*, Ph. D. Thesis, Université Charles de Gaulle, Lille 3, 2007.

[12] P. MARTY. *Induction d'extraction n-aire pour les documents semi-structurés*, Ph. D. Thesis, Université Charles de Gaulle, Lille 3, 2007.

Articles in refereed journals and book chapters

[13] 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, http://dx.doi.org/10.1016/j.tcs.2007.05.010.

[14] D. BÉCHET, A. FORET, I. TELLIER. *Learnability of Pregroup Grammars*, in "Studia Logica", n^{o} 87, november-december 2007, p. 225–252.

[15] 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.

[16] K. ERK, J. NIEHREN. *Dominance Constraints in Stratified Context Unification*, in "Information Processing Letters", vol. 101, n^{o} 4, February 2007, p. 141-147, http://hal.inria.fr/inria-00094787/en.

[17] 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.

[18] I. TELLIER, D. BÉCHET, A. DIKOVSKY, A. FORET, Y. L. NIR, E. MOREAU, C. RÉTORÉ. *Modèles algorithmiques de l'acquisition de la syntaxe*, in "Recherches Linguistiques de Vincennes", 2007.

Publications in Conferences and Workshops

[19] J. CHAMPAVÈRE, R. GILLERON, A. LEMAY, J. NIEHREN. *Towards Schema-Guided XML Query Induction*, in "ICML-2007 Workshop on Challenges and Applications of Grammar Induction", 2007.

[20] F. DENIS, É. GILBERT, R. GILLERON, A. HABRARD, M. TOMMASI. *On Probability Distributions for Trees: Representations, Inference and Learning*, in "poster in NIPS Workshop on Representations and Inference on Probability Distributions", 2007.

[21] 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.

[22] 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.

[23] O. GAUWIN, A.-C. CARON, J. NIEHREN, S. TISON. *Complexity of Earliest Query Answering with Streaming Tree Automata*, in "ACM Workshop on Programming Language Techniques for XML (Plan-X)", To appear, 2008.

[24] G. GODOY, S. TISON. *On the Normalization and Unique Normalization Properties of Term Rewrite Systems*, in "International Conference on Automated Deduction", Lecture Notes in Computer Science, vol. 4603, Springer Verlag, 2007, p. 247-262, http://www.lsi.upc.es/~ggodoy/papers/cade2007.ps.

[25] C. KUTTLER, C. LHOUSSAINE, J. NIEHREN. *A Stochastic Pi Calculus for Concurrent Objects*, in "Second International Conference on Algebraic Biology", Lecture Notes in Computer Science, n^{o} 4545, Springer Verlag, July 2007, p. 232-246, http://hal.inria.fr/inria-00121104/en.

[26] J. NIEHREN, D. SABEL, M. SCHMIDT-SCHAUSS, J. SCHWINGHAMMER. *Observational Semantics for a Concurrent Lambda Calculus with Reference Cells and Futures*, in "23rd Conference on Mathematical Foundations of Programming Semantics", Electronical notes in theoretical computer science, vol. 173, Elsevier, April 2007, p. 313-337, http://hal.inria.fr/inria-00128861/en.

Miscellaneous

[27] J. CHAMPAVÈRE, R. GILLERON, A. LEMAY, J. NIEHREN. *Efficient Inclusion Checking for Deterministic Tree Automata and DTDs*, Submitted, 2008.

[28] R. GILLERON, F. JOUSSE, I. TELLIER, M. TOMMASI. *Conditional Random Fields for XML Trees*, Submitted, 2008.

[29] P. SENELLART, A. MITTAL, D. MUSCHICK, R. GILLERON, M. TOMMASI. *Automatic Wrapper Induction from Hidden-Web Sources with Domain Knowledge*, Submitted, 2008.

[30] S. STAWORKO, J. CHOMICKI. *Priority-based Conflict Resolution in Inconsistent Relational Databases*, Submitted, 2008.

References in notes

[31] 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.

[32] 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.

[33] G. J. BEX, F. NEVEN, T. SCHWENTICK, K. TUYLS. *Inference of Concise DTDs from XML Data*, in "Proceedings of 32nd Conference on Very Large databases - VLDB", 2006, p. 115-126.

[34] L. CARDELLI, G. GHELLI. *TQL: a query language for semistructured data based on the ambient logic*, in "Mathematical Structures in Computer Science", vol. 14, n^{o} 3, 2004, p. 285–327.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

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

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

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

[45] 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.

[46] 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.

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

[48] 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.

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

[50] 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.

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

[52] 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.

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

[54] 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.

[55] 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.

[56] 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.