Singletons associated with DKM

Mickaël Foursov, David Gross-Amblard, Israel César Lerman, Virginie Sans

Rennes

Activity Report

2011

Mickaël Foursov
Assistant 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

- Mickal 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

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

References

Publications in 2011

[1] F. Benmakrouha, M.V. Foursov, C. Hespel, J.-P. Hespel and E. Monnier, *Modélisation de la Glycémie d'un Patient Diabétique : une Application Floue*, Infusystèmes, 2011, Vol. 28, No. 3, pp.27–30.

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 1) database watermarking, and 2) data and services management on the Web. As a newcomer at IRISA, my research project may involve the following aspects :

- short term : obtaining formal security proofs on specific watermarking protocols for numerical databases,
- middle term : moving from the intellectual property protection of data (watermarking) to the assessment of its provenance and trust, in a distributed context,
- long term : taking into account the strategic aspect of data (for example its price or impact), probably using tools from (economical) game theory.

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 [9, 3, 20].

I started a database watermarking working group at the Vertigo team. 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) [7]. A practical counterpart of this work has been proposed to obtain a full database watermarking prototype, Watermill [5, 8, 15, 6].

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) [16, 10, 14, 11, 12, 13, 15] ;
- *Medical images* watermarking under constraints [4] ;
- *Symbolic musical databases* watermarking [2].

Data Provenance and Trust : Classical Web and Web of Objects Faced to the Web, Database technique 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, in new.

Impact of Web data : strategic aspects 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 [17].

Those questions are hard to apprehend, because the value of a data is no longer a locally defined property, but a property that emerge 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 [21] and popularized by Nash's result [18], 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 [19].

3 New Results

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.

A work in that direction is in progress with members of the ANR SCALP project, which goal is to certify cryptographical protocol proofs using proof assistants like Coq. We obtained with David Baelde, Pierre Coutieu, Julien Lafaye, Philippe Audebaud et Xavier Urbain a restricted proof of the Agrawal and Kiernan protocol. We are trying now to extend recent works on so-called strong watermarking protocols.

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.

4 Dissemination of Results

Students

Ph.D students

- (running) Joint direction (33%) with Lylia Abrouk (33%) and Christophe Nicolle (33%) of Damien Leprovost's thesis (Bourgogne Young Entrepreneur Funding) entitled "Community discovery by semantic analysis", started September 2009.
- 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.

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

Neuma [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).

Tadorne [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`) ;
- External participant of SemWeb et 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

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

[5] 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.

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

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

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

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

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

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

[12] 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.

[13] 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.

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

[15] 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.

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

[17] 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.

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

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

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

[21] 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. Fabian M. Suchanek, David Gross-Amblard, Serge Abiteboul : Watermarking for Ontologies. In Proceedings of International Semantic Web Conference (1) 2011 : 697-713.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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/`

Israel César Lerman
Professor emeritus, Univeristé de Rennes 1

1 Association Rules, Clustering and Data Mining

1.1 Association Rules and Data Mining

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 described objects. The latter is generally given by a training set 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 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 take only into account 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 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 [10, 9, 2, 5]. 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.

The latter evaluation is obtained with respect to an adequate independence probabilistic hypothesis between the descriptive attributes. This method includes 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 that 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.

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

[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.

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 an association rule.

1.1.1 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 [10]. The latter is finely discriminant for comparing pairwise associations between descriptive attributes.

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 has been transposed to the asymmetrical implicative case. Its limit behaviour has been studied with respect to an increasing model of the object set $\mathcal{O}$, this model being consistent with the *Data Mining* issue [4].

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 [6] 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 [7] 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 obtained mathematical results 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 EGC2011 conference. And we were invited to submit an article to international post proceedings, published by Springer. The submitted paper "Comparing two discriminant probabilistic interestingness measures for association rules" is mainly focused on the development of the first paper [7]. A project of a second article to submit to an international journal and mainly focused on the second paper [1], is in preparation.

1.1.2 Clustering and Data Mining; New Results

Let us retake for a moment 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 [10, 5]. It leads to the discovery of significant behaviour profiles and subprofiles in the described universe.

Now, let us consider the case where the attribute set $\mathcal{A}$ is endowed with an index of implication, defining a rule association coefficient on $\mathcal{A}$. This index is asymmetrical. Therefore, a requested condition for building a classification tree on $\mathcal{A}$, is to reflect the asymmetry of the association coefficients. 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

[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.

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

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 [8]. 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

- 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 $V$$A$riables 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 part 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", attended 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 and P. Peter).

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

[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.

1.3 Scientific Committees and Editorial Boards

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

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.

1.5 Publications in 2011

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

1.6 Major publications in recent years (2006-2011)

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 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.

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

Virginie Sans
Maître de conférences, Université de Rennes 1

1. AFFILIATION

Enseignante chercheuse de l'Université de Rennes 1, mes recherches s'articulent autour des données semi-structurées et de leur interrogation, cette problématique permet d'appréhender des problèmes plus généraux comme les web services et la prise en contraintes dans les systèmes de gestion de données.

Dans le cadre de mon intégration à l'IRISA, j'ai eu des contacts avec les équipes Tex-Mex, Distrib-Com, LIS et KERDATA. Je devrais très prochainement me positionner sur une de ces équipes. Mon choix se porte préférentiellement sur l'équipe LIS puisque mes récents travaux sur les bases de données Saas et la personnalisation de requêtes s'intégreraient bien dans la thématique SIG : systèmes d'information géographiques de l'équipe LIS.

En tant que nouvelle MCF, j'ai demandé à bénéficier du dispositif de décharge de service en totalité sur la deuxième année, je pourrais ainsi m'intégrer au mieux à l'équipe choisie.

Je suis également membre des GDR ISI et I3, et membre des sociétés IEEE, et ACM.

Page Web : http://perso.univ-rennes1.fr/virginie.sans/
Equipe LIS : http://www.irisa.fr/LIS

2. THEMES DE RECHERCHE

Travaux antérieurs - Vues dans les bases de données XML :

Avant d'intégrer l'IRISA, mes recherches ont portés sur le problème de la mise à jour de données dans les bases de données XML.

Le problème de l'intégration des données hétérogènes et distribuées a été étudié au travers des travaux sur les architectures de médiation (Wiederhold et al., 1992). Une architecture de médiation est constituée d'un médiateur et d'adaptateurs.

Dans cette architecture, chaque source est associée à un adaptateur qui a pour tâches :

- l'extraction des informations de cette source via la transformation des requêtes envoyées par le médiateur en requêtes compréhensibles par la source
- la transformation de ces données dans un format approprié et compréhensible par le médiateur,
- l'envoi des données traduites au médiateur.
- De son côté, le médiateur a comme rôle principal d'intégrer les informations provenant des différents adaptateurs et de formuler, à partir de ces données, un résultat pour une application cliente.

Par exemple, un vacancier véliplanchiste pourrait faire une demande au médiateur afin de savoir quels sont les séjours comprenant le transport et l'hébergement dans un pays où la force du vent permet d'assouvir ses envies de nautisme. Dans ce cas, le médiateur envoie des requêtes XQuery aux adaptateurs dont les sources sont concernées (SGBD relationnel, SGBD Objet, SGBD Semi-structuré, Données capteur). Le médiateur récupère ainsi les réponses au format XML des différents adaptateurs, les traite et retourne à l'utilisateur un résultat à partir de ces données hétérogènes.

Le résultat d'une requête appelée également vue peut être matérialisé afin d'une part, d'optimiser le temps de réponse à des requêtes complexes, et d'autre part, de permettre une meilleure disponibilité des données dans le cas où les sources seraient partiellement ou totalement indisponibles, comme cela peut être le cas pour des données provenant de sources mobiles ou de capteurs.

L'utilisation de vues XML est un moyen d'intégrer des données provenant de sources hétérogènes. Le problème principal de la matérialisation de vues est de garder ces vues consistantes lorsque les sources sont mises à jour.

L'ensemble de ces travaux a donc permis d'aboutir à une expertise dans les domaines suivants :

- La maintenance incrémentale de vues XML à partir de sources non coopérantes (Web ou mobiles),
- Le traitement et l'évaluation de données XML ordonnées
- L'évaluation et l'optimisation de requêtes dans des bases de données XML natives.

Forte de ces travaux, j'ai essayé de trouver un thème de recherche prometteur qui pourrait donner des résultats concrets et qui me permettrait de m'insérer au mieux au sein de l'IRISA. Le thème qui me paraissait le plus répondre à mes vœux est celui des bases de données partagées.

Travaux en cours - Bases de données partagées :

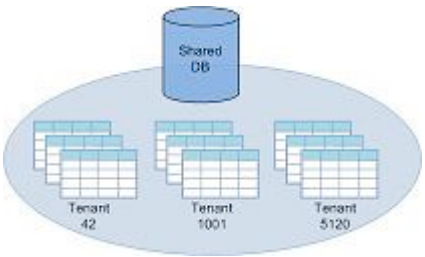
Le modèle Software as a Service (SaaS) est un modèle de déploiement de logiciels par lequel un fournisseur indépendant de logiciels (ISV) répond à une demande de plusieurs clients de disposer d'un même logiciel. L'utilisation du logiciel se fait en tant que service à la demande et peut être accessible depuis l'internet. Un fournisseur SaaS peut héberger l'application sur leurs propres serveurs Web ou proposer de télécharger l'application sur le dispositif consommateur, puis le désactiver après utilisation ou l'expiration du contrat.

Le modèle SaaS a été bien accueillie par les clients y compris les PME et les entreprises, d'une part parce qu'il supprime la charge de la construction et l'exploitation des systèmes informatiques afin d'exécuter des applications, d'autre part parce qu'il offre une certain agilité (pas besoin d'impliquer un service IT par exemple, personnalisation au besoin en fonction du client...).

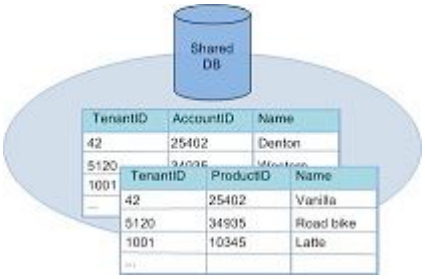
Le modèle SaaS évite ainsi les frais initiaux de licence achats de logiciels. Un fournisseur SaaS fournisse habituellement des services en échange d'un abonnement mensuel, en fonction du nombre d'utilisateurs ou de consommation des utilisateurs.

Dans un système SaaS, la conception de la base de données est l'aspect le plus critique du système. En effet, la conception de base de données doit être "extensible" pour accueillir les locataires. De manière générale, il existe trois modèles de base de données relationnelles qui peut être utilisé en mode SaaS à savoir :

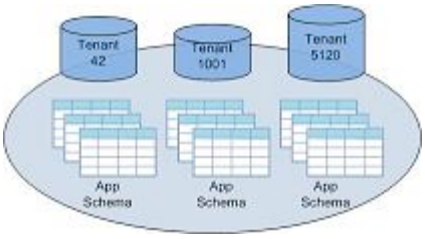
Base de données partagée, schéma distinct



Base de données partagée, schéma commun



Base de données distincte, séparée de schéma



Chaque modèle offre un niveau différent d'isolation entre les données des entreprises locataires et le choix d'un modèle ou d'un d'autre dépend des besoins du client.

La problématique des bases de données partagées dans le cas de données relationnelles est étudiée depuis peu. Cette problématique implique l'étude exhaustive des problèmes de sécurité et de personnalisation de bases de données. A l'heure actuelle, il n'existe pas de travaux portant sur l'utilisation de bases de données native XML en mode Saas.

Les données XML possèdent leurs propres propriétés, bien différentes des données relationnelles, et leur utilisation sur l'internet s'est très largement répandue pour répondre aux besoins divers des utilisateurs.

Dans le cas d'une base de données XML partagée avec un schéma commun, il est possible de stocker les données de plusieurs "petits" locataires dans le même fichier XML afin de gagner en espace disque. Cependant, cela pose des problèmes lors de la restauration de back ups, ou lors de mise à jour de données qui ne doivent porter que sur les données d'un seul locataire, etc...

Il est alors nécessaire de réfléchir à des systèmes d'indexation de données, à des systèmes de personnalisation spécifiques, à des techniques d'optimisation de requêtes propres à être utilisé pour des bases de données XML en mode partagé.

Par ailleurs, une base de données partagée disponible en mode Saas nécessite de pouvoir d'être accessible via l'internet comme un web-service classique. Dans le cas de données XML, il existe des API spécifiques (SOAP, REST, XML-RPC) pour se connecter et échanger avec un web-service.

Il existe deux types de transactions lors des interactions de services Web : la simple requête, qui renvoie des données, et la demande de traitement, qui envoie des données et en récupère d'autres en retour. Ces données sont transmises au format XML.

XML-RPC possède des défauts de spécification (manque de précisions, confusions sur certains aspects (support Unicode, notamment), mots de passe transmis en clair…) par rapport à SOAP, mais ni l'un ni l'autre n'a été spécifiquement conçu pour prendre en compte les spécificités de l'échange de données avec une base de données XML en mode Saas (multi-utilisateurs, personnalisation, sécurité accrue…).

De nombreux problèmes restent donc à soulever dans ce contexte applicatif.

Les travaux de recherche entrepris cette année ont permis de rédiger un article de revue afin d'expliquer les différents modèles Saas pour des bases de données natives XML. Les travaux futurs porteront sur l'indexation des données dans ces systèmes.

Au niveau applicatif pour les bases de données partagées, les SIG semblent être de très bons candidats. Les systèmes d'information géographique (SIG) sont constitués de matériels, de logiciels et de contenus combinés pour créer une base de données permettant le repérage et l'analyse de toute information ayant une composante spatiale. Bien qu'un SIG soit parfois conçu simplement comme un outil cartographique, c'est sa capacité à emmagasiner de l'information et à la relier à un simple point sur la carte qui fait sa force. Dans ce contexte, un SIG peut être considéré de plus en plus comme une infrastructure informatique permettant d'assembler des systèmes multi-utilisateurs volumineux et sophistiqués. Ceci étant, un GIS doit aussi être capable de répondre aux besoins de groupes de travail plus petits et d'utilisateurs individuels.

La possibilité d'accéder aux données SIG, quel que soit leur format, et celle d'utiliser plusieurs bases de données, fichiers de jeux de données, tables SGBD et services Web SIG simultanément sont deux caractéristiques importantes et qui ont font la force d'un SIG aujourd'hui. La mise en œuvre d'un SIG disponible en mode Saas et utilisant des données XML serait une bonne application à des recherches sur les bases de données natives XML partagées.

Par ailleurs, l'aspect Web-services induit dans l'utilisation Saas des bases de données partagées est une thématique déjà étudiée par l'équipe distribCom (Orchestration et techniques formelles pour les web-services, sécurité et flux d'informations).

Ce thème de recherche m'a particulièrement séduit dans la mesure où il me permettrait de collaborer activement avec des membres des équipes Distrib-Com sur l'aspect Web-Services et avec des membres de l'équipe LIS en ce qui concerne les SIG.

Depuis mon recrutement, j'ai produit 2 articles :

- Un article long pour la conférence CSE 2011
  Implementing a multimedia application on iPhone: a case study
  14th IEEE International Conference on Computational Science and Engineering (CSE-2011)

- Un article de revue :
  Multi-Tenancy and native XML Databases
  Cet article sera soumis le 30 janvier à la revue Journal of Database Management (JDM), il est à la date aujourd'hui (18 décembre 2011) en cours de relecture auprès d'une personne de langue naturelle anglaise. En effet, cette revue a une exigence particulière concernant le niveau d'anglais des articles pris en considération pour la publication.

Relecture

J'arbitre des articles dans des conférences, notamment cette année dans les conférences internationales suivantes :

- CITSA 2011 : 8th International Conference on Cybernetics and Information Technologies, Systems and Applications, http://www.iiis2011.org/imeti/website/default.asp?vc=6

- SETIT 2011: The 6th international conference Sciences of Electronic, Technologies of Information and Telecommunications. http://www.setit.rnu.tn/

3. ANIMATION SCIENTIFIQUE

Organisation de conférences :

Dans le cadre de l'animation de la communauté scientifique, j'ai organisé à Cergy Pontoise du 7 au 10 décembre 2010, la conférence SOCPAR (International Conference on Soft Computing and Pattern Recognition) . Cette conférence a accueilli plus de 110 chercheurs de 40 nationalités différentes et a été soutenue par les sociétés IEEE France et IEEE SMC Espagne.

Site SOCPAR 2010 : http://www.mirlabs.org/socpar10

En décembre 2011, j'organise également une session Data Management au sein de la conférence WICT 2011 : Word Congress on Information and Communication Technologies, qui aura lieu à Mumbai en Inde . Cette conférence est également soutenue par plusieurs chapitres IEEE SMC.

Site WICT 2011 : http://www.mirlabs.org/wict11/

Pour l'année 2012, j'ai en projet avec le laboratoire MIRLABS l'organisation d'une conférence sur le campus de Beaulieu. Le projet est actuellement à l'étude et le choix de la conférence dépendra des impératifs de date de l'IRISA ainsi que des orientations scientifiques de la conférence.

Contacts extérieurs :

Parmi les contacts que j'ai contracté cette année en tant que membre de l'IRISA, je suis entrée en contact avec Lefteris Sidirourgos, un enseignant chercheur du CWI (Hollande) qui a travaillé sur le système MonetDB/XQuery. Nous envisageons pour l'année 2012 de déposer une demande de financement Egide pour un partenariat PHC (Partenariats Hubert Curien). En effet, cette collaboration devrait permettre d'approfondir le problème de la mise à jour des données ordonnées dans les bases de données XML natives. Cette demande n'a pas pu être mise en place pour l'année 2011, les délais étant trop courts (fin mai 2011) par rapport au nouveau poste à IBM Almaden que ce collègue compte occuper.

http://homepages.cwi.nl/~lsidir/

Une fois intégrée dans ma nouvelle équipe, je souhaiterais également pour l'échéance de janvier 2012 pouvoir déposer une demande d'ANR Jeunes chercheurs. Cette demande ANR devrait être conjointe avec Nicolas Lumineau du laboratoire LIRIS à l'université de Lyon. Elle portera sur la problématique des bases de données XML partagées.

4. ACTIVITES PEDAGOGIQUES

Mes activités d'enseignement ont été concentrées sur l'enseignement des baes de données et de la programmation Web principalement au sein de la formation MIAGE du niveau L3 à M1.
J'ai enseigné à des étudiants en formation initiale ainsi qu'en alternance.

La synthèse de ces activités est présentée dans le tableau suivant :

Formation	Matières	CM	TD	TP	Eq. TD
L3 Miage	Base de données	14	16		37
L3 Miage	Programmation de clients Web	10	8	16	39
L3 Miage Alternance	Base de données	4		4	10
M1 Miage	Base de données objet	6	12	18	39
M1 Miage	Projet Miage		32		32
M2 Miage	Projet		15		15
M2 Miage	Veille technologique		12		12
MIT 2	Base de données	12	12		30
M2	Encadrement de stages		55		55
Total		**46**	**162**	**38**	**246**

Les cours de L3 Miage et de M1 Miage ayant été assuré par des collèges partis à la retraite depuis, j'ai rédigé en me basant sur les cours préexistants une nouvelle version de ces cours. J'ai ainsi rédigé les supports de CM, TD et TP des cours dont j'avais la charge.

Le cours de programmation de clients Web en L3 Miage devrait notamment subir un certain rajeunissement à la rentrée 2011 avec la mise en place d'une évaluation par projet. En effet, j'ai d'ores et déjà contacté des entreprises du bassin rennais pour qu'elles confient à nos étudiants de projets de mini-sites web. Ces projets seront à la base de la notation du cours de programmation de clients Web. Ces sites web permettront aux étudiants de se sentir plus effectivement impliqués dans leur travail et d'avoir directement des contacts avec des entreprises locales. Cela permettra également de faire connaitre nos formations auprès des entreprises.

Au mois d'avril 2011, j'ai également accompagné la promotion des M1 Miage durant leur voyage d'étude à Prague. Cela m'a permis de discuter plus avant sur leurs attentes et leur aspiration professionnelles et m'a permis de m'intégrer complètement à l'équipe enseignante de la Miage. Parmi les entreprises visitées au cours de ce voyage d'études, j'ai également pris des contacts, ces contacts pourront servir aux étudiants de l'ISTIC afin d'intégrer des entreprises de l'union européenne (UPP , Sage, Oracle) dans le cadre d'un stage ou d'un emploi.

Projets en cours, projets futurs et responsabilités :

Responsabilité pédagogique :
A la rentrée 2011, je prendrais la suite de Sophie Robin, actuellement en charge de la coordination pédagogique des L1. A cet effet, je prendrais en charge le cours de bureautique assuré cette année en L1 par Philippe Ingels.

Projet de création d'une L3 Pro :
Je discute également avec Gilles Lesventes de la possible mise en place d'une L3 Pro en alternance au sein de l'ISTIC. La Licence Professionnelle provisoirement appelée DAWEM (développement d'applications Web et Mobiles), viserait une insertion professionnelle à BAC+3. Cette licence professionnelle serait adaptée au développement du marché du multimédia de l'internet et des mobiles (portable, smartphone, PDA, tablette, etc.).

Cette licence Professionnelle s'appuierait sur une forte demande des entreprises de pouvoir recruter des collaborateurs compétents (niveau 2) dans les domaines du développement d'applications liées aux systèmes multimédias du Web et des mobiles.

Le titulaire de la Licence Professionnelle DAWEM serait alors capable de mobiliser les compétences techniques, organisationnelles pour prendre en charge aussi bien un projet multimédia de développement Web qu'un projet de programmation multimédia dédié aux mobiles (systèmes Windows Mobile, Android, IPhone).

Concernant les étudiants qui pourraient intégrer la formation, les principaux concernés seraient les étudiants du niveau L2 de l'ISTIC qui ne souhaiteraient pas poursuivre une filière longue ainsi que les étudiants qui seraient en difficulté dans une formation L3 plus généraliste.

Evaluation des enseignements :
Forte de mon expérience à l'université de Cergy Pontoise, j'ai proposé à certains enseignants d'étudier la possibilité de disposer d'un système d'évaluation des enseignements comme celui que j'avais mis en place à Cergy Pontoise. Ce système permettrait aux responsables de formation de réduire le temps nécessaire à une évaluation des enseignements qui constituent leur formation et de permettre de fournir lors de l'évaluation AERES des statistiques complètes sur nos formations. Les collègues que j'ai sollicités à cette étape du projet sont à l'heure actuelle : Didier Certain, Mikhail Foursov et Anne Grazon.

Etude de la mise en place d'un référentiel documentaire pour la gestion de projets :
J'ai eu la possibilité cette année d'encadrer un projet de M1, ainsi qu'un projet de M2. Après discussion avec Didier Certain, j'ai proposé de travailler avec lui dans un premier temps sur l'uniformisation et la mise en place d'un référentiel documentaire concernant la gestion de projets faites par les élèves. En effet, les cours de gestion de projet au sein de la MIAGE sont assez disséminés et les étudiants n'ont pas une vision globale de ce qu'est la gestion d'un projet. La mise en place de ce référentiel permettrait aux étudiants d'appréhender le cours de gestion de projets de façon progressive sur les 3 années L3, M1, et M2 et de sortir diplômés avec une expérience et un niveau de professionnalisation qui fait encore défaut à nos étudiants.

5. ENCADREMENT

En termes d'encadrement d'étudiants, j'ai eu cette année la possibilité d'encadrer :

- Un projet de M1 Miage pour une durée de 32 Heures eqTD, j'ai ainsi pu prendre un premier contact avec les exigences des entreprises vis-à-vis de nos étudiants, ainsi que le niveau global des étudiants.

- Un projet de M2 Miage pour une durée de 15 Heures eqTD, c'est à la suite de cet encadrement que j'ai proposé à Didier Certain la mise en place du référentiel documentaire pour la gestion de projets.
- Des stages de M2 (11 élèves), ces encadrements vont me permettre de me familiariser avec les entreprises du bassin rennais et ainsi pouvoir apporter un retour constructif vis-à-vis de nos enseignements et de la façon de les dispenser.

Je suis également en contact avec un doctorant de l'Institut Supérieur d'Informatique et de Multimédia de Sfax (ISIMS), Mohamed Kharrat qui souhaiterait pouvoir effectuer un stage au sein de l'IRISA et qui pourrait déboucher si l'expérience est positive sur la mise en place d'une thèse en cotutelle avec l'Université de Rennes 1.