INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team MOSTRARE

Modeling Tree Structures, Machine Learning, and Information Extraction

Futurs

THEME SYM

Activity Report

2005

Table of contents

1. Team

MOSTRARE *is a joint project with the* LIFL *(UMR 8022 of CNRS and University of Lille 1) and the* GRAPPA *Group (EA 3588 of the University of Lille 3).*

Team Leader
Rémi Gilleron [professor, University of Lille 3]

Administrative assistant
Karine Lewandowski [shared by 3 projects]

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

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

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

Ph. D. student
Iovka Boneva [MESR fellowship, since October 2002]
Laurent Candillier [CIFRE fellowship, since May 2003]
Julien Carme [MESR fellowship, from October 2002 to September 2005]
Denis Debarbieux [MESR fellowship, since October 2002]
Emmanuel Filiot [INRIA Région Nord-Pas-de-Calais fellowship, since October 2005]
Patrick Marty [INRIA and Région Nord-Pas-de-Calais fellowship, since October 2003]
Florent Jousse [INRIA and Région Nord-Pas-de-Calais fellowship, since October 2004]
Laurent Planque [MESR fellowship, since October 2004]

Junior Technical staff
Jean-Philippe Nirel [since October 2005]
Missi Tran [from Jan. to June 2005]

Post-doctoral fellow
Habegger Benjamin [from October 2004 to November 2005]

2. Overall Objectives

2.1. Overall Objectives

The objective of MOSTRARE is to develop adaptive information extraction systems for semi-structured documents that can fully exploit available tree structure. We want to develop novel models for modeling trees, novel machine learning techniques which target such models, and integrate learning algorithms into information extraction systems. The trade-offs between expressiveness, learnability, and efficiency are to be understood.

In our previous project proposal, we summarized these goals in the following two research lines:

Modeling tree structures for information extraction: define and investigate models of tree structures as needed by information extraction; develop corresponding algorithms and software components.

Machine learning for information extraction: develop learning algorithms that induce models of tree structures and apply them to information extraction. Combine learning algorithms for tree and string models so that they apply to diverse data formats, and possibly to heterogeneous data.

3. Scientific Foundations

3.1. Modeling Tree Structure for information extraction

Keywords: *queries in trees*, *semi-structured documents*, *tree automata and logic*, *tree wrapper*, *xml*.

XML or HTML documents are usually parsed into trees with leaves containing texts. The nodes of the tree correspond to XML elements and are labeled by XML tags. Attributes can be collected in extra subtrees with unordered children.

XML trees are *unranked* in that every node may have an unbounded number of children. This is in contrast to the more standard *ranked trees* where the number of children is fixed by the arity of the node's label. XML trees have nodes with children ordered from the left to the right (for elements), and other nodes with unordered children (for attributes).

In this report, we will deal with two kinds of unranked trees, ordered unranked trees where the children of all nodes are ordered (for instance by ignoring or ordering attributes) and unordered unranked trees, where the children of all nodes are unordered (for instance by ignoring the ordering everywhere). In the default case, unranked trees will be ordered.

3.1.1. *Tree automata*

frequently and naturally appear in the XML world. The well-typedness of documents, for instance, can be defined by a requiring membership to the tree language recognized by some tree automaton. DTDs or XML-Schema provide a syntax for defining tree automata for XML trees.

Many different notions of *tree automata for unranked (ordered) trees* have been proposed recently. Brüggemann-Klein, Wood, and Murata's [35] notion of unranked tree automata (UTAs) has been most widely adopted in database theory [56]. The same notion, however, was previously introduced by Thatcher in 1967 [68]. Seidl and Neumann [59] argue for forest grammars and forest automata as an alternative. Frick, Grohe, and Koch [46] opt for tree automata on standard binary encodings of unranked trees.

Tree automata for unordered (unranked) trees have found much recent interest as well. Most tree automata notions in are build in the spirit of UTA's. Lugiez and Dal Zilio [70] propose *sheaves automata* for modeling aspects of XML-Schema. A similar notion of tree automata was proposed by Seidl, Schwentick, and Muscholl [67] for expressing numerical document queries, that is equally expressive to Presburger MSO, MSO on unordered trees extended by Presburger formulas. Earlier proposals for automata for unordered trees date back until 1989 to the algebraic approach by Courcelle [43]. These capture MSO without Presburger formulas.

3.1.2. *Node selection queries in trees*

Node selection queries are particularly relevant as targets for information extraction, but equally important for XML databases and programming language. In an XML document corresponding to a phonebook, one might want to extract

1. all person names, or alternatively
2. all pairs of person names with their phone numbers.

These information extraction problems can be most easily expressed by node selection queries in `XML` trees, say in MSO. The first example requires a *monadic query*. Person names can be found for instance on all nodes that are a first child x of some NAME labeled node.

$$\exists y \quad (\mathtt{Label}_{\mathtt{NAME}}(y) \wedge \mathtt{first}-\mathtt{child}(y,x))$$

The second example requires a *binary query*. Person names with their phone numbers can be found at all pairs of nodes (x_1,x_2) that satisfy for instance the conjunctive MSO formula:

$$\begin{aligned}\exists y_1,y_2 \quad & (\mathtt{Label}_{\mathtt{NAME}}(y_1) \wedge \mathtt{next}-\mathtt{sibling}(y_1,y_2) \wedge \mathtt{Label}_{\mathtt{PHONE}}(y_2) \wedge \\ & \mathtt{first}-\mathtt{child}(y_1,x_1) \wedge \mathtt{first}-\mathtt{child}(y_2,x_2))\end{aligned}$$

Shortly before the creation of the Mostrare project, Gottlob and Koch [49] started promoting *monadic Datalog* as a suitable language for defining monadic queries in visual information extraction. His group in Vienna has supported this claim practically in their Lixto system [32], [48]. On the more theoretical side, they proved that monadic Datalog captures all monadic MSO definable queries, while allowing for efficient algorithms for query answering that require linear time combined complexity.

In recent `XML` programming languages, node selection queries appear in form of tree patterns. Pierce and Hosoya [50] use *regular tree patterns* for expressing n-`ary` queries in the `XML` programming language XDuce. These are closely related to tree automata. Castagna et al. [39], [33] apply similar patterns for the `XML` programming language CDuce.

Tree automata based approaches towards node selecting queries have been advocated for the database community. Neven and Van de Bussche [60] show how to define node selection queries by particular attribute grammars, another appearance of tree automata. Neven and Schwentick [61] propose query automata as an alternative. Seidl and Berlea [34] opt for forest grammars, yet another view on tree automata.

3.2. Machine learning for information extraction

Keywords: *grammatical inference, semi-structured documents, statistical learning, wrapper induction*.

The most relevant machine learning task from the perspective of information extraction is to induce queries for informative chunks in documents from a sample of annotated examples. For instance, consider the example of an XML phonebook. The target query is the binary query for person names and phone numbers. The query designer annotates tuples of values. The wrapper induction system should infer the target query. Most traditional approaches operate on textual documents whereas Mostrare wishes to extend these methods to tree-structured or semi-structured documents.

The scientific challenges for adaptive information extraction from tree structured documents w.r.t. machine learning are at least twofold. First, machine learning algorithms have to be extended to the case of tree structured inputs. Second, the number of annotations by the query designer should be very low. Along these perspectives, we discuss two approaches, grammatical inference and statistical machine learning.

3.2.1. Grammatical inference

Grammatical inference is the discipline of learning formal languages from annotated examples. Positive examples are members of the target language that have been annotated positively, negative examples are non-members that have been annotated negatively. Many ideas date back to Gold [47] who showed that all regular languages can be identified from positive and negative examples.

The RPNI algorithm from Oncina and Garcia [63] for regular inference from positive and negative examples infers the *minimal deterministic finite automaton* recognizing the target language. It requires a sample of positive and negative examples of polynomial size depending on the target automaton, and runs in polynomial time depending on the size of the sample. The correctness of the algorithm depends on the *Myhill-Nerode theorem* which characterizes the unique minimal deterministic finite automaton for regular languages.

The usefulness of grammatical inference for wrapper induction has been illustrated before nevertheless. Chidlovskii's [40] and Hsu [51] use it to learn wrappers modeled as string transducers. For tree wrappers, Kosala et al. base wrapper induction on the induction of subclasses of regular tree languages (local tree languages in [53], contextual tree languages in [65]).

It is known that RPNI extends to ranked trees [64], when generalizing finite words into tree automata, mostly since the *Myhill-Nerode theorem* holds for bottom-up deterministic tree automata. The following questions, however, have been opened at the starting point of the Mostrare project:

1. How can one extend RPNI to unranked trees?
2. Can RPNI be used to infer node selection queries in trees?
3. Does this lead to feasible methods for information extraction?

3.2.2. Statistical machine learning.

Most of the existing wrapper induction techniques are based on statistical machine learning algorithms [54], [58]. This is because of the presence of noise in real data sets, most typically in the textual parts of semi-structured data. Research perspectives deal with wrapper induction for relation extraction from tree structured documents.

In order to define wrapper induction systems for n-ary relations, two methods have been considered. The first one is to combine unary wrappers. This requires either to learn a model for the combination [41] or requires the user's help [52], [58] or to apply heuristics. In the case of tree structured documents, one heuristic is based on the hierarchical structure of documents [32]: first locate an ancestor for a tuple, and then apply from this ancestor a unary wrapper for each component. This is problematic when tuples are in overlapping parts of the trees. The second approach is to directly define n-ary wrappers. The system WIEN [55] deals with n-ary extraction in a double loop that extracts components of a tuple in the inner loop and tuples in the outer one, the system SOFT MEALY [51] uses finite transducers, and the system LIPX [69] uses inductive logic programming. These systems rely on strong hypotheses about the tree organization. As they consider the textual view, WIEN and SOFT MEALY are not adapted when some values are factorized among several tuples or when tuples are interlaced. Also WIEN is not robust against missing values.

The following questions, however, have been open at the starting point of the Mostrare project:

1. How can one infer n-ary queries from tree structured documents with statistical learning algorithms ?
2. How can these wrapper induction systems deal with the different tree representations of n-ary relations ?

4. Application Domains

4.1. Application Domains

Keywords: *Web intelligence*, *business intelligence*, *data integration*, *information retrieval*, *knowledge management*, *multimedia*.

The *information overload* that users of the Web often experience calls for *intelligent Web services* that provide high-quality collected and value-added information. The information from a significant number of sources relevant to the domain of the service has to be extracted, integrated, and made accessible in a uniform manner. Today's information extraction programs are used as basic components in intelligent Web services, information mediator services, in information retrieval, and text mining tools.

In the first project period (2003-2005), we have focused on general tools for *Web information extraction* from homogeneous document collections generated by some databases. We believe that similar technologies can be equally useful in more concrete applications to *business intelligence* and *knowledge management*.

5. Software

5.1. Squirrel : Tree Automata for Query Induction

Keywords: *Mozilla*, *grammatical inference*, *query*, *wrapper induction*.

Participants: Julien Carme [correspondent], Aurélien Lemay, Joachim Niehren.

SQUIRREL is a visually interactive information extraction tool based wrapper induction by grammatical inference develloped by Carme as part of his PhD project [6]. The goal of SQUIRREL is to let Web users create their own wrappers.

SQUIRREL offers a visually interactive user interface for Web information extraction, which can be downloaded and plugged into into the standard Mozilla-Firefox Web Browser. Users have to annote some few elements of the Web side that they want to be extracted. The system then proposes a wrapper, and presents its output on the current Web page to the user. The user may correct the proposed wrapper by adding further positive annotations or by contradicting some of the proposed elements that should not have been selected. The process continues until the user is satisfied by the current wrapper. Though, it might be necessary to annotate examples on several pages, the overall amout of required element annotations remains very small in practice.

Experiments on realistic Web documents confirm excellent quality with very few user interactions – annotations and corrections – during wrapper induction.

With our temporary engineer Han Missi Tran, we have started modularizing and extending the visually interactive user interface of the SQUIRREL system, so that it becomes generic for arbitrary wrapper induction algorithms.

5.2. PAF : Statistical Classification for Information Extraction

Keywords: *HTML documents*, *supervised classification*, *texts*, *wrapper induction*.

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

Marty [57][29] is developing the PAF platform for wrapper induction based on statistical classification (previously called CafeIn). PAF is parameterized by a document representation model and a supervised classification algorithm that can operate on that model. Currently PAF includes a number of feature-sets for textual and structured documents, that can be easily customized or extended.

Recently, PAF has been integrated into the visually interactive Web user interface, abstracted from the original interface of the SQUIRREL system. Furthermore, we have developped algorithms for n-ary relation extraction from Web Document and Marty has started experimenting with n-ary queries.

5.3. OcamlQuery: queries in XML documents

Keywords: *XML documents*, *texts*, *wrapper*.

Participants: Emmanuel Filiot [correspondent], Joachim Niehren.

The *OcamlQuery* software is a prototype for querying XML documents. The system provides a syntax *à la* SQL to express n-ary queries by composition of monadic queries [28]. The system offers facilities to add new monadic extraction systems. Actually monadic queries can be expressed by monadic second order formulae, tree automata or XPath expressions. Several output formats are provided. It also includes a lot of useful built-in predicates and can be used in a toplevel or as a library. The query answering efficiency will be improved by implementing new algorithms and we would like to extend the querying kernel to a transformation kernel. It uses the *Stepwise Tree Automata Library* which provide functions for using stepwise automata [38] with unranked trees: membership, determinisation, union, intersection, complementation.

6. New Results

6.1. Modeling Tree Structures

6.1.1. Automata for unranked trees

Keywords: *monadic second-order logic, monadic Datalog, n-ary queries in unranked trees, tree automata*.

Participants: Julien Carme, Joachim Niehren [correspondent], Laurent Planque, Jean-Marc Talbot, Sophie Tison.

Automata for unranked ordered trees form a foundation for XML schemas, querying and pattern languages. A number of different automata formalism has been proposed for this purpose. Automata induction usually targets for *bottom-up deterministic* tree automata of *minimal size* measured in the number of states.

Martens and Niehren [19] show that the existing automata models for unranked trees fail to yield good notions of bottom-up determinism, which renders them unsuitable as targets for automata induction. They prove that minimal deterministic *unranked tree automata* (UTAs) are not unique and that their minimization problem is NP-hard. They also prove that stepwise tree automata (proposed by Carme, Niehren, and Tommasi [38]) lead to a natural notion of bottom-up determinism that solves the minimization problem for the more standard UTAs, since the Myhill-Nerode theorem becomes valid. Stepwise tree automata thereby lay a solid foundation for grammatical inference on XML-trees (see contributions to machine learning below).

Independently of the Mostrare approach, two alternative solutions have been proposed recently[1]. Among those, Martens and Niehren show that bottom-up deterministic stepwise tree automata always yield the most succinct representations. This underlines their relevance for automata induction: smaller automata are easier to learn.

6.1.2. Queries in Unranked Trees

Keywords: *modal logic, queries, semi-structured data, unranked trees*.

Participants: Emmanuel Filliot, Joachim Niehren, Laurent Planque, Sophie Tison.

Node selection in unranked ordered trees is a fundamental querying problem for XML databases, programming languages, and information extraction. Node selection queries can be understood as languages of annotated trees, whose boolean annotations specify whether or not a node is selected.

Niehren, Planque, Talbot, and Tison [4] investigate representations of *n-ary node selection queries in trees* by *successful runs of tree automata*. They show that run-based n-ary queries capture MSO, contribute algorithms for enumerating answers of n-ary queries, and study the complexity of the problem. They investigate the subclass of run-based n-ary queries by unambiguous tree automata.

Filiot [28] proposes a new class of querying languages for defining n-ary node selection queries as compositions of monadic queries. The choice of the underlying monadic querying language is parametric. He shows that compositions of monadic MSO-definable queries capture n-ary MSO-definable queries, and distinguishes an MSO-complete n-ary query language that enjoys efficient query answering algorithms.

6.1.3. Queries in Unordered Trees

Keywords: *modal logic, queries, semi-structured data, unordered unranked trees*.

Participants: Iovka Boneva, Yves Roos, Jean-Marc Talbot [correspondent], Sophie Tison.

It is often useful to ignore the ordering of elements in semi-structured documents, so that these become unordered trees. The canonical logical language for querying unordered trees are MSO and Presburger MSO. These logics can express numerical queries in semi-structured documents, and model aspects of SCHEMAS.

Boneva, Talbot, and Tison investigate the *spatial logic TQL*[2] withing Boneva's PhD project [5], an alternative querying formalism for unordered trees, invented in the context of the ambient calculus. In their LICS paper

[1] Presented by Raeymaekers and Bruynooghe [66] and Cristau, Loeding, and Thomas [44].

[2] The spatial logic TQL has been presented by Cardelli and Ghelli in [36].

[2], they prove the quantifier-free fragment of TQL to be strictly more expressive than Presburger MSO, and distinguish fragments of TQL whose querying power precisely captures the logics MSO and Presburger MSO respectively. All results are based on their tree automata framework from [14].

Boneva, Talbot and Tison [2] exhaustively classify the querying power of TQL based querying languages in unordered trees. In previous work [10], they studied the expressiveness of the ambient calculus, a formalism for describing distributed and mobile computations, which motivated the invention of TQL.

Talbot, Tison, and Roos [22] cooperate with Hitoshi Ohsaki from Tokyo on equational tree automata modulo associativity and commutativity (AC), following the axiomatic approach towards unorderedness in trees. They prove that the membership problem for monotone AC-tree automata is PSPACE-complete, while solving a hard question that was left open by previous work. The technique used in obtaining the above result yields the answers to two different questions, specifically that the family of monotone AC-tree languages is not closed under complementation, and that the inclusion problem for monotone AC-tree automata is undecidable.

6.1.4. *Queries in Digraphs*

Keywords: *path constraints*, *rewriting*, *semi-structured documents*.

Participants: Anne-Cécile Caron, Denis Debarbieux, Yves Roos, Sophie Tison [correspondent], Yves André [collaborator].

Semi-structured documents with hyper-links or other references are usually modeled as rooted edge-labeled directed graphs. Monadic queries for nodes in digraphs can be expressed by path expressions. Abiteboul and Vianu introduced *path inclusion constraints* for digraphs 1997 in the context of query optimization[3].

Andre, Caron, Debarbieux, Roos, and Tison study *path inclusion constraints* and *path queries* within Debarbieux's PhD project [7]. In [13], they show how to extract inclusion constraints of existing indexes.

6.1.5. *Tree Constraints*

Keywords: *dominance constraints*, *subtype constraints of programming languages*, *underspecified semantics*.

Participants: Joachim Niehren [correspondent], Denis Debarbieux, Yves Roos, Sophie Tison, Yves André [collaborator].

Positive conjunctive logic formulas can express conjunctive queries in trees. In computational logics, such formulas are usually called *tree constraints*. Tree constraints enjoy multiple modeling applications in divers areas.

Niehren has investigated satisfaction problems for *dominance constraints* with colleagues from Saarbrücken and Berlin (Koller, Thater, Mehlhorn, Bodirsky, etc). These are tree constraints that are popular for modeling underspecified semantics of natural language. Motivated by the same application, Niehren has studied the more expressive class of *parallelism constraints* [18], [21] [45] with colleagues from Barcelona and Saarbrücken (Villaret, Levy, Erk).

With his PhD student Priesnitz from Saarbrücken and colleagues from Berkeley and Paris (Aiken, Su, Treinen), Niehren has studied *subtype constraints* that arise during type inference of programming languages [20], [12] [62]. Types are modeled as trees, so that subtype constraints become tree constraints.

6.2. Machine Learning for Information Extraction

6.2.1. *Wrapper induction by grammatical inference*

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

Participants: Aurélien Lemay [correspondent], Julien Carme, Rémi Gilleron, Joachim Niehren, Marc Tommasi, Alain Terlutte [collaborator].

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

[3] Path inclusion constraints for digraphs were introduced by Abiteboul and Vianu [31].

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 their Machine Learning submission [3], [17], Carme, Gilleron, Lemay, and Niehren turn their induction algorithm for monadic queries presented in [37] 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.

This interactive information extraction process on the above induction algorithm for monadic queries has been implemented in the SQUIRREL system, as part of the PhD project of Carme [6] (see below).

6.2.2. Statistical Wrapper Induction

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

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

Gilleron, Marty, Tommasi, and Torre approach wrapper induction by statistical machine learning techniques within Marty's PhD project. In [29], they have extended PAF to extracting n-ary queries in tree structured documents. The system is based on combination techniques. Classifiers are combined in an iterative process where the key point is *data enrichment*: at a given step in the iteration, the document representation is updated with outcomes of classifiers used in previous steps. Experimental results show that the PAF system outperforms existing ones for many tree representations schemes in the case of factorized values and in the case of missing values. They also show that the system can handle the different tree representations of n-ary relations.

Habegger [30] is recently studying the problem of learning n-ary extraction patterns in trees by methods of inductive logic programming.

6.2.3. Statistical classification

Keywords: *semi-supervised classification*, *wrapper induction*.

Participants: Francesco De Comité [collaborator], Rémi Gilleron, Marc Tommasi [correspondent].

Annotating examples for machine learning is always an expensive time consuming process, not only in wrapper induction. Denis et. al. [11] propose a learning model from positively annotated examples that is enhanced by unlabeled examples.

6.2.4. Statistical clustering

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

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

Statistical clustering is of interest for classifying semi-structured documents without disposing annotated examples. Candillier, Tellier, Torre and Bousquet [15] propose a new statistical method and system for *subspace clustering*. Subspace clustering refines more traditional clustering tasks in that every cluster should be definable on a small subset of attributes. They are currently studying the adaptation of subspace clustering methods to semi-structured datasets, by encoding trees though set of attributes-value pairs. They participate in the INEX/PASCAL challenge on document mining co-proposed by Mostrare : the proposed learning algorithm has been classed second out of six in clustering, and first in classification [27]. They also propose a new method for evaluating clustering algorithms on non-subjective bases [26].

7. Contracts and Grants with Industry

7.1. Contracts and Grants with Industry

In 2005, we have intensified cooperations with the Lixto information extraction company in Vienna, a spin-off of G. Gottlob's database and artificial intelligence group at the technical university of Vienna. We have proposed to pursue this cooperation in form of an associated research team. Carme is actually working as a post-doctoral student with this research group.

We have continued our regular exchanges with B. Chidlovskii form the Xerox Research Center Europe XRCE in Grenoble. We will propose two master project – on PDF to XML conversion and on statistical tree automata – with the goal to intensify the cooperation. An RNTL project (ATASH) has been proposed conjointly on automatic document conversion, and has been recently accepted. An RNTLK project WebContent with a consortium of french companies has been recently accepted. Candillier is supported by the company PERTINENCE.

8. Other Grants and Activities

8.1. French Actions

8.1.1. ACI Masse de Données ACIMDD

Participants: Julien Carme, Rémi Gilleron, Aurélien Lemay, Patrick Marty, Joachim Niehren, Alain Terlutte, Isabelle Tellier, Marc Tommasi [correspondent].

We are involved in the French cooperation project "ACI masse de données – ACI-MDD – Accès au Contenu Informationnel pour les Masses de Données et Documents" (2003–2006). The aim of the project is the design of algorithmic tools for Information Retrieval, Information Extraction and Text Classification for semi-structured documents. Our partners are: Patrick GALLINARI (Coordinator - LIP6) and Marie-Christine ROUSSET (LRI and GEMO INRIA project). More information about the project can be found on http://www.grappa.univ-lille3.fr/twiki/bin/view/Acimdd. The marginal budget allocated to the Mostrare project is 70 Keuros over the period 2003-2006.

8.1.2. ACI TraLaLA: Transformation Languages, Logic and Application

Participants: Iovka Boneva, Anne-Cécile Caron [correspondent], Denis Debarbieux, Joachim Niehren, Yves Roos, Jean-Marc Talbot, 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.

8.1.3. ACI Marmota : Stochastic Tree Models and Stochastic Tree Transformation

Participants: Aurélien Lemay, Joachim Niehren, Marc Tommasi [correspondent], Yves Roos.

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). Allocated budget is not known. More information about the project can be found on http://www.grappa.univ-lille3.fr/twiki/bin/view/Acimdd.

8.1.4. RNTL ATASH

Participants: Aurélien Lemay, Joachim Niehren, Marc Tommasi [correspondent], Yves Roos.

We have proposed an exploratory RNTL project ATASH on adaptative tree transformations 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. This project has been recently accepted.

8.1.5. RNTL Webcontent

We are included in a consortium for the development of a platform for Web document processing and semantic Web. We should integrate and adapt our prototypes for Web information extraction. This RNTL Project has been recently accepted.

9. Dissemination

9.1. Scientific Animation

- **Program Committees:**
  S. TISON was member of the editorial board of RAIRO - Theoretical Informatics and Applications.
  R. GILLERON was PC member of CAP'2005 (French conference on machine learning) and PC member of EGC'2006 (french conference on knowledge discovery)
  M. TOMMASI was PC member of CAP'2005
  J. NIEHREN was PC member of FROCOS'2005.
- **Workshop Organization**
  Mostrare co-organizes EGC'2006 (French Conference on knowledge discovery) in Lille.
- **Invited talks**
  J. NIEHREN at the international workshop on unification 2005,
  J. NIEHREN at the meeting of the French database working group (GDR ALP) 2005
  J. NIEHREN at the German meeting of theoretical computer science (Theorietag der Gesellschaft für Informatik) 2005
- **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, and has been director of the doctoral school SPI of the university Lille 1 (2002 - 2005).
  R. GILLERON is head of the research group GRAPPA of the university of Lille 3, member of the scientific committee of Lille 3.

9.2. Teaching and Scientific Diffusion

- REPSONSIBILITIES

Rémi GILLERON	96 hours	masters
	96 hours	administrative responsabilities: member of the direction of Lille 3 director of Lille 3 computer center
Joachim NIEHREN	10 hours	masters
Aurélien LEMAY	192 hours	bacchalor and masters
Isabelle TELLIER	192 hours	bacchalor and masters
Marc TOMMASI	192 hours	bacchalor and masters
Fabien TORRE	192 hours	bacchalor and masters
Anne-Cécile CARON	192 hours	bacchalor and masters
Yves ROOS	192 hours	bacchalor and masters
Jean-Marc TALBOT	192 hours	bacchalor and masters
Sophie TISON	192 hours	bacchalor and masters head of scientific doctoral school in Lille 1 (2002-2005)

- MASTER LECTURES PRESENTED AT THE UNIVERSITY OF LILLE 1
  - Logic et Modelisation:
    - A.-C. CARON, J. NIEHREN, J. M. TALBOT, and P. DEVIENNE (2005-06)
  - Machine Learning for Information Extraction :
    - R. GILLERON (2005-06)
- MASTER PROJECT:
  - V. CORNET on probabilistic document transformation.
  - E. FILIOT on composition of monadic queries.
- DIRECTION OF PHD THESIS SUBMITTED IN 2005:
  - J. Carme on infering regular queries in trees and applications to web information extraction (University of Lille 3). Directed by R. Gilleron.
  - D. Debarbieux on Semistructured data models as a Rooted direct labeled graph (University of Lille 1). Directed by S. Tison and A.C. Caron.
- PHD COMMITTEES:
  R. GILLERON belongs to the committee of D. DEBARBIEUX (Lille), J. NIEHREN belonged to the committee of J. CARME (Lille). S. TISON was member of the committee of J. CARME (Lille), D. DEBARBIEUX (Lille), O. BARAIS (Lille). M. TOMMASI was member of the committee of J. CARME (Lille)
- Habilitation committees: R. GILLERON belonged to the committee of I. TELLIER (Lille), A. RAKOTOMAMONJY (Rouen) and E. GAUSSIER (Xerox - Grenoble). S. TISON belonged to the committee of I. DURAND (Bordeaux), C. DHAENENS (Lille), S. LIMET (Orleans) and J.M. TALBOT (Lille).
- Evaluation committees : R. GILLERON was member of the scientific committee for the evaluation of LRI (computer science department of Orsay, Paris 11).

10. 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, Prague (Czech Republic)", LNCS, vol. 3153, SV, august 2004, p. 863-875, http://www.lifl.fr/~debarbie/DOC/ExtractionAndImplicationOfPathConstraints.pdf.

[2] I. Boneva, J.-M. Talbot, S. Tison. *Expressiveness of a spatial logic for trees*, in "20th Annual IEEE Symposium on Logic in Computer Science", IEEE Comp. Soc. Press, 2005, http://www.lifl.fr/~boneva/papers/lics2005.pdf.

[3] J. Carme, R. Gilleron, A. Lemay, J. Niehren. *Interactive Learning of Node Selecting Tree Transducer*, in "Machine Learning", to appear, 2006, http://www.ps.uni-sb.de/Papers/abstracts/cut.html.

[4] J. Niehren, L. Planque, J.-M. Talbot, S. Tison. *N-ary Queries by Tree Automata*, in "10th International Symposium on Database Programming Languages", LNCS, vol. 3774, SV, September 2005, p. 217–231, http://www.ps.uni-sb.de/Papers/paper_info.php?label=n-ary-query.

Doctoral dissertations and Habilitation theses

[5] I. Boneva. *Logics for unranked and unordered trees and their use for querying semistructured data*, PhD thesis. Université des Sciences et Technologies de Lille - Lille 1. To appear, Ph. D. Thesis, January 2006.

[6] J. Carme. *Inférence de requêtes dans les arbres et applications à l'extraction d'informations sur le Web*, Ph. D. Thesis, Université Charles-de-Gaulle - Lille 3, 2005.

[7] D. Debarbieux. *Modélisation et requêtes des documents semi-structurés: exploitation de la structure de graphe*, Ph. D. Thesis, Université des Sciences et Technologies de Lille, Lille I, 2005, http://www.lifl.fr/~debarbie/these/.

[8] J.-M. Talbot. *Model-checking pour les ambients : des algèbres de processus aux données semi-structurées*, Habilitation à diriger des recherches. Université des sciences et technologies de Lille - Lille I, Ph. D. Thesis, 2005.

[9] I. Tellier. *Modéliser l'acquisition de la syntaxe via l'hypothèse de la primauté du sens*, Habilitation à diriger des recherches. Université Charles-de-Gaulle - Lille 3, Ph. D. Thesis, 2005.

Articles in refereed journals and book chapters

[10] I. Boneva, J.-M. Talbot. *When Ambients Cannot be Opened*, in "TCS", vol. 2, n^o 333, 2005, p. 127-169, http://www.lifl.fr/~boneva/papers/BonevaTalbot-WhenAmbientsCannotBeOpened.pdf.

[11] F. Denis, R. Gilleron, F. Letouzey. *Learning from Positive and Unlabeled Examples*, in "TCS", vol. 348, n^o 1, 2005, p. 70-83.

[12] Z. SU, A. AIKEN, J. NIEHREN, T. PRIESNITZ, R. TREINEN. *First-Order Theory of Subtyping Constraints*, in "ACM Transactions on Programming Languages and Systems", 2005, http://www.ps.uni-sb.de/Papers/abstracts/sub-journal.html.

Publications in Conferences and Workshops

[13] Y. ANDRE, A.-C. CARON, D. DEBARBIEUX, Y. ROOS. *Indexes and path constraints in semistructured data*, in "DEXA Workshop on Logical Aspects and Applications of Integrity Constraints", IEEE Comp. Soc. Press, August 2005, p. 837 - 841, http://www.lifl.fr/~debarbie/DOC/IndexesAndPathConstraints.pdf.

[14] I. BONEVA, J.-M. TALBOT. *Automata and Logics for Unranked and Unordered Trees*, in "20th International Conference on Rewriting Techniques and Applications", LNCS, SV, 2005.

[15] L. CANDILLIER, I. TELLIER, F. TORRE, O. BOUSQUET. *SSC : Statistical Subspace Clustering*, in "4th International Conference on Machine Learning and Data Mining in Pattern Recognition, Leipzig, Germany", P. PERNER, A. IMIYA (editors). , vol. LNAI 3587, SV, july 2005, p. 100–109, http://www.grappa.univ-lille3.fr/~candillier/publis/MLDM05.pdf.

[16] L. CANDILLIER, I. TELLIER, F. TORRE, O. BOUSQUET. *SSC : Statistical Subspace Clustering*, in "5ièmes Journées d'Extraction et Gestion des Connaissances", 2005, p. 177–182, http://www.grappa.univ-lille3.fr/~candillier/publis/EGC05.pdf.

[17] J. CARME, R. GILLERON, A. LEMAY, J. NIEHREN. *Interactive Learning of Node Selecting Tree Transducer*, in "IJCAI Workshop on Grammatical Inference", 2005.

[18] J. LEVY, J. NIEHREN, M. VILLARET. *Well-nested Context Unification*, in "20th International Conference on Automated Deduction", LNAI, vol. 3632, SV, June 2005, p. 149-163, http://www.ps.uni-sb.de/Papers/abstracts/wellnested-cu.bib.

[19] W. MARTENS, J. NIEHREN. *Minimizing Tree Automata for Unranked Trees*, in "10th International Symposium on Database Programming Languages", LNCS, vol. 3774, SV, August 2005, http://www.ps.uni-sb.de/Papers/abstracts/mini.html.

[20] J. NIEHREN, T. PRIESNITZ, Z. SU. *Complexity of Subtype Satisfiability over Posets*, in "14th European Symposium on Programming", LNCS, vol. 3444, SV, 2005, p. 357-373, http://www.ps.uni-sb.de/Papers/abstracts/pdl05.html.

[21] J. NIEHREN, M. VILLARET. *Describing Lambda Terms in Context Unification*, in "5th International Conference on Logical Aspects in Computational Linguistics", LNAI, vol. 3492, SV, April 2005, p. 221-237, http://www.ps.uni-sb.de/Papers/abstracts/clls-cu.html.

[22] H. OHSAKI, J.-M. TALBOT, S. TISON, Y. ROOS. *Monotone AC-Tree Automata*, in "12th International Conference on Logic for Programming Artificial Intelligence and Reasoning", LNCS, SV, 2005, http://www.lifl.fr/~yroos/stc/acta.pdf.

[23] I. TELLIER. *Automata and AB-Categorial Grammars*, in "CIAA 05 (1Oth International Conference on Implementation and Application of Automata)", 2005, p. 287-288.

[24] I. TELLIER. *Inférence grammaticale et grammaires catégorielles : vers la Grande Unification !*, in "7ème Conférence francophone sur l'apprentissage automatique", PUG, 2005, p. 63–78.

[25] I. TELLIER. *When Categorial Grammars meet Regular Grammatical Inference*, in "5th International Conference on Logical Aspects of Computational Linguistics", LNAI, vol. 4492, SV, 2005, p. p.317-332.

Miscellaneous

[26] L. CANDILLIER, I. TELLIER, F. TORRE, O. BOUSQUET. *Cascade evaluation*, dec. 2005, http://www.grappa.univ-lille3.fr/~candillier/publis/NIPS05.pdf, NIPS 2005 Workshop on Theoretical Foundations of Clustering.

[27] L. CANDILLIER, I. TELLIER, F. TORRE, O. BOUSQUET. *Transforming XML trees for efficient classification and clustering*, 2005, INEX/PASCAL worshop.

[28] E. FILIOT. *Composition de requêtes monadiques*, Masters thesis, Technical report, Université des Sciences et Technologies de Lille, 2005.

[29] R. GILLERON, P. MARTY, M. TOMMASI, F. TORRE. *Statistical classification for wrapper induction*, February 2005, http://www.grappa.univ-lille3.fr/~marty/Recherche/Publications/2005/dagstuhl05.pdf, Dagstuhl Seminar: Machine Learning for the Semantic Web.

[30] B. HABEGGER. *Tree-pattern generalization for information extraction from the Web*, GRAPPA Report, 2005.

Bibliography in notes

[31] S. ABITEBOUL, V. VIANU. *Regular Path Queries with Constraints*, in "Proc. of ACM Symposium on Principles of Database Systems (PODS 97), Tucson, Arizona", ACM-Press, May 1997, p. 122 – 133.

[32] R. BAUMGARTNER, S. FLESCA, G. GOTTLOB. *Visual Web Information Extraction with Lixto*, in "28th International Conference on Very Large Data Bases", 2001, p. 119-128.

[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, http://doi.acm.org/10.1145/944746.944711.

[34] A. BERLEA, H. SEIDL. *Binary Queries for Document Trees*, in "Nordic Journal of Computing", vol. 11, n^o 1, 2004, p. 41–71.

[35] A. BRUGGEMANN-KLEIN, D. WOOD, M. MURATA. *Regular Tree and Regular Hedge Languages over Unranked Alphabets: Version 1*, 2001, http://citeseer.ist.psu.edu/451005.html.

[36] 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.

[37] J. CARME, A. LEMAY, J. NIEHREN. *Learning Node Selecting Tree Transducer from Completely Annotated Examples*, in "7th International Colloquium on Grammatical Inference", Lecture Notes in Artificial Intelligence, vol. 3264, Springer Verlag, 2004, p. 91–102, http://www.grappa.univ-lille3.fr/~carme/publi/nst.pdf.

[38] J. CARME, J. NIEHREN, M. TOMMASI. *Querying Unranked Trees with Stepwise Tree Automata*, in "International Conference on Rewriting Techniques and Applications", Lecture Notes in Computer Science, vol. 3091, Springer Verlag, 2004, p. 105 – 118, http://www.ps.uni-sb.de/Papers/abstracts/stepwise.html.

[39] G. CASTAGNA. *Patterns and Types for Querying XML*, in "10th International Symposium on Database Programming Languages", LNCS, SV, August 2005.

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

[41] H. L. CHIEU, H. T. NG. *A maximum entropy approach to information extraction from semi-structured and free text*, in "Proceedings of Eighteenth national conference on Artificial intelligence", 2002, p. 786–791.

[42] H. COMON, M. DAUCHET, R. GILLERON, F. JACQUEMARD, D. LUGIEZ, S. TISON, M. TOMMASI. *Tree Automata Techniques and Applications*, 1997, http://www.grappa.univ-lille3.fr/tata.

[43] B. COUCELLE. *Resolution of Equations in Algebraic Structures*, chap. On Recognizable Sets and Tree Automata, ACADEMIC Press, 1989.

[44] J. CRISTAU, C. LÖDING, W. THOMAS. *Deterministic Automata on Unranked Trees*, in "15th International Symposium on Fundamentals of Computation Theory", 2005.

[45] K. ERK, J. NIEHREN. *Well-Nested Parallelism Constraints for Ellipsis Resolution*, in "11th Conference of the European Chapter of the Association of Computational Linguistics", Association for Compuational Linguistics, 2003, p. 115–122, http://www.ps.uni-sb.de/Papers/abstracts/wellnested.html.

[46] M. FRICK, M. GROHE, C. KOCH. *Query Evaluation on Compressed Trees*, in "18th IEEE Symposium on Logic in Computer Science", 2003, p. 188–197.

[47] E. GOLD. *Complexity of Automaton Identification from Given Data*, in "Inform. Control", vol. 37, 1978, p. 302–320.

[48] G. GOTTLOB, C. KOCH, R. BAUMGARTNER, M. HERZOG, S. FLESCA. *The Lixto data extraction project - back and forth between theory and practice*, in "23rd ACM SIGPLAN-SIGACT Symposium on Principles of Database Systems", ACM Press, 2004, p. 1-12.

[49] G. GOTTLOB, C. KOCH. *Monadic Queries over Tree-Structured Data*, in "17th Annual IEEE Symposium on Logic in Computer Science, Copenhagen", 2002, p. 189–202.

[50] H. HOSOYA, B. PIERCE. *Regular expression pattern matching for XML*, in "Journal of Functional Programming", vol. 6, n^{o} 13, 2003, p. 961-1004, http://doi.acm.org/10.1145/360204.360209.

[51] C.-N. HSU, M.-T. DUNG. *Generating Finite-state Transducers for Semi-structured Data Extraction from the Web*, in "Information Systems", vol. 23, n^{o} 8, 1998, p. 521 – 538.

[52] L. S. JENSEN, W. COHEN. *Grouping Extracted Fields*, in "Proceedings of IJCAI-2001 Workshop on Adap-

tive Text Extraction and Mining", 2001, http://www-2.cs.cmu.edu/~wcohen/postscript/IJCAI_Workshop_-_Wrapster_Associations.htm.

[53] R. Kosala, J. V. D. Bussche, M. Bruynooghe, H. Blockeel. *Information Extraction in Structured Documents using Tree Automata Induction*, in "6th International Conference Principles of Data Mining and Knowledge Discovery", 2002, p. 299 – 310.

[54] N. Kushmerick. *Wrapper induction: Efficiency and expressiveness*, in "Artificial Intelligence", vol. 118, n° 1-2, 2000, p. 15-68.

[55] N. Kushmerick. *Wrapper Induction for Information Extraction*, Ph. D. Thesis, University of Washington, 1997.

[56] L. Libkin. *Logics over unranked trees: an overview*, in "Automata, Languages and Programming: 32nd International Colloquium", LNCS, n° 3580, SV, 2005, p. 35-50, http://www.cs.toronto.edu/~libkin/publ.html.

[57] P. Marty, F. Torre. *Codages et connaissances en extraction d'information*, in "Actes de la Sixième Conférence Apprentissage CAp'2004", M. Liquière, M. Sebban (editors). , Presses Universitaires de Grenoble, 2004, p. 207–222, http://www.grappa.univ-lille3.fr/~marty/Recherche/Publications/2004/EI-CAp2004.pdf.

[58] I. Muslea, S. Minton, C. Knoblock. *Active learning with strong and weak views: a case study on wrapper induction*, in "IJCAI 2003", 2003, p. 415–420.

[59] A. Neumann, H. Seidl. *Locating Matches of Tree Patterns in Forests*, in "Foundations of Software Technology and Theoretical Computer Science", 1998, p. 134-145.

[60] F. Neven, J. V. D. Bussche. *Expressiveness of structured document query languages based on attribute grammars*, in "Journal of the ACM", vol. 49, n° 1, 2002, p. 56–100, http://doi.acm.org/10.1145/505241.505245.

[61] F. Neven, T. Schwentick. *Query automata over finite trees*, in "Theoretical Computer Science", vol. 275, n° 1-2, 2002, p. 633–674.

[62] J. Niehren, T. Priesnitz. *Non-Structural Subtype Entailment in Automata Theory*, in "Information and Computation".

[63] J. Oncina, P. García. *Inferring regular languages in polynomial update time*, in "Pattern Recognition and Image Analysis", 1992, p. 49–61.

[64] J. Oncina, P. García. *Inference of recognizable tree sets*, DSIC-II/47/93, Technical report, Departamento de Sistemas Informáticos y Computación, Universidad de Alicante, 1993.

[65] S. Raeymaekers, M. Bruynooghe, J. V. den Bussche. *Learning (k,l)-Contextual Tree Languages for Information Extraction*, in "Proceedings of ECML'2005", LNAI, 2005.

[66] S. Raeymaekers, M. Bruynooghe. *Minimization of finite unranked tree automata*, Manuscript, 2004.

[67] H. SEIDL, T. SCHWENTICK, A. MUSCHOLL. *Numerical document queries*, in "Proceedings of the Symposium on Principles Of Database Systems", 2003, p. 155-166.

[68] J. W. THATCHER. *Characterizing derivation trees of context-free grammars through a generalization of automata theory*, in "Journal of Comput. and Syst. Sci.", vol. 1, 1967, p. 317–322.

[69] B. THOMAS. *Bottom-Up Learning of Logic Programs for Information Extraction from Hypertext Documents*, in "In proceedings of European Conference on Machine Learning / Principles and Practice of Knowledge Discovery in Databases ECML/PKDD 2003", SPRINGER-VERLAG (editor). , September 2003.

[70] S. D. ZILIO, D. LUGIEZ. *XML Schema, Tree Logic and Sheaves Automata*, in "Proc. of RTA - Rewriting Techniques and Applications", R. NIEUWENHUIS (editor). , LNCS, vol. 2706, SV, 2003, p. 246–263.