INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team MOSTRARE

Modeling Tree Structures, Machine Learning, and Information Extraction

Futurs

THEME 3A

Table of contents

1. Team

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

Head of project-team
Rémi Gilleron [full professor, University of Lille 3]

Administrative assistant
Marie-Agnes Enard [shared by all UR INRIA FUTURS projects in Lille]

Invited researcher
Joachim Niehren [invited by the University of Lille 1 with the support of Région Nord-Pas-de-Calais from April to September 2003, and invited by the University of Lille 3 from October to December 2003]

Staff member Lille 3 University
Aurélien Lemay [assistant professor]
Isabelle Tellier [assistant professor, delegated by INRIA]
Marc Tommasi [assistant professor, delegated by CNRS]

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

Ph. D. student
Iovka Boneva [MESR fellowship, since October 2002]
Julien Carme [MESR fellowship, since October 2002]
Denis Debarbieux [MESR fellowship, since October 2002]
Patrick Marty [INRIA and Région Nord-Pas-de-Calais fellowship, since October 2003]

Student intern
Emmanuel Filiot [master program of ENS Lyon, June and July 2002]

2. Overall Objectives

The MOSTRARE project was successfully proposed to the project committee of UR INRIA FUTURS in February 2003. MOSTRARE is a joint team with the LIFL (UMR 8022 of CNRS and the University of Lille 1) and the GRAPPA Group (EA 3588 of the University of Lille 3). The creation of the project should be decided in the beginning of 2004.

The main objective of the MOSTRARE project is to develop adaptive information extraction systems for semistructured data that make use of the underlying tree structure. MOSTRARE goals are:

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 Structures

Key words: *semistructured documents*, *XML*, *tree automata*, *tree patterns*, *tree logics*, *queries*, *wrappers*.

The evolution of XML into a major data presentation language has reawaken a strong interest in modeling tree structures. The main objectives are to query for nodes in trees (XPath), describe sets of trees and recognize membership (DTD), and to transform trees into others (XSLT). Modeling approaches rely on finite automata, monadic second order logics, modal logics, pattern languages, attribute grammars, or tree transducers.

The main goal of the project is to design adaptive information extraction systems that fully exploit the tree structures of XML or HTML documents. In our approach we want to combine novel models of tree structures and machine learning techniques. Appropriate models of tree languages should satisfy a number of properties imposed by adaptive information extraction. Possible trade-offs between expressiveness, learnability, and efficiency are to be understood. Thus we revisit tree models under these perspectives.

We consider tree automata for unranked and/or unordered trees. We design efficient algorithms for querying semistructured data with these automata. We also study logic query languages such as Datalog, monadic second order logic, and modal logic.

3.2. Learning Models of Tree Languages and Wrapper Induction

Key words: *grammatical inference*, *statistical learning*, *wrapper induction*, *semistructured documents*.

We suppose that a collection of tree-structured documents is given as input together with annotated positions of interest. The task is to learn a tree wrapper from this collection that can identify relevant positions in unseen documents. We search for new algorithms that learn models of languages of semistructured data. We plan to extend results on grammatical inference for regular (tree) languages to the case of unranked and/or unordered trees.

Because of the presence of noise in real datasets, we search for extensions from the string case to the tree case of statistical wrapper induction algorithms. Also, we study combination of wrapper induction algorithms because data often stem from heterogeneous sources.

4. Application Domains

Key words: *Web Intelligence*, *Multimedia*, *Knowledge Management*, *Business Intelligence*, *Information Retrieval*.

The main objective is to develop wrappers for data intensive Web servers, cgi-based Web servers and Web services, that is sets of semistructured documents whose structure is quite uniform. Wrappers are used in information mediator services, in information retrieval tools and in text mining tools. No specific application domain is targeted so far but wrappers are useful in Business Intelligence and Knowledge Management.

5. Software

5.1. Stepwise Tree Automata

Participants: Emmanuel Filiot [correspondant], Julien Carme, Joachim Niehren, Sylvain Tenier.

Key words: *tree automata*, *unranked trees*, *basic operations*, *unary queries*, *semistructured data*.

The Stepwise Tree Automata Library provides functions for using stepwise automata [24] with unranked trees: membership, determinisation, union, intersection, complementation. The libraries are written in Ocaml and are available at http://www.grappa.univ-lille3.fr/~filiot/tata/. The Stepwise Query Automata Library is under development. It is a prototype, written in Ocaml, for querying XML documents.

5.2. Boosted Text Wrapper Induction

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

Key words: *wrapper induction*, *texts*, *html documents*, *boosting*.

The Boosted Text Wrapper Induction program is under development. A prototype is scheduled for December 2003. It is written in Python. From a set of texts or html documents annotated with the information to be extracted, the program outputs a rule based wrapper.

6. New Results

6.1. Modeling Tree Structures

6.1.1. Querying Semistructured Documents

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

Key words: *tree automata*, *unranked trees*, *unary queries*, *monadic second order logic*, *Datalog*.

The problem of selecting nodes from unranked trees is the most basic database query problem in the context of XML. Querying is also an important operation for information extraction; in this context it is often called wrapping.

Querying language for web information extraction should have solid and well understood theoretical foundations, a good trade-off between complexity and expressivity, easy usage in extraction tools, and support machine learning from annotated examples. Gottlob and Koch propose monadic Datalog over unranked trees as a practical language, and monadic second-order logic over unranked trees for theoretical foundations. These languages satisfy all but the final condition on learnability.

We propose *stepwise tree automata* [24], the notion of automata for the algebra of unranked trees. Stepwise tree automata induce a query notion for unranked trees which simplifies over previous automata base query notions of Neven and Schwentick [29][28]. Queries with stepwise tree automata on unranked trees correspond precisely to queries with traditional tree automata on ranked trees. They have the same expressiveness as monadic Datalog and monadic second-order logic over unranked trees.

We are currently studying induction algorithms for stepwise tree automata. First, we want to generalize learning algorithms from the ranked case to the unranked case. Second, we study learning algorithms for unary queries over semistructured documents. The idea is to consider automata queries as tree transducers, so that these can be learned by grammatical inference. This is work in progress and it is a first step toward tree wrapper induction.

6.1.2. Optimizing Queries

Participants: Anne-Cécile Caron, Denis Debarbieux, Yves Roos, Sophie Tison.

Key words: *semistructured data*, *path constraints*, *path rewriting*.

We consider semistructured data as graphs. We are particularly interested in path constraints, trying to use implicit information expressed by these constraints. A first part of our work deals with infering models from data while preserving the set of satisfied constraints. We have first proposed such a model for one datum [18] and we have defined a common model to several data. Secondly, given a set of constraints, we decide properties like boundedness of a query, implication of constraints, etc [20]. Our approach is based on rewriting and automata techniques; it yields us some new complexity results. One of our current objectives is to adapt these techniques for data integration using views.

6.1.3. Modal Logic

Participants: Iovka Boneva, Jean-Marc Talbot, Sophie Tison.

Key words: *semistructured data*, *queries*, *unordered trees*, *ambient calculus*.

We study TQL, a modal logic for querying semistructured data introduced by Cardelli and Ghelli. Semistructured data are embodied as unordered and unranked trees. We have established sharp complexity results for the model-checking problem (ie determine whether a tree satisfies a formula) [25]. Furthermore, we have shown

that satisfiability (ie determine whether there exists a tree satisfying a formula) for various fragments of this logic is not decidable [26], [25].

We carry on this work on one hand looking for decidable fragments of this logic for the satisfiability problem and for fragments having efficient model-checking and querying algorithms. On the other hand, we investigate precise relationships for other logics and automata that are dealing with unordered and unranked trees.

6.2. Learning Models of Tree Languages and Wrapper Induction

6.2.1. Boosted Textual Wrappers

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

Key words: *statistical learning*, *boosting*, *wrapper induction*, *textual data*.

The actual tendency is to represent data on the web using standards like HTML or XML. Data are heterogeneous and moreover, input data of information extraction, can be either scattered into many pieces, or hidden in large parts of purely textual data. Wrappers taking the structural information into account must therefore be combined together or with wrappers on textual data. For classification tasks, some machine learning algorithms realize such combination. We have started to examine how such methods can be used for information extraction. Our work is first to reformulate information extraction as an instance of classification tasks. Second we want to apply known techniques and combine them with structural wrapper induction.

Freitag and Kushmerick have proposed a system (BWI, [27]) for wrapper induction based on boosting techniques. The classification task consists in deciding whether a position in a text is the beginning (resp. the end) of data to extract. However, classifiers for beginnings and ends are learned independently. We investigate several ways to improve their approach. This includes the choice of an adequate representation of the data and the choice of a good base learner for the boosting approach. A prototype is developed and experiments are conducted to examine the performance of several base learners and the quality of text representations. This is ongoing work. It will be continued with the integration of structural informations for the combination of textual and structural wrappers.

6.2.2. Learning from Heterogeneous Data

Participants: Francesco de Comité, François Denis, Rémi Gilleron, Marc Tommasi.

Key words: *positive examples*, *one-class learning*, *textual data*.

Recently there has been significant interest in learning algorithms that combine information from labeled and unlabeled data. This research area is motivated by the fact that it is tedious and expensive to hand-label large amount of training data, specially for text learning tasks. This is the case for information extraction where only information to be extracted is annotated and only a small number of documents is annotated. Therefore we are interested in learning algorithms from positive and unlabeled data apply. We have defined a learning model from positive examples with the help of unlabeled examples. We have designed a text learning algorithm and a co-training algorithm for texts from positive and unlabeled examples [21].

We have considered heterogeneous data in the classical sense. We have designed classification algorithms from tabular data together with text data. Our algorithm **ADTBoost** [19] can handle tests on (continuous or discrete) tabular data as well as tests on text data.

6.2.3. Learning Natural Language

Participants: Daniela Dudau, Isabelle Tellier, Marc Tommasi.

Key words: *grammatical inference*, *language learning*, *linguistic*, *categorial grammars*.

Most algorithms in grammatical inference are based on representation schemes which have some kind of determinism. This limits their application field either on regular languages or on limited subclasses of more expressive language classes. Moreover it is known that the size of a deterministic automaton can be exponential in the size of an equivalent non-deterministic automaton. Our main idea is to define non-deterministic

representation schemes that allow learnability. We have defined a class of non deterministic tree automata [16] and have designed learning algorithms [17]. A work in progress is an extension to unranked trees.

We are also involved in learning formal grammars used in natural language processing. Our original contribution to this domain is to take into account lexical semantic information, supposed to be first acquired, to help the syntax learning process. This strategy takes advantage of the Principle of Compositionality to impose constraints on admissible tree structures underlying the input data. We consider types as lexical semantic information and propose a learning algorithm. A corpus has been created and our method has been empirically evaluated on it [22][23].

7. Contracts and Grants with Industry

MOSTRARE is a new project and we have no contract and no grant with industry so far. We have discussions with Archimed (a regional company which designs information systems for libraries and university web sites with XML technology), Xerox by the way of the Xerox Research Center Europe XRCE in Grenoble, and Lixto (an Austrian company which designs tree based wrappers).

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.

We are involved in a French research project "ACI masse de données – ACI-MDD – Accès au Contenu Informationnel pour les Masses de Données et Documents". This research project (2003–2006) is directly related with the MOSTRARE project. The aim of the project is the design of algorithmic tools for Information Retrieval, Information Extraction and Text Classification for semistructured documents. Our partners are: Patrick GALLINARI (LIP6) and Marie-Christine ROUSSET (LRI and GEMO INRIA project).

8.1.2. Action Spécifique DSTIC

Participants: Iovka Boneva, Anne-Cécile Caron, Denis Debarbieux, Joachim Niehren, Yves Roos, Jean-Marc Talbot, Sophie Tison.

We are member of a French research action "Action Incitative Bases de Données et d'Informations hétérogènes et distribuées". The leader of this action is Véronique Benzaken (LRI). The purpose is to study the links between typing, integrity and security in semistructured data. Our partners are the LRI (V. Benzaken, N. Bidoit), the LIENS (G. Castagna), and the LIUPPA (A. Gabillon).

8.1.3. Action RIP-WEB

Participants: Rémi Gilleron, Patrick Marty, Isabelle Tellier, Marc Tommasi.

We are member of a French research group on Question Answering "RIP-WEB: Recherche d'Information Précise sur le WEB: http://www.limsi.fr/Individu/monceaux/RIP-Web/rip-web.html" whose leader is Brigitte GRAU, in LIMSI, and whose purposes include to evaluate what machine learning techniques can bring to Question Answering systems.

9. Dissemination

9.1. Scientific Animation

- **Program Committees:**

S. TISON was PC member of RTA'2003, member of the editorial board of RAIRO - Theoretical Informatics and Applications and of the "SPECIF Best thesis Award" jury.
R. GILLERON was director of the scientific committee of CAP'2003 (French conference on machine learning).
J. NIEHREN was PC member of the workshop UNIF'2003
J. M. TALBOT was PC member of LPAR'2003 and is member of CSL'2004.

- **Workshop Organization**
  we have organized a workshop in December 2003 on Learning Tree Languages: http://www.grappa.univ-lille3.fr/twiki/bin/view/Public/WorkshopDec2003.
- **French Scientific Responsibilities**
  S. TISON is head of the STC team, vice-director of the LIFL (computer science department in Lille) and director of the doctoral school SPI of the university Lille 1.
  R. GILLERON is head of the GRAPPA team, member of the scientific committee of Lille 3 university, and member of the scientific committee of the RTP STIC CNRS "découvrir et résumer" (French national action of the CNRS on machine learning and data mining).

9.2. Teaching and Scientific Diffusion

- TATA is a numeric textbook on tree automata: http://www.grappa.univ-lille3.fr/tata/
- master thesis lectures: J. NIEHREN, J. M. TALBOT and S. TISON on Logic and Modelisation; I. TELLIER and M. TOMMASI on Machine Learning for Information Extraction.
- master projects: P. MARTY on Supervised Classification and Information Extraction; S. TENIER on Information extraction from semi-structured documents: MLN a deterministic approach; F. DUPONT on Learning Natural Language via Lambek grammars.
- development: A. HERBRETEAU, E. GOURNAY, S. TENIER on the implementation of stepwise queries.
- student internships: E. FILIOT (magister Lyon) on An implementation of stepwise tree automata in CAML
- PhD jury: R. GILLERON sit on the jury of G. SIOLAS, B. PIWOWARSKI (reviewer), K. N. VERMA, J. BESSOMBES (president), O. PERRIQUET (president), L. RALAIVOLA. I. TELLIER sit in the jury of Y. LE NIR. S. TISON sit in the jury of G. LEMAUR, S. MAHMOUDI, S. NEUT (president), T. TOUILI.

10. Bibliography

Major publications by the team in recent years

[1] S. ABITEBOUL, P. BUNEMAN, D. SUCIU. *Data on the Web.* Morgan Kaufmann Publishers, 2000.

[2] R. BAUMGARTNER, S. FLESCA, G. GOTTLOB. *Visual Web Information Extraction with Lixto.* in « The VLDB Journal », pages 119-128, 2001.

[3] L. CARDELLI, G. GHELLI. *A query language based on the ambient logic.* in « Proceedings of the 9th European Symposium on Programming ESOP'01 », series Lecture Notes in Computer Science, volume 2028, pages 1–22, 2001.

[4] 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.

[5] G. GOTTLOB, C. KOCH. *Monadic Queries over Tree-Structured Data.* in « Proceedings of the 17th IEEE Symposium on Logic in Computer Science (LICS 2002) », series Lecture Notes in Computer Science, pages 189–202, Copenhagen, 2002.

[6] R. KOSALA, M. BRUYNOOGHE, J. V. DEN BUSSCHE, H. BLOCKEEL. *Information Extraction from web documents based on local unranked tree automaton inference.* in « Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-2003) », pages 403–408, 2003.

[7] N. KUSHMERICK. *Finite-state approaches to Web information extraction.* in « Proc. 3rd Summer Convention on Information Extraction », 2002.

[8] M. MÜLLER, J. NIEHREN, R. TREINEN. *The First-Order Theory of Ordering Constraints over Feature Trees.* in « Discrete Mathematics and Theoretical Computer Science », number 2, volume 4, 2001, pages 193-234.

[9] F. NEVEN, T. SCHWENTICK. *Query automata over finite trees.* in « Theoretical Computer Science », number 1-2, volume 275, 2002, pages 633–674.

Articles in referred journals and book chapters

[10] E. ALTHAUS, D. DUCHIER, A. KOLLER, K. MEHLHORN, J. NIEHREN, S. THIEL. *An Efficient Graph Algorithm for Dominance Constraints.* in « Journal of Algorithms », number 1, volume 48, 2003, pages 194–219, Special Issue of SODA 2001.

[11] F. DENIS, R. GILLERON, F. LETOUZEY. *Learning from Positive and Unlabeled Examples.* in « Theoretical Computer Science », to appear.

[12] F. DENIS, A. LEMAY, A. TERLUTTE. *Learning regular languages using RFSAs.* in « Theoretical Computer Science », to appear.

[13] J. NIEHREN, T. PRIESNITZ. *Non-Structural Subtype Entailment in Automata Theory.* in « Information and Computation », number 2, volume 186, 2003, pages 319-354, Special Issue of TACS 2001.

Publications in Conferences and Workshops

[14] M. BODIRSKY, D. DUCHIER, S. MIELE, J. NIEHREN. *An Efficient Algorithm for Weakly Normal Dominance Constraints.* in « ACM-SIAM Symposium on Discrete Algorithms 2004 », to appear.

[15] I. BONEVA, J.-M. TALBOT. *When Ambients Cannot be Opened.* in « Proceedings of Sixth International Conference on Foundations of Software Science and Computation Structures », series Lecture Notes in Computer Science, number 2620, pages 169–184, 2003.

[16] J. CARME, R. GILLERON, A. LEMAY, A. TERLUTTE, M. TOMMASI. *Residual Finite Tree Automata.* in « Proceedings of the seventh int. conf. developments in Language Theory DLT'03 », series Lecture Notes in Computer Science, number 2710, pages 171–182, 2003.

[17] J. CARME, A. LEMAY, A. TERLUTTE. *Identification à la limite de langages réguliers d'arbres à résiduels premiers disjoints.* in « Proceedings of CAP'03 », pages 217–235, 2003.

[18] A.-C. CARON, D. DEBARBIEUX, Y. ROOS. *Modéles de données semi-structurées et contraintes d'inclusion.* in « RSTI série RIA-ECA », volume 17, Extraction et Gestion des Connaissances, Hermes, pages 461–472, 2003.

[19] F. DE COMITE, R. GILLERON, M. TOMMASI. *Learning Multi-label Alternating Decision Trees from Texts and Data.* in « Proceedings of Intern. Conference on Machine Learning and Data Mining », series Lecture Notes in Artificial Intelligence, number 2734, pages 35–49, 2003.

[20] D. DEBARBIEUX, Y. ROOS, S. TISON, Y. ANDRE, A.-C. CARON. *Path Rewriting in Semistructured Data.* in « Proceedings of Words'03: 4th International Conference on Combinatorics on Words », pages 358–369, september, 2003.

[21] F. DENIS, R. GILLERON, A. LAURENT, M. TOMMASI. *Text Classification and Co-training from Positive and Unlabeled Examples.* in « Proceedings of the ICML-2003 workshop: the Continuum from labeled data to unlabeled data in Machine Learning and Data Mining », pages 80–87, 2003.

[22] D. DUDAU, I. TELLIER, M. TOMMASI. *A learnable Class of CCG from Typed Examples.* in « Proceedings of the eigth conference on Formal Grammar », pages 77–88, 2003.

[23] D. DUDAU, I. TELLIER, M. TOMMASI. *Une classe de grammaires catégorielles apprenable à partir d'exemples typés.* in « Proceedings of CAp'03 », pages 169–184, 2003.

Internal Reports

[24] J. CARME, J. NIEHREN, M. TOMMASI. *Querying Unranked Trees with Stepwise Tree Automata.* Technical report, Grappa Group, Lille 3 University, 2003, http://www.grappa.univ-lille3.fr/ftp/reports/stepwise.pdf.

Miscellaneous

[25] I. BONEVA, J. M. TALBOT. *On Model-checking, Satisfiability and Safety Problems for the TQL Logic.* submitted.

Bibliography in notes

[26] W. CHARATONIK, J. M. TALBOT. *The Decidability of Model Checking Mobile Ambients.* in « Tenth Annual Conference of the European Association for Computer Science Logic - CSL 2001 », pages 339–354, 2001.

[27] D. FREITAG, N. KUSHMERICK. *Boosted Wrapper Induction.* in « Proceedungs of the 17th National Conference on Artificial Intelligence AAAI-2000 », pages 577-583, 2000.

[28] G. GOTTLOB, C. KOCH. *Monadic Datalog and the Expressive Power of Languages for Web Information Extraction.* in « Proceedings of the 21th Symposium on Principles of Database Systems (PODS) », pages 17-28, 2002.

[29] F. NEVEN, T. SCHWENTICK. *Query automata over finite trees.* in « Theoretical Computer Science », number 1-2, volume 275, 2002, pages 633–674.