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 8022 *(*CNRS*, Lille 1 and Lille 3 universities)*

Head of the team
Rémi Gilleron [professor, University of Lille 3, HdR]

Administrative assistant
Karine Lewandowski [shared by 3 projects]

Staff member INRIA
Joachim Niehren [senior researcher (DR2), UR Futurs, HdR]

Staff member Lille 3 University
Aurélien Lemay [assistant professor]
Isabelle Tellier [assistant professor, HdR]
Marc Tommasi [assistant professor, HdR]
Fabien Torre [assistant professor]

Staff member Lille 1 University
Anne-Cécile Caron [assistant professor]
Yves Roos [assistant professor]
Jean-Marc Talbot [assistant professor until september 2006, HdR]
Sophie Tison [professor, HdR]

Ph. D. student
Iovka Boneva [MESR fellowship, from October 2002 to June 2006]
Laurent Candillier [CIFRE fellowship, from May 2003 to September 2006]
Jérôme Champavere [MESR fellowship, since October 2006]
Emmanuel Filiot [INRIA and Région Nord-Pas-de-Calais fellowship, since October 2005]
Olivier Gauwin [INRIA Cordi fellowship, since december 2006]
Florent Jousse [INRIA and Région Nord-Pas-de-Calais fellowship, since October 2004]
Patrick Marty [INRIA and Région Nord-Pas-de-Calais fellowship, since October 2003]

Technical staff
Mathieu Keith [junior engineer since november 2006]
Jean-Philippe Nirel [junior engineer from October 2005 to June 2006]
Missi Tran [engineer from January to June 2006]

Post-doctoral fellow
Denis Debarbieux [assistant professor until september 2006]

2. Overall Objectives

2.1. Overall Objectives

The objective of MOSTRARE is to develop adaptive document processing methods for `XML`-based information systems. Adaptiveness imports when documents evolve frequently such as on the Web. The particularity of MOSTRARE is that we devellop 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 infering them. Once available, we want to integrate these learing algorithms into innovative information extraction systems, semantic Web platforms, and document processing engines.

The following two paragraphs summarize our two main research objectives:

Modeling tree structures for information extraction.

We wish to 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 generalize 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.

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 querying 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.

To illustrate the main difficulty of programming XML transformations, consider the example of PDF to XML conversion, under the assumption that the output's DTD is given [37]. In a first step, one can use an existing PDF to HTML converter. In a second step, it remains to convert HTML into XML. The DTD of the HTML input document, however, will be either unknown or uninformative. Furthermore, the input will contain errors that are to be accounted for.

Alternatives programming language for defining XML transformations have been proposed by the programming language community, for instance XDuce [46], Xtatic [44], [50], and CDuce [33], [34], [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 sublanguage 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 [48], [58]. 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 [53], [51].

The automata community usually approaches tree transformations by tree transducers [42], i.e., tree automata producing output structure. Macro tree transducers, for instance, have been proposed recently for defining XML transformations [51], [54]. From the view point of logics, tree transducers have been studied for MSO definability [43].

3.2. Machine learning for XML document transformations

Keywords: *annotation*, *grammatical inference*, *statistical learning*, *transformation*, *trees*, *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.

3.2.1. Grammatical inference

should be useful for inferring tree transducers that represent XML transformations. So far only very basic tree transducers have been shown to be learnable, by previous work of the Mostrare project [12]. These are node selecting tree transducer (NSTTs) which preserve the structure of trees while relabeling their nodes deterministically. Previous work on grammatical inference for transducers remains limited to the case of strings [36], [55]. The case of trees remains to be explored.

Stochastic tree transducers have been studied in the context of natural language processing [45], [47]. 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 [41].

3.2.2. 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.

Probabilistic context free grammars (pCFGs) [52] are used in the context of PDF to XML conversion [37], [38]. Such methods infer a CFG as a generative model on which to add probabilities in a second step. Such two step 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) [49]. One main advantage of CRF is to take into account long distance dependencies in the observed data. CRF have also been applied in many situations like in bioinformatics [39] for gene prediction. Essentially, CRF have been used to model sequences of words. CRF suppose a graph that relates conditional independence of random variables and features that are used to estimate conditional probabilities. CRF have been used for sequences with internal graph structure represented as a linear chain.

So called *structured output* has very recently become a hot research topic in machine learning [57], [56]. 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. For instance, let us consider the task of syntactic parsing which consists in finding the syntactic tree of a sentence. One classical way is to find the tree that maximizes the probability of the sentence - the tree is usually modeled by a pCFG. In the *structured output* point of view, this task is considered as a classical classification task, but the difference is that the label of the sentence is not a "simple" discrete value but its syntactic tree: the set of possible labels is the set of all the possibles syntactic trees. Giving a learning set of couples (tree,sentence), the structured output classification task consists in finding the tree-label of a new sentence.

4. Application Domains

4.1. Introduction

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 [40] survey novel integration tasks that appear with the Semantic Web and the usage of ontologies. Therefore, the semi-automatic generation of XML transformations is a challenge in the database community and in the semantic Web community.

Also, XML transformations are useful for document processing. For instance, there is need of designing transformations from documents organized w.r.t visual format (HTML, DOC, PDF) into documents organized w.r.t. semantic format (XML according to a DTD or a schema). The semi-automatic design of such transformations is obviously a very challenging objective.

4.2. A Web Service for Information Extraction

Keywords: *Web service*, *monadic queries*, *n-ary queries*, *wrapper induction*.

Participants: Aurélien Lemay [correspondent], Mathieu Keith, Patrick Marty, Fabien Torre.

A Web service for information extraction is currently under development. The Web service will be included in a platform for the semantic Web which was developped by all partners of the Webcontent project. Our Web service will include wrapper induction programs for monadic queries and n-ary queries. These programs correspond to SQUIRREL prototype and PAF prototype which were described in the previous MOSTRARE reports. The construction of a wrapper induction program will be interactive: the user can interact with the program by giving informations to be extracted or by correcting wrong extractions.

4.3. TreeCRF: conditional random fields for trees

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

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

TreeCRF is a stochastic system which allows to label element, attribute and text nodes of XML trees. It is available as a freely available JAVA library[1]. It provides automatic generation of features from pairs (XML input tree, its labeling) allowing to define the model. Efficient implementations for inference and training algorithms are provided in the library. After training, labelings of new XML input trees can be computed.

5. New Results

5.1. Modeling XML document transformations

5.1.1. Node Selection Queries

Keywords: *XPath*, *XQuery*, *automata*, *database theory*, *logic*, *semi-structured documents*.

Participants: Sophie Tison [correspondent], Emmanuel Filliot, Joachim Niehren, Jean-Marc Talbot, Olivier Gauwin, Anne-Cécile Caron.

XQuery is the W3C standard for defining tree transformation. Each expression of XQuery defines a composition of basic queries. Basic queries by FLWR expressions return the result of an n-ary node selecting query in some output tree. They rely on path expressions, storing selected nodes of the n-tuples in variables. The expressions for selecting n-tuples of nodes have been pushed down to XPath 2.0 very recently, which is a proper sublangage of XQuery 1.0.

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, Niehren, Talbot, and Tison [24], [23] distinguish a fragment of Core XPath 2.0 that we call the polynomial-time path language (PPL). They show that PPL remains first-order complete even though enjoying polynomial time query answering (and thus model checking).

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

Monadic second-order (MSO) logic is more expressive than FO and thus XPath 2.0. The famous theorem of Thatcher and Wright (1968) states that tree automata can express the same queries than MSO. The traditional theorem holds with respect to ranked trees, but has been lifted to unranked trees as in XML meanwhile. It is also well known that n-ary queries represented by deterministic tree automata can be answered in polynomial time.

Martens and Niehren [17] study minimization of XML Schema and tree automata for unranked trees. First, they study unranked tree automata that are standard in database theory, assuming bottom-up determinism and that horizontal recursion is represented by deterministic finite automata. They show that minimal automata in that class are not unique and that minimization is np complete. Second, they study more recent automata classes that do allow for polynomial time minimization. Among those, we show that bottom-up deterministic stepwise tree automata (invented in the Mostrare project) yield the most succinct representations. Third, they investigate abstractions of ML schema languages. In particular, they show that the class of one-pass preorder typable schemas allows for polynomial time minimization and unique minimal models.

Erk and Niehren [13] 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.

5.1.2. Programming Languages

Keywords: *Concurrency*, *functional programming*.

Participant: Joachim Niehren [correspondent].

Kuttler, Lhoussaine, and Niehren [14], [30] propose to model the dynamics of gene regulatory networks as concurrent processes in the stochastic pi calculus. As a first case study, they show how to express the control of transcription initiation at the lambda switch, a prototypical example where cooperative enhancement is crucial. This requires concurrent programming techniques that are new to systems biology, and necessitates stochastic parameters that we derive from the literature.

Niehren, Schwinghammer and Smolka [18] introduce a new lambda calculus with futures, Lambda(fut), that models the operational semantics of concurrent statically typed functional programming languages with mixed eager and lazy threads such as Alice ML, a concurrent extension of Standard ML. Lambda(fut) 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.

5.2. Machine learning for XML document transformations

5.2.1. Wrapper induction by grammatical inference

Keywords: *grammatical inference*, *monadic queries*, *ordered trees*, *tree automata*, *wrapper induction*.

Participants: Aurélien Lemay [correspondent], Rémi Gilleron, Joachim Niehren, Yves Roos, Jérôme Champavère.

Carme, Gilleron, Niehren, and Lemay investigate wrapper induction for Web information extraction by methods of grammatical inference. They consider Web documents in HTML as unranked ordered trees, and wrappers – the extraction target – as node selection queries in unranked trees. Users of a Web information extraction system are supposed to annotate example HTML documents, visually by the help of some Web browser. They may label informative nodes positively and others negatively. The tasks of the extraction system is then to infer the correct node selection query from the sample of annotated examples.

In [12], Carme, Gilleron, Lemay, and Niehren turn their induction algorithm for monadic queries into a visually interactive learning process that can also deal with document with just a few annotation (complete annotations are no longer required). Experiments on realistic Web documents confirm excellent quality with very few user interactions – annotations and corrections – during wrapper induction.

In [32], Lemay, Niehren and Gilleron consider n-ary queries. They propose an induction algorithm based on grammatical inference techniques. Preliminary experimental results are quite promising. Nevertheless, the work will be pursued to introduce pruning techniques and heuristics in order to get an even more efficient system and to allow the interactive design of n-ary wrappers.

Latteux, Lemay, Roos and Terlutte [15], [16], [31] study learning of finite automata from positive examples. They consider Residual Finite State Automata (RFSA) which are non deterministic automata that share some properties with DFA (in particular, DFA are RFSA and RFSA can be much smaller). Latteux, Lemay, Roos and Terlutte introduced the class of biRFSA which are RFSA whose reverse are RFSA. This class is not learnable in general but they identified two non trivial subclasses that are learnable, the second one being learnable in polynomial time.

5.2.2. Statistical wrapper induction

Keywords: *HTML data*, *attribute-value representation*, *supervised classification*, *textual data*, *wrapper induction*.

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

Gilleron, Marty, Tommasi, and Torre approach wrapper induction by statistical machine learning techniques within Marty's PhD project. they have defined a system PAF to extracting n-ary queries in tree structured documents. The system is based on combination techniques. In [26], they have extended PAF to an interactive system allowing to define, with very few interactions with the user, n-ary queries over `HTML` Web pages. It is worth noting that the system can be applied whatever is the organization of the target n-ary relation in the input Web page.

5.2.3. Statistical clustering

Keywords: *Expectation-Maximization*, *subspace clustering*, *unsupervised classification*.

Participants: Fabien Torre [correspondent], Isabelle Tellier, Laurent Candillier.

Laurent Candillier has defended his PhD thesis in september 2006 [10]. His work has achieved two main results. First, a new subspace clustering algorithm for attribute-value databases has been defined. This algorithm has been tested on many problems in which it has been proved to perform very well [22]. It could also be applied to semi-structured data, after an appropriate encoding of XML data. The algorithm participated in the 2005 INEX/PASCAL challenge on document mining, where it has been classed second out of six in clustering. An adaptation of decision-tree learning algorithms applied to the same encoding for the supervised learning task, tested by the same authors, has been classed first in classification. The participants of the challenge having obtained the best results have written the chapter "Mining XML documents" of the book "Data Mining Patterns : New Methods and Applications", accepted to appear next year (the co-authors of this chapter are Laurent Candillier, Ludovic Denoyer Patrick Gallinari, Marie-Christine Rousset, Alexandre Termier and Anne-Marie Vercoustre). The second main achievement of Candillier and co-authors' work is the proposition of a new evaluation method for non supervised-learning. This method proposes to consider clustering as a pre-treatment for a task (for example supervised learning) which can be rigorously evaluated. The comparison of how the task is performed with or without the clustering as a pre-treatment mesures the information this clusering has brought. Many Experiments have proved that this method is robust and largely domain independant [21], [20].

5.2.4. Probabilistic XML tree labeling

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

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

Conditional random fields are graphical models defining conditional probability distributions. They have been successfully applied for labeling tasks in the case of sequences. We have defined XML Conditional Random Fields, a framework for building conditional models for labeling XML documents in [28], [29]. We have defined efficient algorithms for inference and parameter estimation. A prototype TreeCRF has been implemented. We have applied XML Conditional Random Fields to tree labeling tasks in information extraction and schema matching. Experiments yield very good results.

6. Contracts and Grants with Industry

6.1. Contracts and Grants with Industry

6.1.1. 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. A PhD CIFRE supported by XRCE and supervised by Rémi GILLERON and Boris CHIDLOVSKI will begin in january 2007.

6.1.2. 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. We should integrate our Web service for information extraction, currently under development, in the platform. and adapt our prototypes for Web information extraction. We are also involved in academic works on the semantic web: semantic annotations, ontology inference and ontology mapping.

6.1.3. Others

The PhD of Laurent Candillier was supported by the company PERTINENCE in Paris.

7. Other Grants and Activities

7.1. French Actions

7.1.1. 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, subtyping verification, execution models for streamed data). The marginal budget allocated to the Mostrare project is 53 Keuros over the period 2004-2007.

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

7.1.2. ARA MDCA 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). More information about the project can be found on http://www.grappa.univ-lille3.fr/marmota.

7.1.3. ARC Mosaique

Participant: Isabelle Tellier [correspondent].

This ARC (Common Research Action between INRIA projects) gathers several French teams working on the syntactic formalisation of natural language. Some of them have developped syntactic ressources, but the problem faced is that these ressources are neither comparable (because they are based on different grammatical formalisms) nor reusable by any other formalism than their own. None of them have a very large covering. So, the purpose of this project is to capitalize as much as possible the efforts already made, by developping bridges between various formalisms, or by proposing some higher level formalism, able to generalize several others. The first year of the ARC has mainly been dedicated to an exhaustive presentation of the available ressources, and of the particularities of each of them. The use of XML formats and tree models in this context links this project with Mostrare's goals.

8. Dissemination

8.1. Scientific Animation

- **Program Committees:**

  S. TISON was member of the editorial board of RAIRO - Theoretical Informatics and Applications, was PC member of LPAR'2006,PLANX'2007, and FOSSACS'2007.

  R. GILLERON was PC member of EGC'2006 and EGC'2007 (french conference on knowledge discovery)

  F. TORRE was PC member of CAP'2006 (french conference on machine learning)

  I. TELLIER was PC member of CORIA'2007 (french conference on information retrieval); was member of the editorial committee for the special issue "TAL: systemes question-reponse".

  J. NIEHREN was PC member of MFCS'2006, LPAR2006, ROMAND'2006 and CSLP'2006.

- **Workshop Organization**

  Anne-Cécile CARON co-organizes EGC'2006 (French Conference on knowledge discovery) and BDA'2006 (French Conference on databases) in Lille.

  Sophie TISON co-organizes a Workshop on Tree Automata, funded by the European Science Foundation, in Bonn (june 2006).

- **Invited talks**

  Joachim Niehren was invited to the dagstuhl seminar on constraint satisfaction

  Rémi Gilleron, Jean-Marc Talbot and Joachim Niehren were invited for presentation at the tree automata workshop in Bonn.

- **French Scientific Responsibilities**

  S. TISON is, vice-director of the LIFL (computer science department in Lille), head of the research group STC of the LIFL. She is member of the national evaluation committee (MSTP-DS9) for teaching and research.

  R. GILLERON is member of the scientific council for the program ARA - MDCA de l'ANR.

  I. TELLIER is member of the CNU 27 (french evaluation committee for assistant professors in computer science)

8.2. 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	bachelor and 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 et Modelisation: A.-C. CARON, J. NIEHREN, and S. TISON
  - Machine Learning for Information Extraction: I. TELLIER (2006-07)
- MASTER PROJECTS:
- DIRECTION OF PHD THESIS SUBMITTED IN 2006:
  - I. BONEVA, Logics for unranked and unordered trees, supervised by J. M. TALBOT and S. TISON.
  - L. CANDILLIER, on unsupervised learning by subspace clustering, supervised by I. TELLIER and F. TORRE.
- HABILITATION THESIS IN 2006:
  - M. TOMMASI, Machine Learning for tree structures.
- PHD COMMITTEES:

  R. GILLERON belonged to the committee of L. CANDILLIER; I. TELLIER belonged to the committees of L. CANDILLIER, E. MOREAU (Nantes), and R. EYRAUD (Saint Etienne, reviewer); S. TISON belonged to the committees of D. MARCHAL, A. MULLER, J. LEMESRE, I. BONEVA, and C. MIACHON (Orsay, reviewer);
- Habilitation committees: R. GILLERON belonged to the committees of M. TOMMASI (Lille), F. YVON (Paris); S. TISON belonged to the committee of M. TOMMASI (Lille).

9. Bibliography

Major publications by the team in recent years

[1] Y. ANDRÉ, A.-C. CARON, D. DEBARBIEUX, Y. ROOS, S. TISON. *Extraction and Implication of Path Constraints*, in "Proceedings of the 29th Symposium on Mathematical Foundations of Computer Science (MFCS'04)", Lecture Notes in Computer Science, vol. 3153, Springer Verlag, 2004, p. 863-875.

[2] I. BONEVA, J.-M. TALBOT. *When Ambients Cannot be Opened*, in "Theoretical Computer Science", vol. 333, n^{o} 2, 2005, p. 127-169.

[3] 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.

[4] 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.

[5] J. CARME, R. GILLERON, A. LEMAY, J. NIEHREN. *Interactive Learning of Node Selecting Tree Transducers*, in "Machine Learning", Appears 2007, vol. 66, n^{o} 1, 2006, p. 33-67.

[6] 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.

[7] 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", 2006.

[8] W. MARTENS, J. NIEHREN. *On the Minimization of XML Schemas and Tree Automata for Unranked Trees*, in "Journal of Computer and System Science", In press 2007, 2006.

Year Publications

Doctoral dissertations and Habilitation theses

[9] I. BONEVA. *Logics for unranked and unordered trees and their use for querying semistructured data*, Ph. D. Thesis, Universite Lille 1, 2006.

[10] L. CANDILLIER. *Contextualisation, Visualisation et Evaluation en Apprentissage Non Supervise*, Ph. D. Thesis, Universite Charles de Gaulle, Lille 3, 2006.

[11] M. TOMMASI. *Habilitation thesis: Machine Learning for Tree Structures*, Ph. D. Thesis, Universite Charles de Gaulle, Lille 3, 2006.

Articles in refereed journals and book chapters

[12] 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.

[13] K. ERK, J. NIEHREN. *Dominance Constraints in Stratified Context Unification*, in "Information Processing Letters", in press, 2007.

[14] C. KUTTLER, J. NIEHREN. *Gene Regulation in the Pi Calculus: Simulating Cooperativity at the Lambda Switch*, in "Transactions on Computational Systems Biology", vol. 4230, n^{o} VII, 2006, p. 24-55.

[15] M. LATTEUX, A. LEMAY, Y. ROOS, A. TERLUTTE. *Identification of biRFSA languages*, in "Theoretical Computer Science", vol. 356, n^{o} 1-2, 2006, p. 212-223.

[16] M. LATTEUX, Y. ROOS, A. TERLUTTE. *Minimal NFA and biRFSA Languages*, in "RAIRO - Theoretical Informatics and Applications", To appear, 2007.

[17] W. MARTENS, J. NIEHREN. *On the Minimization of XML Schemas and Tree Automata for Unranked Trees*, in "Journal of Computer and System Science", in press, 2007.

[18] J. NIEHREN, J. SCHWINGHAMMER, G. SMOLKA. *A Concurrent Lambda Calculus with Futures*, in "Theoretical Computer Science", vol. 364, n^{o} 3, 2006, p. 338-356.

[19] I. TELLIER. *Learning Recursive Automata from Positive Examples*, in "Revue d'Intelligence Artificielle", vol. New Methods in Machine Learning, n^{o} 20/2006, 2006, p. 775-804.

Publications in Conferences and Workshops

[20] 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.

[21] L. CANDILLIER, I. TELLIER, F. TORRE, O. BOUSQUET. *Evaluation en Cascade d'Algorithmes de Clustering*, in "8eme Conference francophone sur l'Apprentissage automatique (CAp'2006)", 2006, p. 109–124.

[22] L. CANDILLIER, I. TELLIER, F. TORRE, O. BOUSQUET. *SuSE: Subspace Selection embedded in an EM algorithm*, in "8eme Conference francophone sur l'Apprentissage automatique (CAp'2006)", 2006, p. 331–345.

[23] E. FILIOT, J. NIEHREN, J.-M. TALBOT, S. TISON. *Composing Monadic Queries in Trees*, in "International PLAN-X Workshop", Basic Research in Computer Science, 2006.

[24] E. FILIOT, J. NIEHREN, J.-M. TALBOT, S. TISON. *Polynomial Time Fragments of XPath with Variables*, Submitted, 2007.

[25] R. GILLERON, P. MARTY, M. TOMMASI, F. TORRE. *Extraction de relations dans les documents Web*, in "Revue RNTI - Actes de EGC'06", 2006, p. 415–420.

[26] 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", 2006.

[27] B. HABEGGER, D. DEBARBIEUX. *Learning n-ary tree-pattern queries for web data integration*, in "5th International Conference on Ontologies, Databases, and Applications of Semantics", 2006.

[28] F. JOUSSE, R. GILLERON, I. TELLIER, M. TOMMASI. *Champs conditionnels alÃ©atoires pour l'annotation d'arbres*, in "8eme Conference francophone sur l'Apprentissage automatique (CAp'2006)", 2006, p. 171–186.

[29] F. JOUSSE, R. GILLERON, I. TELLIER, M. TOMMASI. *Conditional Random Fields for XML Trees*, in "ECML Workshop on Mining and Learning in Graphs", 2006.

[30] C. KUTTLER, C. LHOUSSAINE, J. NIEHREN. *A Stochastic Pi Calculus for Concurrent Objects*, in "1st International Workshop on Probabilistic Automata and Logics", 2006.

[31] M. LATTEUX, A. LEMAY, Y. ROOS, A. TERLUTTE. *Identification des langages biAFER*, in "8eme Conference francophone sur l'Apprentissage automatique (CAp'2006)", 2006, p. 33–48.

[32] A. LEMAY, J. NIEHREN, R. GILLERON. *Learning n-ary Node Selecting Tree Transducers from Completely Annotated Examples*, in "International Colloquium on Grammatical Inference", Lecture Notes in Artificial Intelligence, vol. 4201, Springer Verlag, 2006, p. 253-267.

References in notes

[33] 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.

[34] 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.

[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. *Supervised learning for the legacy document conversion*, in "Proceedings of the 2004 ACM Symposium on Document Engineering", 2004, p. 220-228.

[38] 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.

[39] A. CULOTTA, D. KULP, A. MCCALLUM. *Gene prediction with conditional random fields*, Technical report, n^{o} UM-CS-2005-028, University of Massachusetts, Amherst, April 2005, http://www.cs.umass.edu/~culotta/pubs/culotta05gene.pdf.

[40] 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.

[41] 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.

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

[43] 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.

[44] 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.

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

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

[47] 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.

[48] 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.

[49] J. D. LAFFERTY, A. MCCALLUM, F. C. N. PEREIRA. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.*, in "Proceedings of the Eighteenth International Conference on Machine Learning (ICML)", 2001, p. 282-289.

[50] 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.

[51] 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.

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

[53] 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.

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

[55] 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.

[56] 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.

[57] 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.

[58] 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.