Singletons associated with DKM

Mickaël Foursov, David Gross-Amblard, Israel César Lerman, Zoltán Miklós, Virginie Sans[1]

Rennes

Activity Report 2012

[1] Virginie Sans is on maternity leave for this report period.

Mickaël Foursov
Associate Professor, Université de Rennes 1

1 Overall Objectives

Keywords : Dynamical systems, generating series, identification of the input/output functionals, symbolic computations.

Algebraic modeling consists in constructing a local bilinear model (at time t_0), from an unknown dynamical system considered as a black box, up to an order k, in such a way that the difference between the outputs of the unknown system and of the model be of the order of $O((t-t_0)^k)$. The construction is based on generating series, a generalization of the transfer function to nonlinear systems.

The input/output behavior of nonlinear dynamical systems with $m-1$ inputs and one output can be locally written (in a neighborhood of t_0), under certain conditions, in term of a certain formal power series in m noncommutative variables called the generating series. This series is a generalization of the transfer function to nonlinear dynamical systems. It is constructed on an alphabet of m letters, each letter coding either an input of the system or the drift.

The first step of identification consists in computing the generating series up to a given order k, from input/output sets of an unknown system. This computation is effectuated by finding the multi-derivatives of the inputs as solutions of the system of linear equations obtained by successive differentiation of the output, by Gauss transformations and splittings [8, 10]. The truncated generating series is then prolonged to the order k, as a rational series of minimal rank [9].

The second step consists in constructing a realization of the rational series by a bilinear system (linear both in the state and in the inputs). It is this bilinear system that represents our model in the sense that the outputs of this system and of the unknown system have the same Taylor series expansion up to order k [9].

The advantages of this modeling are the following :

- It is generic, i.e. the identification of the coefficients of the generating series can be done using formal input/output sets, or using the input/output sets parameterized by the initial state.
- It is incremental, in the sense that when one passes from an approximation of order k to the one of order $k+1$, it suffices to compute only the coefficients of the words of length $k+1$.
- This modeling by generating series is well-adapted to cascade systems. Indeed, the generating series associated to the system formed from two systems connected in a cascade (with one input and drift) can be formally computed in terms of the series associated to each system.

It also presents some disadvantages :

- A combinatorial explosion is possible when the number of input is large.
- In the case of numerical input/output sets, the extraction of the values of the derivatives is limited.

This modeling can be applied to the insulinemia/glycemia behavior after an insulin injection or infusion. Using the appropriate insulinemia/glycemia correlated sets (for different insulin concentrations), we obtain a model allowing a prediction of the glycemia for insulin concentrations. This prediction is reliable with an error of 5% over an interval of fifteen minutes, for available samples. The main future goal is to propose a regulation method.

2 Scientific Foundations

Keywords : Formal power series in noncommutative variables, rational power series, dynamical systems, bilinear systems, Hankel matrices, symbolic computations, identification.

> Formal power series in noncommutative variables is a powerful tool for approximation and identification of dynamical systems. There exist at least two representations of formal power series : one uses its Hankel matrix and the other one a weighted automaton of its residuals. A generalization of the notion of Padé-type approximants in noncommutative variables provides two approximants associated to these representations. Applications of formal power series to the treatment of dynamical systems consists in constructing bilinear approximants of dynamical systems and in identifying the generating series of unknown systems from the input/output sets.

The notion of formal power series in noncommutative variables was introduced by M.P. Schützenberger, in relation to automata and formal languages.Another application of formal power series lies in the treatment of dynamical systems. M. Fliess [12] developed the idea that the generating series of a system can be used to code the input/output behavior of the system. This idea, together with the idea that the natural realization of a rational series is a bilinear system, led to the creation the algebraic modeling [9]. Algebraic modeling of an unknown dynamical system is based on the computation of its generating series truncated at an order k, on its approximation by a rational series of minimal rank and on its realization by a bilinear system [11].

3 Application Domains

Keywords : identification, modeling, regulation, diabetes.

The diabetes is a major disease caused by the inability of the pancreas to regulate the blood glucose concentration (glycemia). It is characterized by large variations of glucose concentrations due to insufficient production of endogenous insulin. The patients need exogenous insulin administration in order to keep up the metabolic control. The currently most widely–used therapeutic method consists in a series of 3 to 5 daily insulin injections with quantities based on 4 to 8 daily glucose measures. Up to now, this therapy could not restore the metabolism to its normal level, and large fluctuations happen to numerous patients.

The diabetes affects about 16 million people worldwide. The diabetes–related expenses can reach 23% of all health expenses of a country. As an example, France spent 4.86 billion euros in 1998 and 5.71 billion euros in 2000 for the treatment of diabetes and its complications. Approximately 60% of these costs are due to complications. Whereas it does not seem to be possible to diminish the cost of diabetes management, one can try to optimize the therapy in order to decrease the number of diabetes–related complications.

The insulin administration is classically made by a sub-cutaneous injection. More recently, continuous infusion techniques were developed. An infusion imitates better the pulsatile secretion of insulin. However, there still exist some technical difficulties to generalize their use. Quite recently, implantable insulin pumps were designed and thousands of them are already used world–wide.

However, in spite of this significant progress, none of the existing models of diabetic behavior seems to be sufficiently precise in order to be widely used in the clinical treatment of diabetes. As all the currently–used methods are linear, we are interested in developing a nonlinear regulation methods in order to see whether they can regulate the diabetics without any human intervention.

We are working on the application of algebraic modeling to the problem of treatment of diabetics. Taking the insulin infusion rate as the input and the blood glucose level as the output, we consider the patient as a black box whose model has to be obtained from the available measures. We think that the recent breakthroughs in the development of continuous insulin infusion devices will provide us with the necessary data for continuous real–time regulation of diabetics.

4 New Results

integration using fuzzy logic–based approaches

The algebraic modeling method works with continuous inputs/outputs and computes a generic bilinear system which approximates an unknown system up to order k, in a neighborhood of a point. Thus the formal identification is done at first; the application to a real system is realized later. This bilinear system is constructed from the successive derivatives of the inputs and the output, obtained from the numerical data. However, it may be technically impossible to compute the derivatives of orders greater than 3. Moreover, the identification is local, it is effectuated at several points and the approximating curves are not smoothed. Smoothing is particularly interesting for the problems where an off–line identification is well–adapted. Even though fuzzy models are operational in numerous practical situations, it is difficult to estimate the produced error. To identify an unknown system, the mixing of two methods consists in using the algebraic method around the measurement points and merging the local approximations using the fuzzy logic. The connecting points are those where the equation for error estimation is verified. Several questions are posed: number and distribution of points, refinement of the fuzzy system parameters.

5 Dissemination of results

5.1 Animation of the scientific community

- Mickaël Foursov serves as a referee for the scientific journals “Physics Letters A”, “Journal of Mathematical Physics” and “Symmetry, integrability and geometry: methods and applications”, “International Journal of Control”.

5.2 University teaching

- Mickaël Foursov is the director of studies of Master Miage (double major : Computer Science and Business Management).
- Mickaël Foursov is responsible for the 3rd year of studies for a Bachelor’s in Miage.

References

Publications in 2012

[1] F. Benmakrouha, M.V. Foursov, C. Hespel, J.-P. Hespel, *Stabilité glycémique en situation perturbée et adaptation thérapeutique*, Infusystèmes, 2012, Vol. 29, No. 2, pp 15-18.

Major publications in recent years

[2] Mikhail V. Foursov and Christiane Hespel, *Formal power series and polynomial dynamical systems*, in Proceedings of "Transgressive Computing '06", pp.257-270.

[3] Mikhail V. Foursov and Christiane Hespel, *Weighted Petri nets and polynomial dynamical systems*, in Proceedings of 17th International Symposium on Mathematical Theory of Networks and Systems (MTNS'06), pp.1539-1546.

[4] Farida Benmakrouha, Mikhail V. Foursov, Christiane Hespel and Jean-Pierre Hespel, *Glycaemic stability of the diabetic patient and therapeutic adjustment*, in Proceedings of 8th IEEE International Conference on Bioinformatics and BioEngineering (BIBE 2008).

[5] Mikhail V. Foursov and Christiane Hespel, *About the Decomposition of Rational Series in Noncommutative Variables into Simple Series*, in Proceedings of 6th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2009).

[6] Mikhail V. Foursov and Christiane Hespel, *On approximation of nonlinear generating series by rational series*, in Proceedings of 3rd International Conference on Complex Systems and Applications (ICCSA 2009).

Reference works and articles

[7] C. Hespel, *Iterated derivatives of a nonlinear dynamic system and Faà di Bruno formula*, Math. Comp. Simul., **42** (1996), pp.642–657.

[8] C. Hespel and G. Jacob, *First steps towards exact algebraic identification*, Discrete Math. **180** (1998), pp.211–219.

[9] C. Hespel, *Une étude des séries formelles non commutatives pour l'approximation et l'identification de systèmes dynamiques*, HDR thesis, Université de Lille–1, 1998.

[10] C. Hespel and G. Jacob, *On algebraic identification of causal functionals*, Discrete Math. **225** (2000), pp.173–191, 2000.

[11] F. Benmakrouha, M. Foursov, C. Hespel and E. Monnier, *La modélisation algébrique : méthode, avantages, inconvénients, applications*, technical report IRISA No 1407, 2001.

[12] M. Fliess, *On the concept of derivatives and Taylor expansions for nonlinear input/output systems*, in "IEEE Conference on Decision and Control" (San Antonio, Texas), 1983, pp.643–648.

David Gross-Amblard
Professor, Université de Rennes 1

1 Overall Objectives

My recent work is focused on database security and data management on the Web. After a year at IRISA, I successfully ended my remaining external projects (formal security proofs on specific watermarking protocols for numerical databases, published at ITP / defense of my former Ph.D student, Damien Leprovost). My main results at IRISA in 2012 are the successful obtention of a CominLabs funding (POSEIDON project, headed by Caroline Fontaine) along with a PhD student (Anis Bkakria, co-directed by Frederic Cuppens and Nora Cuppens-Boulahia), the recruitement of Zoltan Miklos as associate professor in the data management field, and the organization of a scientific event around Alan Turing's centenary.

My present research project aims at integrating open (as in OpenData), participative (as in Wikipedia), externalized (as in Cloud) and socialized aspects (as in recommandation systems) into classical data management systems. These new viewpoints lead to a deep modification of query optimization, data distribution and data security.

2 Scientific Foundations

Database Watermarking Watermarking techniques allow for invisible and robust information hiding in a digital document, for example the document owner's identity. Many watermarking methods exist for multimedia documents like images, sound files and video. Recently, database watermarking techniques have emerged [10, 3, 21].

I started a database watermarking working group at the Vertigo team, CEDRIC Lab, CNAM Paris. We have proposed a database watermarking model where data hiding must preserve the quality (the result) of a user-defined set of important queries. In this setting, two questions arise : (i) knowing the hiding *capacity* of a given database, that is the largest size of a hidden message, (ii) computing watermarked databases efficiently, that respects the intended result of queries.

From the theoretical point of view, I focused on the relationship between the syntactical form of the query to preserve, and the watermarking capacity. We have shown that, without hypothesis, this watermarking capacity can be null. On the contrary, if the data set fulfills reasonable assumptions, *the watermarking capacity is guaranteed, for any* `SQL` *(for relational databases) or* `XPath` *(for XML) queries.* Moreover, corresponding watermarks can be obtained efficiently. These results are published in ACM Principles of Database Systems (2003) [8]. A practical counterpart of this work has been proposed to obtain a full database watermarking prototype, Watermill [6, 9, 16, 7].

This activity has been followed in three directions :

- *Geographical databases watermarking.* This work has been done with GREYC, LAMSADE, and COGIT Labs (French National Cartography Institute) [17, 11, 15, 12, 13, 14, 16] ;
- *Medical images* watermarking under constraints [5] ;
- *Symbolic musical databases* watermarking [2].
- *Formal proof* (à la Coq) of a stylized database watermarking system [4].

Data Provenance and Trust : Classical Web and Web of Objects Faced to the Web, Database techniques have included semi-structured data, navigational query languages, massively distributed query evaluation strategies, to cite a few aspects. Moreover, the Web allows any user to become a data provider, using forums, blogs, tweets, social networks, OpenData architecture or collaborative platforms. Sophisticated on-line content can then be realized by combining data from various distant sources and services calls. In these scenarios, users may require protection methods for the intellectual property of their personal productions, and trust / provenance indicators for the data they query. I would like to consider the following questions :

– How to integrate tools for intellectual property protection in a flow of Web documents, naturally dedicated to exchange, transformation and combination with other documents during their lifecycle.
– How to integrate provenance and trust of data first in a controlled distributed context, then on the generic Web, social networks or sensor networks. The study of trust in a distributed context already produced a prominent literature. Nevertheless, recent works view distributed data as a problem of knowledge management on a large scale. The corresponding tools are then distributed deductive databases (Bloom, WebDamLog), using data production rules. To determine the trust of data produced by such rules, or the trust of the rules themselves, is a new issue.

This approach is now followed in a larger view for the design of security policy, as part of the CominLabs POSEIDON [1] project.

Strategic aspects in participative environment My work on database watermarking naturally leads to the question of the *value of data* : does my information has an (economical, scientific, ...) value for potential users ? What is the best way / time to publish information ? Several recent works focus on these questions, trying to model common behaviors associated with data advertisement systems like Google Smart Pricing and Yahoo Quality Based Pricing [18].

Those questions are hard to apprehend, because the value of a data is no longer a locally defined property, but a property that emerges from user interactions. These users seek to maximize the value of their data according to their own objectives and knowledge of the overall system. From a methodological point of view, these questions are well modeled by game theory. This theory, initially proposed by Von Neumann [22] and popularized by Nash's result [19], allows for the modeling of the behavior of autonomous actors. Its computational counterpart is now very popular, where actors are seen as machines with limited resources [20]. Applications range from crowdsourcing applications to open data publication.

3 New Results

Formal Proofs for Database Watermarking One of the long term goals of the watermarking community is to obtain complete security proofs of watermarking protocols, in a similar spirit as cryptographical protocol proofs. It is sometimes noted that existing proofs for watermarking are limited to specific classes of attacks and simply lead to an "arm race". A better situation is to obtain a proof with the following property : any victorious attacker must have solved an NP-complete problem efficiently, or must have violated a commonly accepted cryptographical hardness hypothesis.

We obtained with David Baelde, Pierre Coutieu, Julien Lafaye, Philippe Audebaud et Xavier Urbain a restricted proof of the Agrawal and Kiernan database watermarking protocol. The result is an ITP publication [4].

Ontology Watermarking Another result is the proposition of a new watermarking algorithm for populated ontologies, that is ontologies with instances of concepts. Those ontologies are currently very successful for the semantic Web, as shown by the huge YAGO and DbPedia ontologies. This work with Fabian Suchanek and Serge Abiteboul, obtained during my visiting period at the WebDam ERC project, is the first to use deletion as a method of watermarking for databases.

1. `http://www.cominlabs.ueb.eu/themes/project/`

4 Dissemination of Results

Students

Ph.D students

- (running) Joint direction (33%) with Frédéric Cuppens (33%) and Nora Cuppens-Boulahia (33%) of Anis Bkakria's thesis (Labex CominLabs funding, POSEIDON project), on "security politics for outsourced data", started September 2012.
- Joint direction (33%) with Lylia Abrouk (33%) and Nadine Cullot (33%) of Damien Leprovost's thesis (Bourgogne Young Entrepreneur Funding) entitled "Community discovery by semantic analysis", started September 2009, defended November 30, 2012. Now postdoc in the Axis team at Inria Rocquencourt.
- Joint direction (95%) with Michel Scholl (5%) of Julien Lafaye's thesis (Polytechnique funding), entitled "Database watermarking with constraint preservation", started September 2004, defended November 7, 2007. Now working for the IT company Scimetis.
- Joint direction (30%) with Bernd Amann (70%) of Camélia Constantin's thesis (French research ministry funding), entitled "Web services ranking by utility", started September 2004, defended November 27, 2007. Camélia is now a research assistant at the LIP6 Lab, Paris VI University.

Research Master students

- Julien Lafaye (2004)
- Camelia Constantin (2004)
- Ammar Mechouche (2005)
- Jean Béguec (2006)

Engineer students

- Camélia Constantin (2003), Meryem Guerrouani (2005), Guillaume Chalade (2006), Karine Volpi (2006), Robert Abo (2006), Mai Hoa Guennou (2007).

Funded projects

Labex CominLabs POSEIDON (member) This project, started in 2012, concerns the security of outsourced data (2 PhD thesis, 1 18-month postdoc, funded for 49 KE, non-staff costs).

PEPS CNRS STRATES (head) This 2010 project, funded for 10 KE studied keyword pricing in search engines, with two economists from École d'économie de Paris.

ANR CONTINT Neuma (member) [2] This 3-years project, started end 2008, funded for 620 kE, focuses on wide musical symbolic databases. This project gathers musicologists from CNRS (IRPFM), along with computer sciences labs (LAMSADE, LE2I) and an IT company (ARMADILLO).

ACI Sécurité Tadorne (head) [1] This 4-years project started in 2005, funded for 61 kE, concerns database watermarking under constraints. Participant labs are CEDRIC, GREYC, LAMSADE and COGIT (French National Cartography Agency) ;

National collaborations

- Visitor of the Wisdom group (`http://wisdom.lip6.fr`), a database group gathering the database groups from LIP6, LAMSADE and CEDRIC labs (PPF - plan pluri-formation) ;
- External participant of SemWeb and SCALP projects.
- Co-authors and collaborators : Serge Abiteboul, Fabian Suchanek, Cristina Bazgan, Bernd Amann, Philippe Rigaux, Richard Chbeir, Anne Ruas, Julien Lafaye, Camelia Constantin, Michel de Rougemont.

Invited talk

- PresDB 2007 (International Workshop on Databases Preservation, Edinburgh, March 23, 2007), "Database watermarking : protection by alteration".

Program committee

- Future Program chair of the national database conference BDA 2014.
- PC member of the workshop on Open Data WOD 2012 and 2013.
- PC member of international conferences CSTST 2008 and ICDIM 2008 ;
- Demo chair of the national conference Bases de données avancées (BDA) 2008 ;
- PC Chair of SWAN 2006 (1st Workshop on Security and Trust of Web-oriented Application Networks) ;
- PC member of the national conferences Bases de données avancées (BDA) 2005, 2008 and 2009 ;
- Reviewer for journals JOT (2012), JCSS (2005), TKDE (2005, 2006), Information systems (2007), TDSC (2005), TISSEC (2005), WWWJournal (2005), Acta Informatica (2005), Infosec (2004) and TODS (2003), external reviewer for conferences ACNS 2007, ASIACCS 2007, ICDE 2007, ICDIM 2006 et 2007, ASIAN 2005, PODS 2005, SOFSEM 2005, VLDB 2005, EDBT 2004, VLDB 2003.

Références

[1] Projet Tadorne (tatouage de données contraintes). `http://cedric.cnam.fr/vertigo/tadorne`.

[2] The NEUMA Project. `http://neuma.irpmf-cnrs.fr`.

[3] R. Agrawal and J. Kiernan. Watermarking Relational Databases. In *International Conference on Very Large Databases (VLDB)*, 2002.

[4] D. Baelde, P. Courtieu, D. Gross-Amblard, and C. Paulin-Mohring. Towards provably robust watermarking. In L. Beringer and A. P. Felty, editors, *ITP*, volume 7406 of *Lecture Notes in Computer Science*, pages 201–216. Springer, 2012.

[5] R. Chbeir and D. Gross-Amblard. Multimedia and Metadata Watermarking Driven by Application Constraints. In *IEEE Multi Media Modelling conference (MMM)*, 2006.

[6] C. Constantin, D. Gross-Amblard, and M. Guerrouani. Watermill : an Optimized Fingerprinting System for Highly Constrained Data. In *ACM MultiMedia and Security Workshop*, New York City, New York, USA, January 1–2 2005.

[7] C. Constantin, D. Gross-Amblard, M. Guerrouani, and J. Lafaye. Logiciel Watermill. `http://watermill.sourceforge.net`.

[8] D. Gross-Amblard. Query-Preserving Watermarking of Relational Databases and XML Documents. In *Symposium on Principles of Databases Systems (PODS)*, pages 191–201, 2003.

[9] M. Guerrouani. Tatouage de documents xml contraints. Technical report, Rapport scientifique CEDRIC - Mémoire d'ingénieur CNAM, 2005.

[10] S. Khanna and F. Zane. Watermarking maps : hiding information in structured data. In *Symposium on Discrete Algorithms (SODA)*, pages 596–605, 2000.

[11] J. Lafaye. Enhancing security of Web Services Workflows using Watermarking. Technical report, Rapport scientifique CEDRIC - Master Thesis Report, 2004.

[12] J. Lafaye. An analysis of database watermarking security. In *IAS*, pages 462–467. IEEE Computer Society, 2007.

[13] J. Lafaye. On the complexity of obtaining optimal watermarking schemes. In *6th International Workshop on Digital Watermarking (IWDW'07)*, pages 462–467, Guangzhou, China, December 2007.

[14] J. Lafaye, J. Béguec, D. Gross-Amblard, and A. Ruas. Invisible graffiti on your buildings : Blind and squaring-proof watermarking of geographical databases. In D. Papadias, D. Zhang, and G. Kollios, editors, *SSTD*, volume 4605 of *Lecture Notes in Computer Science*, pages 312–329. Springer, 2007.

[15] J. Lafaye and D. Gross-Amblard. XML streams watermarking. In *IFIP WG 11.3 Working Conference on Data and Applications Security (DBSEC)*, 2006.

[16] J. Lafaye, D. Gross-Amblard, C. Constantin, and M. Guerrouani. Watermill : An optimized fingerprinting system for databases under constraints. *IEEE Trans. Knowl. Data Eng. (TKDE)*, 20(4) :532–546, 2008.

[17] A. Mechouche. Tatouage de données géographiques. Technical report, Rapport scientifique CEDRIC - Rapport de master, 2005.

[18] B. Mungamuru and H. Garcia-Molina. Predictive pricing and revenue sharing. In C. H. Papadimitriou and S. Zhang, editors, *WINE*, volume 5385 of *Lecture Notes in Computer Science*, pages 53–60. Springer, 2008.

[19] J. F. Nash. Equilibrium points in n-person games. *Proc. of the National Academy of Sciences*, 1950.

[20] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge university Press, 2007.

[21] R. Sion, M. Atallah, and S. Prabhakar. Rights protection for relational data. In *International Conference on Management of Data (SIGMOD)*, 2003.

[22] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

David Gross-Amblard
Major Publications in Recent Years

International journals

1. **David Gross-Amblard. Query-Preserving Watermarking of Relational Databases and XML Documents. *ACM Transactions on Database Systems (ACM TODS)*, 36(1) :3 (2011).**
2. **Julien Lafaye, David Gross-Amblard, Camélia Constantin and Meryem Guerrouani. Watermill : an optimized fingerprinting system for highly constrained data. IEEE Transactions on Knowledge and Data Engineering (TKDE) (accepted 9/2007), April 2008 (Vol. 20, No. 4) pp. 532-546.**
3. **David Gross-Amblard and M. de Rougemont. Uniform generation in spatial constraint databases and applications. In *Journal of Computer and System Sciences (JCSS)*, 72(4) : 576-591, June 2006.**

National journals

1. **Sonia Guéhis, David Gross-Amblard, Philippe Rigaux. Un modèle de production interactive de programmes de publication. Ingénierie des Systèmes d'Information (Networking and Information Systems),revue des sciences et technologies de l'information (RTSI) série ISI, 13 (5), pp. 107-130, octobre 2008.**
2. **Camelia Constantin, Bernd Amann and David Gross-Amblard. Un modèle de classement de services par contribution et utilité. In *Revue des sciences et technologies de l'information* (numéro special "Recherche d'information dans les systemes d'information avances") (1633-1311) - 12(1), pp.33-60, 2007.**

International conferences with peer review

1. David Baelde, Pierre Courtieu, David Gross-Amblard and Christine Paulin-Mohring. Towards Provably Robust Watermarking. In Interactive Theorem Proving, Princeton, USA, August 2012.
2. Fabian M. Suchanek, David Gross-Amblard, Serge Abiteboul : Watermarking for Ontologies. In Proceedings of International Semantic Web Conference (1) 2011 : 697-713.
3. Sonia Guehis, David Gross-Amblard and Philippe Rigaux. Publish By Example. In Proceedings of IEEE International Conference on Web Engineering (ICWE'08), 14-18 Juillet 2008, Yorktown Heights, New York.
4. Julien Lafaye, Jean Béguec, David Gross-Amblard and Anne Ruas. Invisible Graffiti on your Buildings : Blind & Squaring-proof Watermarking of Geographical Databases. In *10th International Symposium on Spatial and Temporal Databases (SSTD)*, July 16-18, 2007, Boston. LNCS 4605, pages 312-329.
5. Julien Lafaye and David Gross-Amblard. XML Streams Watermarking. In *20th Annual IFIP WG 11.3 Working Conference on Data and Applications Security (DBSec2006)*, Sophia Antipolis, France, 7/31 - 8/02 2006, pages 74–88.
6. Camélia Constantin, Bernd Amann, David Gross-Amblard. A Link-Based Ranking Model for Services. In *Cooperative Information Systems (CoopIS) International Conference, 2006*, pages 327-344.
7. Multimedia and Metadata Watermarking Driven by Application Constraints, avec Richard Chbeir, In IEEE Multi Media Modelling conference (MMM), 8 pp., 2006.

National conferences with peer review, informal proceedings

1. Publication de données par l'exemple. Sonia Guéhis, David Gross-Amblard et Philippe Rigaux. In *Journées nationales Bases de données avancées (BDA 2007)*, Marseille, France, 23/26-10 2007.

2. Invisible Graffiti on your Buildings : Blind & Squaring-proof Watermarking of Geographical Databases. Julien Lafaye, Jean Béguec, David Gross-Amblard and Anne Ruas. In *Journées nationales Bases de données avancées (BDA 2007)*, Marseille, France, 23/26-10 2007.
3. Camélia Constantin, Bernd Amann, David Gross-Amblard. A Link-Based Ranking Model for Services. In *Journées nationales Bases de données avancées*, Lille, France, 10/17-20 2006.

Softwares

1. Camélia Constantin, David Gross-Amblard, Meryem Guerrouani et Julien Lafaye. *`Watermill` : database watermarking with optimized constraint preservation.*
   `http://watermill.sourceforge.net`
2. Julien Lafaye et Jean Béguec. Geographic data watermarking library Watergoat (OpenJump plugin).
   `http://cedric.cnam.fr/~lafaye_j/index.php?n=Main.WaterGoatOpenJumpPlugin`
3. Sonia Guehis. Web publishing-by-example DocQL suite.
   `http://www.lamsade.dauphine.fr/~guehis/docql/`

Contribution to the Activity Report of the Department *Data and Knowledge Management*

February 10, 2013

Israel César Lerman
Professeur émérite, Université de Rennes 1, Irisa
Département Data and Knowledge Management, Irisa

1 Association Rules, Clustering and Data Mining

1.1 Association Rules and Data Mining

1.1.1 Overview; Position of the Problem

Building a relevant interestingness measure for association rules is a fundamental problem in *Data Mining* [GHe07]. We assume a context defined by a data table crossing a set $\mathcal{A}$ of descriptive attributes with a set $\mathcal{O}$ of objects described. The latter is generally given by a training set provided from a universe $\mathcal{U}$ of objects. The most important and basic case is that where $\mathcal{A}$ is constituted by Boolean attributes. Extension to other types of descriptive attributes is also studied in many research works.

Let a and b be two Boolean attributes from $\mathcal{A}$, a statistical *association rule* (also called *implication rule*) is denoted symbolically by $a \rightarrow b$. Intuitively, it means: "If the attribute a is *true* on a given object o belonging to $\mathcal{O}$, then, generally but not absolutely, b is *true* on o. In these conditions, the matter is to assess this statistical tendency. As in logics, a and b are called *premise* and *conclusion*, respectively. This evaluation is obtained by means of a numerical index. Many indices have been proposed in the literature. All of them consider only the two attributes a and b to be compared. One important facet of the originality of our approach consists in taking into account the strength of the

[GHe07] F. GUILLET and H.J. HAMILTON eds. *Quality measures in data mining, Studies in Computational Intelligence, vol. 43.* Springer, 2007.

association $a \rightarrow b$ in a relative manner, with respect to the set $\mathcal{A} \times \mathcal{A}$ of all ordered attribute pairs.

Likelihood Linkage Analysis Classification approach leads to a powerful and fine tool for clustering and data analysis of complex data [5, 4, 13, 12, 2, 7]. All mathematical types of data can be processed by this method. It is based on two principles:

1. Set theoretic and relational mathematical representation of the descriptive attributes with respect to the object set $\mathcal{O}$;

2. Probabilistic evaluation of the associations between descriptive attributes and of the similarities between objects or categories.

In [5] a very large range of data types are clearly specified, according to item 1. The probabilistic evaluation - mentioned in item 2 - is obtained with respect to an adequate independence probabilistic hypothesis between the descriptive attributes. This method provides a probabilistic association coefficient between Boolean attributes. The latter is symmetrical and for an ordered pair of Boolean attributes (a, b), it expresses a measure of statistical equivalence degree between a and b. We can denote this symbolically by $a \leftrightarrow b$.

The idea to adapt this symmetrical index to the asymmetrical implicative case mentioned above, was proposed, studied and applied [GRA79,LGR81]. It is mainly a local version of this index, restricted to the comparison of a single ordered pair (a, b) of Boolean attributes which is considered in the cited references. However, this local form of the probabilistic index tends - when the object set size increases - towards one of two values 0 and 1, 0 in the repulsive case and 1 in the attractive one. These two cases are defined with respect to a statistical independence hypothesis.

Now, generally, the data size is extremely large in *Data Mining* and then, it is imperious to have a discriminant probabilistic index for interestingness measure of association rules.

1.1.2 Association Rules and Data Mining; New Results

For pairwise comparison of an attribute set $\mathcal{A}$, a simple and natural normalization technique is applied in the LLA agglomerative hierarchical clustering. A probability scale is obtained from standardized indices [13]. The latter is finely discriminant for comparing pairwise associations between descriptive attributes.

[GRA79] R. GRAS. *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques, Doctorat d'État.* PhD thesis, Université de Rennes 1, 1979.

[LGR81] I.C. LERMAN, R. GRAS, and H. ROSTAM. Élaboration et évaluation d'un indice d'implication pour des données binaires i et ii. *Mathématique et Sciences Humaines*, (74-75):5–35, 5–47, 1981.

Mathematical and statistical justifications were provided for this normalization technique [LER84]. On the other hand, experimental analysis has validated this approach. We have called it *global reduction of the similarities*. This method was transposed to the asymmetrical implicative case. Its limit behaviour was studied with respect to an increasing model of the object set $\mathcal{O}$, this model being consistent with the *Data Mining* issue [6].

Obtaining a probabilistic discriminant measure of the *Likelihood of the Link* for association rules is also an objective in [RM08]. For this approach the data are summarized by means of a hypothetical sample of size 100. Then, the notion of *TestValue* is applied to the latter sample. Note that for the deduced measure denoted $TV_{percent}$ the basic notion of a statistical data unit is no longer respected.

An extensive theoretical, methodological and experimental analysis [8] has been carried out in order to compare different approaches where a probabilistic index of the *Likelihood of the Link* takes part. This analysis is based on increasing models of the number of objects. On the other hand, variations of the level and the nature of the link between *premise* and *conclusion* for a given association rule, are considered in this analysis. The mathematical and experimental results confirm the validity of our normalization method.

Two major aspects of the previous work gave rise to two important contributions to the EGC2011 conference [9] and [1]. The first paper is focused on the mathematical and statistical comparisons between the *Likelihood of the Link* and $TV_{percent}$ measures. The second paper is more concerned by methodological and experimental analysis. The latter validates the mathematical results obtained and leads to a greater depth in investigating the convergence phenomenon with respect to the simulated models of object set size increasing, mentioned above.

Our contribution was selected among the ten best ones of the EGC'2011 conference. This has led to the article *Comparing two discriminant probabilistic interestingness measures for association rules* [10]. This paper is a substantial development of [9]. A project of a second article focused on the second paper [1], to submit to an international journal, is in preparation.

1.1.3 Clustering and Data Mining; New Results

Let us take up again the case where the set $\mathcal{A}$ of Boolean attributes is endowed with a symmetrical association coefficient. The agglomerative construction of a classification tree is based on a symmetrical notion of association measure between the built up clusters in the agglomerative process [13, 7]. It leads to

[LER84] I.-C. LERMAN. Justification et validité statistique d ' une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées. *Publications de l'Institut de Statistique des Universités de Paris*, (3-4):XXIX, 27–57, 1984.

[RM08] R. RAKOTOMALALA and A. MORINEAU. The tvpercent principle for the counterexamples statistic. In F. Guillet R. Gras, E. Suzuki and F. Spagnolo, editors, *Statistical Implicative Analysis*, pages 449–462. Springer, 2008.

the discovery of significant behaviour profiles and subprofiles in the universe described.

Now, let us consider the case where the attribute set $\mathcal{A}$ is endowed with an index of implication, defining an association rule coefficient on $\mathcal{A}$. This index is asymmetrical. Therefore, a requested condition for building a classification tree on $\mathcal{A}$, is to reflect this asymmetry. The formation of an implicative tree is proposed in [GL93]. In this, the link between two clusters is directed (for example, from left to right). In [3] we provide a global analysis of this directed tree structure. Indeed, it is important to realize clearly the transposition of the classical symmetrical construction to the asymmetrical one. All the facets of this construction are studied: logics, combinatorics, statistics and algorithmic.

The sought structure called *directed hierarchy* is reexamined in a complete framework in [11]. In this work we establish in a constructive way a bijective correspondence between a directed hierarchy and a specific notion of ultrametric distance called *directed ultrametric*. This result establishes the transposition to the asymmetrical case of a very known result (the Johnson correspondence) obtained in the classical and much simpler symmetrical case.

1.2 Software

$CHAVLH$ (Classification Hiérarchique par Analyse de la Vraisemblance des Liens en cas de données Hétérogènes) [PLL05] is the software which implements the Likelihood Linkage hierarchical agglomerative clustering. For a description of an object set $\mathcal{O}$ the following types of descriptive attributes are provided:

1. Numerical;
2. Boolean;
3. Nominal categorical;
4. Ordinal categorical;
5. Categorical, endowed with an ordinal or numerical similarity between its values.

For the latest type, the attribute is called *preordonance* attribute.

Such a description is represented by a classical data table crossing the object set $\mathcal{O}$ with an attribute set $\mathcal{A}$. Clustering $\mathcal{O}$ can be carried out when the attribute set $\mathcal{A}$ is constituted by attributes of

---

[GL93] R. GRAS and A. LARHER. L'implication statistique, une nouvelle méthode d'analyse des données. *Mathématiques (, Informatique) et Sciences Humaines*, (120):5–31, 1993.

[PLL05] P. PETER, H. LEREDDE, and I.C. LERMAN. Notice du programme CHAVLH (Classification Hiérarchique par Analyse de la Vraisemblance des Liens en cas de variables Hétérogènes). Dépôt APP (Agence pour la Protection des Programmes) IDDN.FR.001.240016.000.S.P.2006.000.20700, Université de Rennes 1, Décembre 2005.

- one single type;
- different types.

Clustering the attribute set $\mathcal{A}$ requires a single type for all of the attributes. However, *preordonance* coding can be considered for all of the descriptive attributes [OA91].

The software $AVARE$ (Association entre VAriables RElationnelles) calculates the symmetrical association coeficients table between such attributes [OA00]. This software has been integrated in $CHAVLH$ in 2011 by Philippe Peter.

Two other types of a data table can be handled by $CHAVLH$:

- Pairwise dissimilarity table between objects, directly provided by expert knowledge or other sources;
- Horizontal juxtaposition of contingency tables.

$CHAVLH$ is very used. More particulary, it has been applied in many research works at the IRISA institute. It has played an important role in the validation of the results of the thesis of Noel Malod-Dognin: "Protein Structure Comparison: From Contact Map Overlap Maximization to Distance-based Alignement Search Tool", defended in 2010.

$CHAVLH$ is implemented in "GenOuest Bioinformatics Platform" of *Symbiose* project, as a clustering tool. Interfacing project is envisaged in order to optimize its use.

Since July 2007, an ergonomic and simplified version of $CHAVLH$, called $LLAhclust$ (Likelihood Linkage Analysis $h$ierarchical $clust$ering), is implemented in the **R** software environment (I. Kojadinovic (École Polytechnique de l'Université de Nantes), I.-C. Lerman, P. Peter and N. Le Meur de l'Irisa).

1.3 Scientific Committees and Editorial Boards

I.-C. Lerman was a PC member of the *EGC2011* conference, *Extraction et Gestion de Connaissances*, January 2011, Brest, France.

I.-C. Lerman was a PC member of the *EGC2012* conference, *Extraction et Gestion de Connaissances*, January 2011, Bordeaux, France.

I.-C. Lerman is a member of the editorial board of the journal "Mathématiques et Sciences Humaines, *Mathematics and Social Sciences*", Paris.

I.-C. Lerman was in 2011 "Special Reviewer" of the Journal of Classification", New York.

[OA91] M. OUALI-ALLAH. *Analyse en préordonnance des données qualitatives. Application aux données numériques et symboliques.* PhD thesis, Université de Rennes 1, décembre 1991.

[OA00] M. OUALI ALLAH. Programme de calcul de coefficients d'association entre variables relationnelles. *La Revue de Modulad*, (25):63–74, 2000.

1.4 National Collaborations

- Sylvie Guillaume, Université de Clermont, Auvergne, LIMOS, Clermont Ferrand ;
- Pascale Kuntz, Université de Nantes, Laboratoire d′Informatique de Nantes Atlantique, Equipe COD, Site Polytech ′ Nantes ;
- Philippe Peter, Université de Nantes, Laboratoire d′Informatique de Nantes Atlantique, Equipe COD, Site Polytech ′ Nantes.

Major publications in recent years (2006-2012)

References

[1] S. GUILLAUME and I.-C. LERMAN. Analyse du comportement limite d'indices probabilistes pour une sélection discriminante. In A. Khenchaf et P. Poncelet, editor, *Revue de l'Information et des Nouvelles Technologies, RNTI E.20, EGC'2011*, pages 657–664. Hermann, 2011.

[2] I.-C LERMAN. Coefficient numérique général de discrimination de classes d'objets par des variables de types quelconques. *Revue de Statistique Appliquée*, (LIV(2)):33–63, 2006.

[3] I.-C. LERMAN. Analyse logique, combinatoire et statistique de la construction d'une hiérarchie binaire implicative; niveaux et noeuds significatifs. *Mathématiques et Sciences Humaines, Mathematics and Social Sciences*, (184):47–103, 2008.

[4] I.-C LERMAN. Analyse de la vraisemblance des liens ; une méthodologie d'analyse classificatoire de données relationnelles : le cas symétrique d'abord, le cas orienté ensuite. Séminaire, IRISA-INRIA, October 2012.

[5] I.-C LERMAN. Facets of the set theoretic representation of categorical data. Publication Interne 1988, IRISA-INRIA, January 2012.

[6] I-C. LERMAN and J. AZÉ. A new probabilistic measure of interestingness for association rules, based on the likelihood of the link. In F. Guillet and H.J. Hamilton, editors, *Quality measures in data mining, Studies in Computational Intelligence, vol. 43*, pages 207–236. Springer, 2007.

[7] I.-C. LERMAN and K. BACHAR. Comparaison de deux critères en classification ascendante hiérarchique sous contrainte de contiguïté. *Journal de la Société de Statistique de Paris et Revue de Statistique Appliquée*, (149, 2):45–74, 2008.

[8] I.-C. LERMAN and S. GUILLAUME. Analyse comparative d'indices discriminants fondés sur une échelle de probabilité. Rapport de Recherche PI Irisa 1942, RR Inria 7187, IRISA-INRIA, Février 2010.

[9] I.-C. LERMAN and S. GUILLAUME. Comparaison entre deux indices pour l' évaluation probabiliste discriminante des règles d'association. In A. Khenchaf et P. Poncelet, editor, *Revue de l'Information et des Nouvelles Technologies, RNTI E.20, EGC'2011*, pages 647–656. Hermann, 2011.

[10] I.-C LERMAN and S. GUILLAUME. Comparing two discriminant probabilistic interestingness measures for association rules. In G. Venturini F. Guillet, B. Pinaud and D. Zighed, editors, *Advances in Knowledge Discovery and Management 3*, page to appear. Springer, 2013.

[11] I.-C. LERMAN and P. KUNTZ. Directed binary hierarchies and directed ultrametrics. *Journal of Classification*, (28):272–296, October 2011.

[12] I.-C. LERMAN and P. PETER. Representation of concept description by multivalued taxonomic preordonance variables. In G. Cucumel P. Brito, P. Bertrand and F. Carvalho (eds), editors, *Selected Contributions in Data Analysis and Classification*, pages 271–284. Springer, 2007.

[13] I.C. LERMAN. Analyse de la vraisemblance des liens relationnels une méthodologie d ' analyse classificatoire des données. In Younès Benani and Emmanuel Viennet, editors, *RNTI A3, Revue des Nouvelles Technologies de l'Information*, pages 93–126. Cèpaduès, 2009.

Zoltán Miklós
Maitre de conferences, Université de Rennes 1

1 Overall objectives

My recent work has focused on semantic interoperability establishment techniques on the Web as well as in business-oriented context. I have joined IRISA in September, thus I am gradually building up research collaboration with Pr. David Gross-Amblard. At the same, I am phasing out and completing earlier research efforts. I co-supervised PhD students at EPFL: Surrender Reddy Yerva (since 2008, defence foreseen for spring 2013), and Nguyen Quoc Viet Hung (since 2010), whom I continue to co-supervise (defence in 2014). Our work with Hung opens a perspective on interoperability establishment, where we consider a network context for the interoperability establishment between database schemas or Web forms. Our works with Surender addressed several questions of entity identification in Web data (Web page collections, twitter messages).

2 Scientific Foundations

2.1 Past research activities

Novel uses of information technology and in particular the Web create new challenges for computer science and for data management in particular. Applications need to handle large amounts of data, often originating from multiple sources. In this development, we believe that the major challenges are (1) to design efficient algorithms that can handle large instances of computational problems, (2) to manage uncertainty and incompleteness of the data and (3) to cope with the semantic and structural heterogeneity of data.

We believe that in order to make progress on these challenges, one needs to rely on the deep theoretical results, adapt and extend them to novel settings. At the same time experimental and application-oriented research is also essential, as certain aspects of information systems are too complex to model them theoretically. My research work is on the borderline between artificial intelligence and data management. I believe that these two fields mutually enrich each other, and these connections will be even more tight in the future.

Constraint satisfaction, database theory, hypergraph decompositions I worked on the subject of hypertree decompositions that is a technique to manage large instances of constraint satisfaction problems. The concept was originally introduced in the context of databases, to efficiently evaluate conjunctive queries. My work in this area focused on generalizations of the hypertree decomposition concept. My main results in this area

are reported in [J4] (that is an extended version of [C6]). Other publications include [B1],[W5] and my PhD thesis.

Data integration, semantic interoperability I worked on several aspects of data integration and semantic interoperability. In particular, I studied ontology-based data integration, while later, my focus was more the establishment of semantic correspondences, in the absence of commonly agreed ontologies.

My main contributions involved developing a view-model for RDF, with the help of a rule-based language [C8]. We have in the ELENA and Prolix projects, for various integration problems in the application domain of eLearning [J5],[C8], [W8]. My work also involved developing software prototypes, that I have built on top of a P2P system (edutella) [O4], [O5]. I worked on a similar approach later also in the Team project, in the software engineering context.

My recent work at EPFL was linked to the NisB project [**?**], whose proposal I have co-written. The project addresses semantic interoperability establishment for business networks. The project focuses on gradual interoperability establishment [O2], with the help of small reusable units of interoperability [M2], in a network of databases. I have a number of publications in preparation related to this project, in particular related to the reconciliation process.

Web data, entity resolution Entity-oriented information systems aim to handle entities (such as persons, companies, geographic locations) as first-class citizens, as opposed to the current document oriented Web. I supervised in the course of the OKKAM project the development of a large-scale infrastructure for entity search, in particular its storage component [C5], [O3], [**?**]. My main contributions in this area are the development of entity matching and classification techniques for Web data [J2], [J1]. Our research prototype won the WePS-3 challenge on entity matching [W2]. Other aspects of these problems are studied in [W3], [C2]. I also contributed to a research that addresses quality evaluations for ontologies, w.r.t. a given taxonomy (WordNet), by suggesting appropriate mathematical techniques [J3], [C3]. My recent works on Web data are related to the PlanetData NoE, which deals with large-scale data management, linked data and sensor data management [W1].

Other subjects Besides my main interests I also worked on complementary subjects, that have also close relation to my main research interests. I contributed to the OPELIX project to a software framework for information commerce on the Internet [C9]. In this context, I designed and analyzed an access-control model for distributed event-based system [W10] and also a novel trust delegation mechanism [W9]. Later, in the Team project, we analyzed access control [W4] and privacy mechanisms [C4] for P2P systems.

3 New results

Our recent submission [M1] (paper under review) demonstrated that the network context can successfully used to reduce the necessary human effort in the semantic interoperability establishment process. A variant of the this work is [C1], where we demonstrate a similar reduction effect, but here we utilize the effort of the crwod (i.e. we simulate a croudsourcing scenario), instead of a single expert. Additionally we have also submitted a demo paper, that concerns again an another variant of the problem, namely a setting where a small group of experts collaboratively solve the same problem. This is a different setting as the crowdsourcing scenario, as in that case the participants do not communicate among each other.

4 Dissemination of results

4.1 Journal and Conference Reviewing

I was acting as a referee for the following journals and conferences. I list here also other related refereeing activities.

Journals: Journal of the ACM, Artificial Intelligence (Elsevier), Journal of Computer and System Sciences (Elsevier), ACM Transactions on Database Systems (TODS), Information Processing Journal (Elsevier), The VLDB Journal, IEEE Transactions of Data and Knowledge Engineering, Journal of Web Semantics, Computer Networks (Elsevier), IEEE Transactions on Parallel and Distributed Systems, Future Generation Computer Systems (Elsevier), Information Systems (Elsevier), IEEE Intelligent Systems Magazine

Conferences: ECIS'2013 (PC member), ESWC'2012 (PC member), VLDB 2010, 2011, Computer Science Logic (CSL'10), Foundations of Software Technology and Theoretical Computer Science (FSTTCS'2010), 13th International Workshop on the Web and Databases (WebDB 2010), Workshop on Semantic Data Management (VLDB'10 workshop) (PC member), IEEE International Conference on Web Services 2010 (ICWS'10), Graph Theoretic Concepts in Computer Science (WG'2010), Word Wide Web conference (WWW'2010, 2009), ODBASE'2009, ACM SIGMOD 2009, International Conference on Distributed Computing Systems (ICDCS'09), ACM Conference on Distributed Event-Based Systems (DEBS'09), International Workshop on Managing data with Mobile Devices (MDMD'09), 8th International Semantic Web Conference (ISWC'2008), International Conference on Peer-to-Peer Computing (P2P'2008, 2010), 19th International Conference on Database and Expert Systems Applications (DEXA'2008), International Colloquium on Automata, Languages and Programming (ICALP'2008), Logic in Computer Science (LICS'2008), Principles of Database Systems (PODS'2008), Computer Science Logic (CSL'2007), Mathematical Foundations of Com-

puter Science (MFCS'2007), Symposium on Theoretical Aspects of Computer Science (STACS'2006)

Doctoral consortium: I was a scientific advisor at the doctoral consortium at the Future Internet Symposium (FIS'2010).

Project proposals: Expert evaluator of EU project proposals on behalf of the European Commission for the FP7-ICT-2011-Call 8 SO 4.4 (Intelligent Information Management), February 2012.

4.2 Invited Talks

I gave the following invited talks at university seminars and workshops (in the recent years):

- Towards understanding the cost/benefit tradeoff for data management techniques, Université Lille 1, March 2012
- Towards emergent semantics: concepts and computational techniques, Université de Bourgogne, Dijon, March 2012
- From query processing to information integration: the power of decomposition techniques, Université 2, Montpellier, France, February 2012
- From query processing to information integration: the power of decomposition techniques, Université Paris 6, Paris, France, November 2011
- Understanding tractable decompositions for conjunctive queries, Israel Institute of Technology (Technion), Haifa, Israel, June 2011
- Divide and conquer techniques for data management problems, TU Berlin, Germany, May 2011
- From problem decompositions to semantic interoperability: a journey between databases and artificial intelligence, Université Paris Sud, France, March 2011

4.3 Major publications in recent years

4.3.1 Book or Book chapter

[B1] Georg Gottlob, Gianluici Greco, Zoltán Miklós, Francesco Scarcello, and Thomas Schwentick. *Graph Theory, Computational Intelligence*

and Thought. Essays Dedicated to Martin Charles Golumbic on the Occasion of His 60th Birthday, volume 5420 of *LNCS*, chapter Tree Projections: Game Characterization and Computational Aspects, pages 87–99. Springer, 2009.

4.3.2 International Journal with PC

[J1] Surrender Reddy Yerva, Zoltán Miklós, and Karl Aberer. Entity-based Classification of Twitter Messages. *International Journal of Computer Science & Applications*, 9(1):88–115, 2012.

[J2] Surender Reddy Yerva, Zoltán Miklós, and Karl Aberer. Quality-aware similarity assessment for entity matching in Web data. *Information Systems*, 37:336–351, 2012.

[J3] Zhao Lu, Zoltán Miklós, Liang He, Song M. Cai, and Jun Z. Gu. A novel multi-aspect consistency measurement for ontologies. *Journal of Web Engineering*, 10(1), 2011.

[J4] Georg Gottlob, Zoltan Miklos, and Thomas Schwentick. Generalized Hypertree Decompositions: NP-hardness and Tractable Variants. *Journal of the ACM*, 56(6):1–32, September 2009.

[J5] Bernd Simon, Peter Dolog, Zoltán Miklós, Daniel Olmedilla, and Michael Sintek. Conceptualizing smart spaces for learning. *Journal of Interactive Media in Education (JIME)*, (9), 2004.

4.3.3 International Conference with PC

[C1] Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Zoltán Miklós, and Karl Aberer. On Leveraging Crowdsourcing Techniques for Schema Matching Networks. In *DASFAA 2013*, 2013.

[C2] Surender Reddy Yerva, Zoltán Miklós, and Karl Aberer. What have fruits to do with technology? The case of Orange, Blackberry and Apple. In *International Conference on Web Intelligence, Mining and Semantics (WIMS'2011)*. ACM press, 2011.

[C3] Zhao Lu, Zoltán Miklós, Songmei Cai, and Junzhong Gu. Measuring Taxonomic Consistency of Ontologies Using Lexical Semantic Relatedness. In *The Fifth International Conference on Digital Information Management (ICDIM'2010)*. IEEE, 2010.

[C4] Rammohan Narendula, Thanasis G. Papaioannou, Zoltán Miklós, and Karl Aberer. Tunable Privacy for Access Controlled Data in Peer-to-Peer systems. In *Proceedings of the 22nd International Teletraffic Congress (ITC 22)*, 2010.

[C5] Zoltán Miklós, Nicolas Bonvin, Paolo Bouquet, Michele Catasta, Daniele Cordioli, Peter Fankhauser, Julien Gaugaz, Ekaterini Ioannou, Hristo Koshutanski, Antonio Mana, Claudia Niederée, Themis Palpanas, and Heiko Stoermer. From Web Data to Entities and Back. In *The 22nd International Conference on Advanced Information Systems Engineering (CAiSE'10)*, volume 6051 of *LNCS*, pages 302–316. Springer, 2010.

[C6] Georg Gottlob, Zoltán Miklós, and Thomas Schwentick. Generalized hypertree decompositions: NP-hardness and tractable variants. In *26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS'07)*, pages 13–22, 2007.

[C7] B. Simon, S. Sobernig, F. Wild, S. Aguirre, S. Brantner, P. Dolog, G. Neumann, G. Huber, T. Klobucar, S. Markus, Z. Miklós, W. Nejdl, D. Olmedilla, J. Salvachua, M. Sintek, and T. Zillinger. Building Blocks for a Smart Space for Learning. In *Proceedings of the 6th IEEE International Conference on Advanced Learning Technologies*, 2006.

[C8] Zoltán Miklós, Gustaf Neumann, Uwe Zdun, and Michael Sintek. Querying Semantic Web Resources Using TRIPLE Views. In *In the Proceedings of the 2nd International Semantic Web Conference (ISWC'03)*, volume 2870 of *LNCS*, pages 517–532. Springer, 2003.

[C9] Manfred Hauswith, Mehdi Jazayeri, Zoltán Miklós, Ivana Podnar, Elisabetta Di Nitto, and Andreas Wombacher. An Architecture for Information Commerce Systems. In *In the Proceedings of ConTEL'2001-6th International Conference on Telecommunications*, 2001.

4.3.4 International Workshop with PC

[W1] Ian Rolewicz, Michele Catasta, Hoyoung Jeung, Zoltán Miklós, and Karl Aberer. Building a front end for a sensor data cloud. In *The First International Workshop on Cloud for High Performance Computing (C4HPC), Special Session of the 11th International Conference on Computational Science and Its Applications (ICCSA'2011)*, 2011.

[W2] Surender Reddy Yerva, Zoltán Miklós, and Karl Aberer. It was easy, when apples and blackberries were only fruits. In *Third WePS Evaluation Workshop: Searching Information about Entities in the Web*,

2010. **Our software was the winner of the evaluation campagine.**

[W3] Surender Reddy Yerva, Zoltán Miklós, and Karl Aberer. Towards better entity resolution techniques for Web document collections. In *1st International Workshop on Data Engineering meets the Semantic Web (DESWeb'2010) (co-located with ICDE'2010)*, 2010.

[W4] Rammohan Narendula, Zoltán Miklós, and Karl Aberer. Towards Access Control-Aware P2P Data Management Systems. In *In the Proceedings of the 2nd International Workshop on Data Management in Peer-to-peer Systems (DAMAP'09)*, 2009.

[W5] Zoltán Miklós. On the parallel complexity of tractable structural CSP decompositions. In *In the Proceedings of the 3rd Workshop on Algorithms and Complexity in Durham (ACiD'07)*. College Publications, 2007.

[W6] Zoltán Miklós and Stefan Sobernig. Translation between RDF and XML: a case study in the educational domain. In *Workshop on Interoperability of Web-Based Educational Systems in conjunction with 14th International World Wide Web Conference (WWW'05)*, pages 27–34, 2005.

[W7] B. Simon, D. Massart, F. van Assche, S. Ternier, E. Duval, S. Brantner, D. Olmedilla, and Z. Miklós. A Simple Query Interface for Interoperable Learning Repositories. In *Workshop on Interoperability of Web-Based Educational Systems in conjunction with 14th International World Wide Web Conference (WWW'05)*, 2005.

[W8] S. Decker, M. Sintek, A. Billig, N. Henze, P. Dolog, W. Nejdl, A. Harth, A. Leicher, S. Busse, J. Guy Suess, Z. Miklós, J. Ambite, M. Weathers, G. Neumann, and U. Zdun. Triple - and RDF rule language with context and use cases. In *In W3C Workshop on Rule Languages for Interoperability*, 2005.

[W9] Zoltán Miklós. A decentralized authorization mechanism for e-business applications. In *In Proceedings of the 13th International Workshop on Database and Expert Systems Applications (DEXA 2002) - International Workshop on Trust and Privacy in Digital Business - TrustBus*, pages 446–450, 2002.

[W10] Zoltán Miklós. Towards an Access Control Mechanism for Wide-Area Publish/Subscribe Systems. In *In the Proceedings of the 22nd International Conference on Distributed Computing Systems, Workshops (ICDCSW '02) - DEBS - International Workshop on Distributed Event-Based Systems (DEBS'02)*, pages 516–524, 2002.

4.3.5 Other

[O1] Surender Reddy Yerva, Zoltán Miklós, Flavia Grosan, Tandrau Alexandru, and Karl Aberer. TweetSpector: Entity-based retrieval of Tweets. In *SIGIR'2012*, 2012. (demo paper).

[O2] Golnaz Vakili, Thanasis G. Papaioannou, Zoltán Miklós, Karl Aberer, and Siavash Khorsandi. Analyzing the Emergence of Semantic Agreement among Rational Agents. In *13th IEEE Conference on Commerce and Enterprise Computing (short paper)*, 2011.

[O3] Ekaterini Ioannou, Saket Sathe, Nicolas Bonvin, Anshul Jain, Srikanth Bondalapati, Gleb Skobeltsyn, Claudia Niederée, and Zoltán Miklós. Entity Search with NECESSITY (demo paper). In *12th International Workshop on the Web and Databases (WebDB 2009)*, 2009.

[O4] Sandra Aguirre, Stefan Brantner, Gernot Huber, Sascha Markus, Zoltán Miklós, Alberto Mozo, Daniel Olmedilla, Joaquin Salvachua, Bernd Simon, Stefan Sobernig, and Thomas Zillinger. Corner Stones of Semantic Interoperability Demonstrated in a Smart Space for Learning (demo paper). In *2nd Annual European Semantic Web Conference (ESWC'2005)*, 2005.

[O5] Bernd Simon, Zoltán Miklós, Wolfgang Nejdl, Michael Sintek, and Joaquin Salvachua. Smart Space for Learning: A Mediation Infrastructure for Learning Services. In *In Proceedings of the Twelfth International Conference on World Wide Web (Alternate Paper Tracks)*, 2003.

4.3.6 Manuscripts

[M1] Nguyen Quoc Viet Hung, Tri Kurniawan Wijaya, Zoltán Miklós, Karl Aberer, Eliezer Levy, Victor Shafran, Avigdor Gal, and Matthias Weidlich. Reconciling Matching Networks of Web-Schemas. In *(submitted)*.

[M2] Karl Aberer, Avigdor Gal, Michael Katz, Eliezer Levy, Zoltán Miklós, Nguyen Quoc Viet Hung, Tomer Sagi, and Victor Shafran. A Generalized Cover Problem for Schema Matching. (submitted to a conference with PC), 2012.