

Research Activity Report

2004-2007

Laboratoire ERIC
Université Lumière Lyon 2
5, avenue Pierre Mendès-France
Bât L.
69600 Bron France

Tel. +33 478 772 376
Fax. +33 478 772 375
Web. <http://eric.univ-lyon2.fr>

2008 - 02 - 28

LIST OF MEMBERS OF THE TEAM

Management:

Djamel Abdelkader ZIGHED Professor, Director
Sabine LOUDCHER Assistant professor, Deputy Director

Administrative staff

Valerie GABRIELE
Julien CREVEL

Professors

Stephane LALLICH
Jean-Hugues CHAUCHAT

Assistant Professors

Fadila BENTAYEB
Omar BOUSSAID
Jérôme DARMONT
Nouria HARBI
Ricco RAKOTOMALALA
Julien VELCIN
Jacques VIALLANEIX

Assistants

Anne Muriel ARIGON
Virginie LEFORT

PhD Students

Emna BAHRI
Anouck BODIN-NIEMCZUK
Sonia BOUATTOUT
Ahmad EI SAYED
Cécile FAVRE
Rémi GAUDIN
Marouane HACHICHA

Hakim HACID
Hadj MAHBOUBI
Nora MAIZ
Simon MARCELLIN
Efthimios MAVRIKAS
Elie PRUDHOMME
Taimur QURESHI

Ony RAKOTOARIVELO
Jean-Christian RALAIVAO
Rashed Khalil SALEM
Anna STAVRIANOU
Julien THOMAS
Zhihoua WEI

Table of content

1 Summary	9
2 Scientific Works	11
2.1 The role of ERIC at LYON 2	11
2.2 The scientific position of ERIC	11
2.2.1 Historical point of view	12
2.2.2 Current scientific point of view	12
2.3 Presentation of KDD	14
2.4 Scientific progress and achievements 2004-2007	17
2.4.1 Contributions to complex data warehousing	17
2.4.2 Work on information retrieval from complex data warehouses	19
2.4.3 Works on data preparation	20
2.4.4 Works on Data Mining	21
2.4.5 Works on Validation-Integration and Deployment	22
2.5 Future directions	23
3 Scientific valuations	25
3.1 Publications	25
3.2 Editorial positions	25
3.3 Scientific animations	26
3.3.1 Conferences and workshops	26
3.3.2 Working group	26
3.3.3 Seminars	26
3.4 Applied research	27
3.5 Freeware development	27
3.6 Synergy between research and teaching	28
4 Ressources	31
4.1 Financial balance	31
4.2 Human Resources at December, 31st 2007	32
4.2.1 Permanent staff	32
4.2.2 Assistants	32
4.2.3 Theses in progress	33
4.2.4 Theses	33
4.2.5 Habilitated	34
4.2.6 Administration staff	34

4.2.7	Summary at 31st December 2007	34
4.2.8	People having completed their contracts or left the lab.....	34
5	Publications 2004-2007	37
5.1	International Journals	37
5.2	French journals.....	38
5.3	International Conferences.....	39
5.4	National Conferences	46
5.5	Chapters of book.....	52
5.6	PhD and HDR.....	53

APPENDICES

I.	Personal files of activities	57
II.	Editorial Activities	127
III.	Organisation of Scientific events.....	129
a.	Conferences, Workshops and working research groups	129
b.	Seminars of the Master ECD	130
c.	Seminars of the ERIC Lab.....	132
IV.	Applied Research Projects	135
V.	International collaborations	141

To our beloved and lamented Nicolas.

1 SUMMARY

The Research Team in Knowledge Engineering (Equipe de Recherche en Ingénierie des Connaissances, ERIC) was founded in 1995, initially as a “Young Team” (Jeune Equipe, JE) (1995-1999), later as a “Host Team” (Equipe d’Accueil, EA), since 1999¹. ERIC’s staff is currently composed of three full professors (CNU 27)², eight assistant professors and a part-time secretary. The laboratory hosts twenty PhD students, three assistants (Attachés Temporaires d’Enseignement et de Recherche, ATER) and, on average, three guest professors per year invited for around one month stay. The ERIC laboratory is situated at Porte des Alpes campus of Bron. It shares its premises with the Computer Science and Statistics Department (Département d’Informatique et Statistique DIS) from the Faculty of Economics and Management (Faculté de Sciences Economiques et de Gestion).

The work carried out within the laboratory focuses on Knowledge Discovery from Databases (KDD), tackling problems of both scientific and technological nature, such as:

- taking complex data into account: data that are heterogeneous (tables, multimedia, graphs, etc.), with little structure, of great volume, having or not a temporal nature, etc.;
- taking into account the empirical nature of data mining and its impact on quality measurement and its optimization in machine learning, identification of efficient representation spaces, the combination and aggregation of predictors, etc.;
- involving domain knowledge, either for semantic enhancement of data, or for decision-making systems deployment.

Putting these aspects into perspective in a KDD process makes it possible on the one hand to unify them and, on the other hand, to reveal new challenges and new approaches.

In quantitative terms, here is a summary of what has been achieved at ERIC in the last four academic years:

- 9 researchers have been able to carry out or complete their doctorate at ERIC;
- 4 colleagues have successfully completed the authorization to supervise research activities;
- ERIC continues to attract young PhD students with ministerial research grants and industry grants, with 20 doctoral theses currently in progress;
- 3 business start-up projects have been originated from the laboratory;

¹ "Young Team" and "Host Team" are statuses attributed by the state department for the research teams.

² "CNU 27" is the part of the National Council of Universities that brings together teaching researchers in the field of computer science

- 9 foreign colleagues have been invited as guest professors for a minimum stay of one month;
- 5 international teaching and scientific agreements;
- participation in national research projects, generating a total of 200 K€;
- partnerships with the private sector, generating 146 K€;
- the quality and diversity of ERIC's publications reflect the many achievements of a highly dynamic team: publications in international (19) or national (9) journals, papers published at international (90) or national (72) conferences, book chapters (15) or the supervision of works (HDR) (5), diffusion of software, organisation of major conferences, contacts with foreign universities, and so on. The expertise developed by the researchers at ERIC is recognized, as shown by the number of contracts signed, representing about 50% of its resources;
- ERIC maintains strong and explicit links with teaching activities, particularly via the Master's degree in Computer science taught at Lyon 2 and the preparation of an Erasmus Mundus Master's degree in partnership with 6 universities from 4 European countries (Italy, Spain, Romania and France).

For the future, we plan to:

- maintain as much as possible the synergy between teaching and research;
- reinforce the themes for which ERIC is recognized, that is, Knowledge Discovery from Databases (KDD);
- support research activities at three levels: theoretical, software development, and research, applied particularly in the fields of Human, Social and Economic Sciences;
- reinforce the team's editorial and scientific organizational policy by increasing the number of publications in specialized international journals;
- develop our relationships with the local, national and international scientific community around both research activities and teaching, such as co-directorship of doctoral thesis, double diplomas or the European Master's degree;
- promote our research at the industrial level by supporting pre-industrial projects.

2 SCIENTIFIC WORKS

2.1 The role of ERIC at LYON 2

It no longer needs to be proven that computer models, and computer science in general, have become not just a tool, but also a methodological framework for dealing with issues raised in different domains. For example, when psychologists aim to understand cognitive mechanisms, or when sociologists aim to analyze the behavior of a social group, or when geographers want to reconstruct reliefs with computer-generated graphics, or when archeologists want to identify and date remains from the past. This awareness has become virtually generalized both in research and in teaching, particularly in human and social sciences.

An illustration of the high level of involvement of colleagues belonging to computer science and mathematics in the specific domain of social and human science of our university is the creation of various teaching programs that are at the intersection between different disciplines. For instance, let's mention the interaction of Computing and Applied Mathematics with other disciplines such as, Human and Social Sciences (MISASHS Bachelor degree)¹, Economics and Management (IDEA Bachelor degree)². We also offer a Master degree in Computing³ based on three professional specializations⁴ and a combined professional and research specialization⁵.

2.2 The scientific position of ERIC

The scientific position of the research conducted at ERIC can be made on two points of view based on the historical evolution of the technology and the current scientific challenges, which we will present in a succinct manner.

¹ Mathematics, Computer Science and Statistics Applied to Human and Social Sciences

² Decision-based Computer Science and Applied Econometrics

³ The Master in Computer Science is common to Lyon 1, Lyon 2, the Ecole Normale Supérieure (ENS Lyon), the Ecole Centrale de Lyon (ECL) and the National Institute for Applied Sciences (INSA – Lyon).

⁴ Computer Engineering for Economics Decision Making and Evaluation (IIDEE); Social and Economic Statistics and Computing (SISE) and Organization and Business Information Systems Protection (OPSIE).

⁵ Knowledge Discovery in Databases (ECD).

2.2.1 Historical point of view

A study¹ of *Disk/Trend*, an American company based in California and specialized in industrial espionage, revealed that the average cost of storing a megabyte on a hard drive went from \$11.54 in 1988 to \$0.04 in 1998 and \$0.003 in 2007. This exponential decrease in the cost of storing has led to an exponential increase in the volume of data stored by companies. On average, according to the same studies, this volume doubles every 9 months. This trend for accumulating data does not seem to have reached its asymptomatic level and is even accelerated by the effect of the development of transmission networks such as the Internet, which are becoming increasingly powerful. Today, such networks can effectively attain a rate of 10 gigabytes per second, whilst their cost continues to decrease. Since 1975, the cost, in Mbit/s.km, has been divided by 1000. Now, it is possible to claim that access to data and their storage are based on reliable and cheap technology. In other words, the challenge of the 1960s-1970s, which was to master information systems, has been successfully addressed. So now, where do the new challenges stand? They have been moved to the field of information access and knowledge discovery in large databases. For further proof, one needs only to look at the greatest computer industry success of the last decade, Google and its internet search engine. It is thus possible to affirm that what is problematic today is neither the access to data nor their storage, but instead the semantic content of the data.

The activities at the ERIC laboratory lie under this vast and ever-expanding field of research. The aim is to propose and develop novel methodologies and IT tools that make it possible to gain access to the semantic content of the largest databases. This is what we refer to as Knowledge Discovery in Databases (KDD).

2.2.2 Current scientific point of view

Decision-making is at the heart of all activities, be they human, social, biological, economic, or other. It is based on identifying a situation and, depending on a desired state that we can consider to be the aim, undertaking the appropriate actions. Let us take three simple examples:

- A doctor who examines a patient and who observes an anomaly will prescribe a treatment whose aim is to eliminate the pathology so as to improve the state of health of the patient.
- A coastguard who, thanks to his radars, observes a suspicious movement out at sea, will launch a verification process in order to see whether it is a boat that needs to be intercepted.
- A judge, in view of the account of the facts recognized by the accused, will identify the type of crime and decide on the appropriate punishment.

¹ Gamze Zeytinci, CSIS-550 History of Computing Spring-2001; Evolution of the Major Computer Storage Devices From Early Mechanical Systems to Optical Storage Technology.

For multiple reasons such as cost, efficacy, rapidity, complexity, volume and so on, a significant part of the decision-making process, and particularly the identification process, is entrusted to computers. Let us consider the security services responsible for the Internet and whose aim is, for example, to intercept any communications judged to be “sensitive”. How can this type of task be contemplated, when we know that there are almost 700 million Internet users and that, on average, an Internet user produces around ten messages (e-mails, blogs, queries, etc.) per day? In this context, it is impossible to imagine any human organization capable of guaranteeing surveillance of the content of 7.5 billion messages every day, not to mention the web sites whose content can be targeted. It is also possible to observe this situation in other fields, such as marketing or health. For example, the generalization of breast cancer screening to all French women in the 50-74 years range is becoming a problem. In fact, this would require an X-ray infrastructure capable of treating almost 11 million examinations per year, for only 2,000 available radiologists in France and who are already at saturation point when they only treat around 60% of these cases. In such situations, resorting to the use of the power of computers seems a natural way of attaining scalability. But in order to be able to use computers, it is necessary to be able to “explain” to a computer how to recognize an electronic communication of a possibly suspicious nature out of all the millions of messages? And, in the health area for instance, how to identify and localize suspicious cases on the basis of a medical file containing images (mammograms), clinical and/or biological examinations, reports, and so on? Let us suppose that we would like to develop an information system capable of helping public safety managers identify “sensitive” messages. The designer of the identification assistance system will collect, from experts in content analysis, all the rules that lead to a diagnosis. These rules describe the content analysis process of a message and the deduction rules for ultimately deciding whether or not a message is sensitive. To achieve this, there are two conditions, both based on strong hypothesis:

- ❖ The first postulates that the identification process can be described as a series of operations on symbolic structures that we call inference rules and that we assimilate with the knowledge used by experts to identify the categories of a message.

- ❖ The second postulates that the expert is able to explicit his knowledge in the form of rules according to a precise formalism that could be coded into a machine.

If these conditions are both satisfied, then knowledge can be introduced into the machine in the form of rules in order to form a knowledge base. The machine is then given a programme capable of interpreting the rules in the same way as a compiler. This program is called an “inference engine” and it will apply the rules at its disposition with a view to make an inference from the facts that are submitted. If the expert system (knowledge base + inference engines) is judged to be relevant, it is then possible to duplicate it into a population of so-called “intelligent” agents that can be then deployed on the Internet

to identify and signal the presence of any suspect message. To the extent that such an approach aims to mimic the expert's reasoning to face real cases, this approach could be referred to as psycho-mimetic. Unfortunately, this approach comes up against two key problems that bring into doubt, at least in part, the hypothesis that we have just proposed. It can effectively occur that, in new fields, there are no experts, and thus no knowledge that can be transferred into a computer. What is the solution? Is it necessary to wait until the people confronted with the surveillance situations in question have acquired the expertise required by a trial-and-error process so that it can then be communicated to the machine? It is also possible for an expert to be perfectly able to identify the situations required, but totally incapable on the other hand of explaining the cognitive process he follows to make the identification. After all, we are all capable of recognizing someone in a crowd, but does that also mean that we are capable of explaining how we do it?

To overcome these deficiencies, we use the KDD methods. The knowledge is not provided by the expert, but produced by the machine after machine learning from past situations. For example, a doctor provides all the data concerning the patients with and without cancer that have already been treated and, thanks to KDD methods, we will try to determine the diagnosis rules that could be applied to new cases awaiting diagnosis. Once validated, this knowledge could, in turn, be integrated into the Expert system. This Expert system will also be able to incorporate fragments of knowledge obtained from human experts. This approach can be applied to all decision-making fields.

[The ERIC laboratory is working on the KDD process that we are now going to describe in a little more detail.](#)

2.3 Presentation of KDD

Without going into too much technical detail, it can be said that KDD is a process that makes use of the methods and the tools produced by a variety of computing fields: databases, artificial intelligence, statistics, optimization, etc. with a view to exploring voluminous and heterogeneous amounts of data looking for structured, unvarying elements that, once extracted and validated, could be considered to be knowledge.

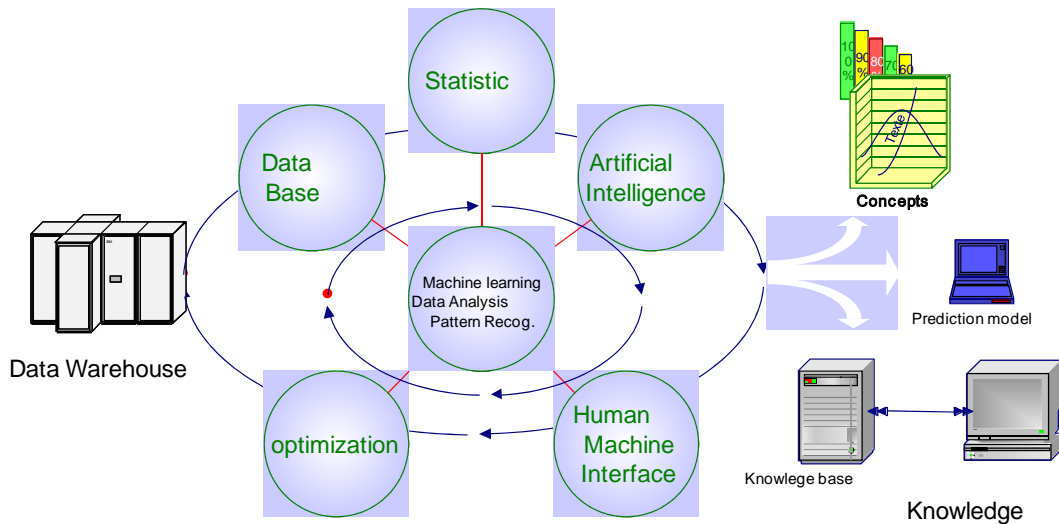


Figure 1 : KDD : From Data to Knowledge, Technologies involved

KDD is a 4-step process depicted in figure 2: Acquisition, Data Preparation, Data Mining and Validation that we will now describe very briefly:

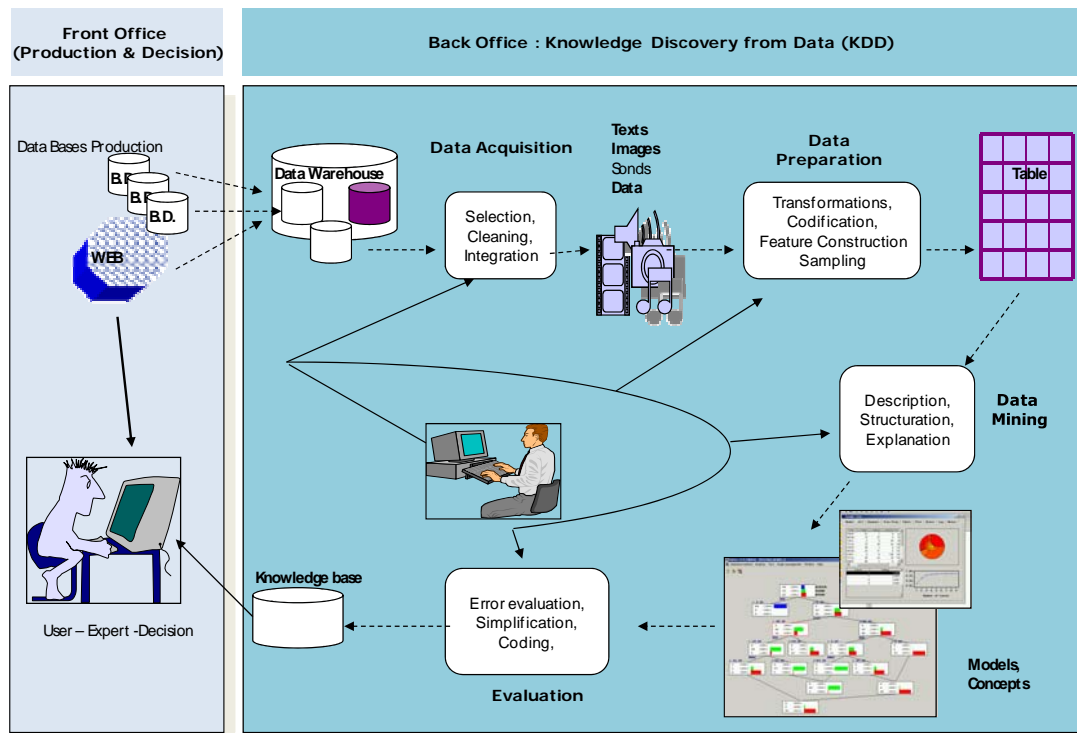


Figure 2: KDD workflow

Acquisition: the aim of acquisition is to retrieve from data warehouses the data liable to help performing a KDD task. Acquisition can be achieved from very large sources, those dispersed over a large geographical area and those saved on a wide range of computer environments (relational databases, XML databases, flat files or those in special formats such as DICOM or MPEG7). In

addition to access and selection, it is also necessary to organize and integrate the data into a local environment that is suited to data mining: for example, XML databases, OLAP warehouse, etc.

Data preparation: the data acquired can be of different types: tables of figures, textual data, images, etc. The aim of the preparation phase is to structure the data in such a way so as to make it possible to implement data mining methods. The form that is the most generally appropriate is that of a double entry data table. This can be a table containing individual observations-variables, a contingency table, a similarity table, etc. It should be specified that this operation is far from being simple and generally forms a real bottleneck. For example, if the source data are texts, it is necessary to define a series of linguistic pre-processes such as lemmatization, stop-word removal or stemming, resorting to ontology as a means of unifying the vocabulary, the extraction of concepts, etc. It is also necessary to identify the statistic individuals: are they texts, paragraphs, concepts etc. This work generally requires expertise and is associated with the field of application. The same problems arise when the data are in image form. What attributes must be extracted in order to describe these data with no apparent mathematical structure? What types of unit should be taken into account: whole images, mini-images produced by cutting up or automatic segmentation, etc. When the fundamental examples are composed of data of a variety of different types (images, text, curves, tables of figures, etc.), this is known as complex data. It is then necessary to not only define ad hoc coding that is pertinent for these data, but also to unify them. The medical file of a patient is the perfect example of this situation as it often contains X-rays, electrocardiogram printouts, quantitative data on biological readings, textual data describing a clinical examination, etc. As a result, it is necessary to be able to construct similarity measurements between individuals by taking into account all the available data. It is at this level that the question of incomplete and/or imprecise data is raised: how should they be taken into account in data mining? This phase is crucial. With the selection phase, it represents 80% of the time spent in a KDD cycle.

Data mining: this stage is at the heart of the KDD approach. We must make use of a type of algorithm, either to describe the lines and/or columns in the tables, or to structure the lines and/or columns of the table in clusters or, finally, to establish a prediction method by means of explanatory methods. We can mention, at random: factorial data analysis, clustering methods known as unsupervised learning techniques or association rules algorithms, etc. The nature of the data and coding then introduce variants of the algorithms which enrich the range of mining methods. For example, if the data are fuzzy or symbolic, it is appropriate to propose specific algorithms for each of the types of the mentioned methods.

Validation: At this stage, it is time to evaluate the results in relation to their reliability, their value for the user and, if applicable, to raise the problem of integrating them into a knowledge base. In this case,

it is then necessary to code them with a formalism that is appropriate so as to be able to use them in a real situation.

2.4 Scientific progress and achievements 2004-2007

When reading the personal activity files (Appendix I) of our researchers and looking through the list of publications for the period in question, we can observe that the work conducted at ERIC covers a relatively broad spectrum of subjects, covering the entire KDD cycle. Putting this work into perspective nevertheless, allows us to observe that particular attention has been paid to taking complex data into account in the KDD process. This specific context has made it possible to reveal new issues and challenges, both theoretical and technological, in order to respond to the needs of real applications. We will thus highlight this originality when presenting the contributions of the researchers from ERIC in the field of Knowledge Discovery from Complex Data (KDCCD). We will then, via the entire KDD cycle, look at which problems are covered, where the theoretical input is situated, how the methodological contributions and applications are processed. We will then find, in the appendices and for each researcher (including doctoral students) a summary that covers this work in a more specific manner.

2.4.1 Contributions to complex data warehousing

Standard data warehouses were developed from the model of relational databases. They in turn gave rise to technology such as OLAP (On Line Analytical Processing), making it possible to navigate through the immense wealth of knowledge they contain and extract summaries. What, then, would be the scientific and technical challenges laid down in the context of complex data? In addition to the problem of the representation of data with little or no structure, is it possible to imagine interrogation and exploration models that would be the equivalent of data in table form? What would then be the performances of this type of system once put into practice? In the following section we will describe the work carried out in order to try to answer these questions.

2.4.1.1 Representation and navigation in complex data warehouses

To tackle these scientific issues, we propose a complete process of warehousing and on-line analysis of complex data. The integration and modeling consist of physically incorporating complex data into a database acting as ODS (Operational Data Storage). In the integration phase, we define conceptual, logical and physical models. We use XML as a formalism to describe the logical and physical models. The conceptual model is translated at the logic level in the form of a DTD or XML schema. From the logic model obtained, we generate a collection of XML documents such as physical models. The XML

documents generated are valid and can be stored in an XML-native or relational database via mapping. Moreover, we have proposed an approach in order to build an OLAP cube described by an XML schema. This XML cube is automatically generated from user requirements expressed by a dimensional conceptual model (DCM) and a corpus of complex data represented by XML documents. The DCM and XML documents are both expressed using XML schemas (XSD) and then transformed into attribute trees in order to be compared. Some matching algorithms make it possible, thanks to operators of fusion by pruning or grafting, to treat the attribute trees in order to generate an XML schema of an XML cube and their corresponding XML documents. This XML cube provides an analysis context and may be analyzed by OLAP or data mining techniques. We have also developed another data integration approach based on a mediation system using ontology to describe each data source. Starting from these local ontology, the aim is to build a global ontology allowing the mediator to propose the relevant data for the construction of an OLAP cube. This global ontology is built using classification of the set of terms belonging to the local ontology.

As a result of this research, two main types of software have been produced. (1) SMAIDoC is a multi-agent system, articulated around five agents, for achieving the integration of complex data into a relational or XML-native database. (2) X-Warehousing is a Java platform dedicated to the automatic generation of XML cubes. These cubes are obtained by starting with user requirements expressed through a multidimensional conceptual model, which can be matched with the XML documents containing the complex data.

2.4.1.2 Optimization and evaluation of complex data warehouse performances

In this context of using XML language as a support for warehousing complex data, data warehouse performance remains as much as ever a crucial issue. The main physical data structures that are used for optimizing the data access time when performing complex analytical queries are indices, materialized views and partitions. Selecting an optimal set of such objects is an NP-hard problem that has been quite extensively addressed. However, scalability remains an issue, since existing approaches either rely on human expertise or exploit costly data structures. Furthermore, relationships between indices and materialized views are never taken into account, while these data structures are mutually beneficial.

To address these issues, we have designed a generic, automatic approach that applies data mining techniques to a workload (set of queries) that is representative of data warehouse usage, to deduce a quasi-optimal configuration of indices and/or materialized views. This approach significantly reduces the search space for suitable indices and views, and thus improves scalability. Then, cost models help select the most efficient indices and materialized views in terms of a performance gain/overhead ratio. These models take index-view interrelationships into account in order to achieve the best trade-off.

Our performance optimization research has been supported by, and applied to, two projects for optimizing access to complex data: MAP (relational, biomedical data warehouse) and CLAPI (XML warehouse of spoken language corpora).

Moreover, to assess and compare the efficiency of performance optimization techniques, they must be tested in different test cases. This task is usually carried out experimentally with the help of benchmarks. The Transaction Processing Performance Council (TPC) issues standard benchmarks, but their database schema and workload are fixed (only warehouse size varies). They are thus of little relevance in an engineering and design context.

Hence, to experimentally validate our performance optimization approach, we have designed several generic benchmarks. Their main design principle is adaptability: our benchmarks make it possible to generate various data warehouse configurations, as well as associated decision-support workloads. DWEB (Data Warehouse Engineering Benchmark) is the most mature of these tools. It is currently the only operational benchmark for evaluating the performances of data warehouses that is freely available online.

2.4.2 Work on information retrieval from complex data warehouses

Information retrieval from complex data warehouses (Recherche d'Information dans les Entrepôts de données Complexes, RIEC) raises questions of a specific nature. In all information retrieval processes it is effectively necessary to be equipped with either a similarity measure between objects, or a topological structure of these same objects without having to explicit the underlying measure of similarity. In cases where the data are in table form, there are many similarity indices. But, when the data are not structured, such as in chemical formulae, texts, images, temporal series, videos or multimedia documents, it is not easy to define the proximity between objects. The wide range of solutions proposed up till now take only one type of data into account at a time: either textual, or images, or the structure of chemical formulae, etc. Few works have focused on measurements that take into account the heterogeneity that is found in data.

We have explored different strategies for constructing a similarity measurement between complex objects. For example, we have done this by aggregating the standard similarity measurements from each type of data. Aggregation is then done by linear combination of the partial similarities that are the result of each type of data. We have also adopted other approaches, combining both a topological point of view and a probabilistic point of view. To do so, we constructed, for example, a topological structure in each homogenous subset of data, using standard similarity measurements. The global similarity between two objects will be all the greater if these objects are neighbors in the specific subspaces. For example, two patients will be more close (similar) if they are similar in the clinical data

field, similar in the image data field, similar in the biological data field, etc. Thus, two patients will be declared similar if they are similar in each of the topological subspaces induced. From this, we can construct an indicator which, once standardised, would be similar to the probability that the two patients be similar. If this probability is equal to 1, then the two individuals could be considered to be identical or almost identical. These approaches are still in progress, and have been efficiently implemented and tested in real applications, of which we can mention:

- The spoken language corpus, which contains textual, audio and visual data. This research project was carried out in the context of a joint project with the ICAR¹ laboratory (UMR Lyon 2 ENS lettres), and has benefited from financial support from the State Department for Higher Education.
- The corpus of legal texts associated with international workplace law. This research project was carried out in collaboration with the University of Geneva and the International Labor Organization.
- A basic corpus of images indexed by texts available as benchmarks in the Information Retrieval community.

2.4.3 Works on data preparation

In KDCCD, more than elsewhere, data preparation for mining is an arduous task. The main problem is the choice of representational space. Originally, the incidences of objects saved in databases are expressed with a formalism that does not lend itself, or very little, to the mathematical processing that is the foundation of most data mining methods. It is often necessary to use unified coding, generally in the form of a vector. How then can textual, image, video and temporal data be transformed into vectors? Is it necessary to put everything into the form of a vector and align everything in a single table? Such choices are not anodyne as they can, in turn, generate other problems. For example, the choice of coding texts with a digital vector requires linguistic processes that are often very sophisticated and where intuition, the *a priori* knowledge of the field and even arbitrary choices are commonly called upon. And all without being certain that one has made the right choice. Of the problems that can arise, for example, we can quote a resulting vector that is very large in size, which would have a significant impact on calculation times, but would also have an effect on the coherence of the interpretation of proximities between observation points. In addition to the case of textual and/or image data, we can also mention genomic data, where the variables space is generally much larger than that of individuals. How can we then reduce dimensionality with only minimum loss of information? By selection? Elimination? Projection? And, above all, how can we evaluate the relevance of the new representational

¹ Interaction Corpus Learning Representation

space? Furthermore, incoming data sometimes arrive either incomplete or with background noise. For example, the content of electronic mail covers several of these problems. If it is not possible to correct the anomalies, can we at least take this incompleteness, uncertainty and imprecision into account in our analyses?

We have done a considerable amount of work on real applications in the fields of marketing, use of legal texts, identification of plankton from images, identification of the structure of curves in flows of time series... where these questions were key issues.

The expertise acquired by the laboratory in this field is considerable and has produced particularly interesting results, both theoretical and methodological. For example, a non parametric statistical test for measuring the separability of clusters with a view to supervised learning, strategies for detecting atypical objects in multidimensional spaces, similarity measurements for texts making it possible to free oneself from explicit “vectorisation”, or the use of taxonomic approaches such as Kohonen cards as a means of reducing dimensionality.

2.4.4 Works on Data Mining

Usually, data mining is done using table structures prepared in the previous phase. This is generally the most visible part of the mining process as it is the stage where the knowledge is produced in the form of models: logical rules, algebra equation, probability model, topology structure, etc. To do this, we use learning methods, be they supervised or not, exploratory methods such as search algorithms for association rules, factorial analyses, or methods of modeling such as Bayesian networks etc. We are now going to describe some of the works carried out.

- The implementation of mining methods in the context of KDCD has revealed both theoretical and practical problems. Effectively, even if at the mining stage the data are structured in table form, their volume can be extremely large and, as a result, be a problem for calculation times. In this context, we have been obliged to think in two directions. The first starts with the premise that a file, even one that is very large in size, remains all the same a sample obtained from a larger population. As a result, if the same processing were carried out in an iterative manner on this same file enhanced, iteratively, with new cases, the resulting model would very likely be different in terms of its structure and error rate. Hence the idea of using this idea by working on small samples whose size we increase by random addition of cases until the variance in the error rate, for example, becomes almost null. We will thus use the information available in an efficient manner without having to support its heavy weight. The second direction aims to make better use of the computer technology available. This involves seeing how to make data mining methods migrate towards system structures and software that can support the heavy aspect. In this context, we have made strong pairings between database management systems

capable of working on views (tables) of almost limitless size and mining algorithms such as decision trees. We are thus able to benefit from the structure of the data, particularly bitmap indices, to implement many data mining algorithms in a more efficient and scalable manner. We are thus orienting ourselves towards the introduction of new data mining operators alongside standard SQL operators within data warehouse management systems which can, through this, lead to integrated software platforms.

- We are working on real data and have often been confronted with the under-representation of certain classes of interest. In this case, the implementation of supervised learning methods requires taking into account and control of the asymmetry of the classes. In line with this, we performed an almost exhaustive project on the measurements generally used in decision trees, the extraction of association rules, etc. studying both the theoretical properties of these measurements and their performances on benchmarks led to new measurements of generalized entropy. We have also proposed new axioms for these measurements, which, moreover, give more significant results on practical cases.

- At the end of the data mining process, users prefer to have at their disposition intelligible prediction models such as those obtained from decision trees or graphs, and which express the results in the form of logical rules. The use of these algorithms nevertheless comes up against problems when, for example, the variables are widely distributed, requiring groups of modalities, or when the variable to be predicted is of a particular type, such as for example a survival curve or a vector. We have proposed several extensions for the methods based on decision trees.

2.4.5 Works on Validation-Integration and Deployment

The models obtained from learning must be validated before being used as knowledge by the user or a decision system. Most learning methods propose evaluation procedures for the quality of the models and these procedures are generally based on error rates in resubstitution or on sample tests. Reducing error rates has led to new strategies for learning such as bagging, boosting or semi-supervised learning. The valuation procedures that have been conducted within the laboratory have shown, in a clear manner, the advantages of using these re-sampling techniques around learning algorithms and the extensions that we have added.

The knowledge produced automatically is often merely a fragment of knowledge for constructing real assistance systems for decision-making. In this context, and in order to increase the performance of such systems, we have developed and tested methodologies for integrating into a same knowledge base the knowledge obtained from the field and/or the expert. This integration is made by means of

ontology. These ontologies are used both as expansion tools for the fragments of knowledge from different sources and as a receptacle for knowledge.

2.5 Future directions

ERIC has developed methodological expertise which clearly shows that it is capable of mastering the entire KDD cycle in industrial application situations. The teams are able to both identify the scientific challenges so as to undertake the theoretical work necessary, and to identify the technological issues that need to be treated so as to produce the practical tools that the end user will be able to appropriate easily.

Our future work will continue in this direction in the aim of reinforcing both aspects. Effectively, the needs in complex data mining are still in their early days. Industrial demand will increase and the tools validated will be much sought-after in the coming years.

In addition to the projects and the theoretical questions already being dealt with, it seems of great strategic importance for us that we develop our research into KDCC along two complementary issues:

- The representation of complex data and the integration of knowledge. In addition to storage problems, it is also necessary to work on organization, integration and semantic enhancement by means of ontology which, themselves, can be enhanced by data mining processes.
- One of the trouble spots that reduces the efficacy of learning algorithms, be they supervised or not, is the quality of the representation space into which the objects to be processed are plunged. We can, without any risk of error, say that a machine would be capable of learning any task provided that the training examples are described with the “right” parameters. The works of Vapnik and Valiant, with proposals that are totally different, have opened the way for the concept of learnability, and our work on separability is akin to theirs in some ways. Learning strategies that focus simultaneously on reducing error via learning and variance suggest a direction for research that may lead to unification of the problem of learning and characterization of what should be, *a priori*, possible to learn or not depending on the data available.

Finally, the profiles of the two professorships to be filled during the next recruitment campaign will be in harmony with these perspectives.

3 SCIENTIFIC VALUATIONS

3.1 Publications

The table below summarizes the scientific quantitative balance for the period 2004-2007. A complete list of publications over the period 2004-2007 is given in section 5.

Publications	2004	2005	2006	2007	Total
International journals	6	4	4	5	19
National journals	4	2	2	1	9
International conferences	14	21	33	22	90
National conferences	17	16	20	19	72
Books	0	2	1	2	5
Chapters of books	1	2	3	9	15
Total	42	47	63	58	210
Permanent researchers	11	11	11	11	/
PhD Theses achieved	3	1	4	1	9
Qualifications to supervise PhD research (HDR) achieved	1	0	2	1	4
Number of students enrolled in research master Knowledge Discovery in Databases (ECD)	31	30	25	35	121

Table 1: Scientific balance for the period 2004-2007

3.2 Editorial positions

- Zighed D.A. (ERIC, Lyon 2) and Venturini G. (LI, Tours) are co-directors of the RNTI journal (Revue des Nouvelles Technologies de l'Information) published by Cépaduès. These publications (<http://www.antsearch.univ-tours.fr/rnti>) are mainly focused on the data mining and knowledge extraction fields. The whole RNTI listing since 2004 can be found in the appendix II.
- Darmont J. (ERIC, Lyon 2) is a member of the following editorial boards:
 - International Journal of Biomedical Engineering and Technology (IJBET, see <http://www.inderscience.com/browse/index.php?journalCODE=ijbet>)
 - Idea Groupe Inc. Editorial Advisory Review Board

- Editorial Review Board of the Advances in Data Warehousing and Mining (ADWM, see http://users.monash.edu.au/%7Edtaniar/book_series_warehousing.html)

- F. Bentayeb, O. Boussaid, J. Darmont and S. Loudcher are all members of the steering committee of the Conference “Entrepôts de Données et Analyse en ligne (EDA)”.

3.3 Scientific animations

The ERIC laboratory is involved in many scientific animations, including seminars, thematic work groups, conference and workshop organization. Among the majors are:

3.3.1 Conferences and workshops

- Conference “Extraction et Gestion des Connaissances” (EGC 2000-2008).
- Conference “Entrepôts de Données et Analyse en Ligne” (EDA 2005-2008).
- Workshop on Mining Complex Data in association with ICDM IEEE International Conference (2005-2006) and PKDD International conference (2007).
- Workshop “Qualité des Données et des Connaissances”, in association with the Conference “Extraction et Gestion des Connaissances” (2007-2008).
- Workshop “Fouille de Données Complexes”, in association with the Conference Extraction et Gestion des Connaissances (2004-2008).
- Workshop “Mesure de similarité sémantique” (SimSem 2008).
- Workshop “Systèmes Décisionnels” (ASD 2006-2008).

The details of these events are given in the appendix III.

3.3.2 Working group

The ERIC laboratory has created and currently leads the working group about complex data mining (Fouille de Données Complexes). More details can be found on the website:

<http://eric.univ-lyon2.fr/~gt-fdc/>

3.3.3 Seminars

The ERIC laboratory organizes regular seminars gathering professional and/or academic researchers on average twice a month (<http://eric.univ-lyon2.fr/index.php?section=6&soussection=14> and <http://dea-eed.univ-lyon2.fr/?page=seminaire§ion=0>).

These public seminars involve speakers of different horizons and have the following objectives:

- create a link between the laboratory members and other researchers, including some foreign researchers;

- bring together both the students and the laboratory members with Data Mining professionals;
- get a different perspective on issues falling within the laboratory research activity scope;
- better understanding of the subjects related to the concerns of the laboratory;
- allow the researchers, especially, the PhD students, to present their recent works;

These seminars are fully listed in the appendix III.

3.4 Applied research

The ERIC laboratory is also involved in the creation of start-up companies. It offers them the required scientific expertise through different kinds of collaborations. Companies are also supplied with an additional support from CREALYS, which aims to encourage the creation of innovative companies.

ERIC is currently involved in the creation of three companies operating in different market sectors:

- MAP (2003-2004): Archiving, structuring and querying medical data for computer-aided diagnosis and prescription;
- TradingBots (2006-2007): Developing tools dedicated to financial data analysis for prediction and decision support;
- TAPEO (2007-2008): Managing virtual portfolios of business shares owned by users communities on the Web.

3.5 Freeware development

TANAGRA is an open source data mining software intended for academic and research activities. The project began in January 2004. The software is freely available on the web (<http://eric.univ-lyon2.fr/~ricco/tanagra/>). TANAGRA implements various statistical and machine learning algorithms. There are about 130 methods at this time (January 2008).

The main goal of the project is to propose to the researchers and the teachers a tool respecting the standards of the data mining domain. The users use the software for academic studies, for their research activities, they use it also for their publications. TANAGRA is now recognized by the community. It is referred in the comparative surveys and the studies on real data (e.g. X. Chen, Y. Ye, G. Williams, X. Xu, “A Survey of Open Source Data Mining Systems”, Industrial Track Workshop, PAKDD-2007, 3-14.).

Last but not least, the project is also a Web site with many tutorials about data mining and exploratory data analysis, in French and in English. There are about 70 tutorials on line. The web pages

related to the tutorials are the most visited pages of our Web site. Over the 2007 year, we had on average nearly 4000 monthly visitors (#130 visitors per day).

3.6 Synergy between research and teaching

ERIC has always connected its teaching activities with its research work, more particularly within the Masters degrees. We have set up complete education courses able, at the same time, to find safe professional opportunities and populate the laboratory with researchers and PhD students. In terms of Bachelors, the two main channels are:

- Bachelor in Decision support systems and Applied Econometrics (IDEA) in the faculty of economics and management;
- Mathematics, Computer Science and Statistics Applied to Human Sciences (MISASH).

At the Masters level, we have set up a training that evolves around the Computer Decision support Systems and Statistics, which offers 4 possible specializations for the second year:

- Statistics, Computer science and Socio-Economics (SISE). Its content guides students in the fields related to the statistical data processing for marketing or industry (pharmacy, etc.). This specialty is also offered in the context of a joint Master with the University of Kharkov (Ukraine).
- Engineering Science for Decision support systems and Economic Evaluation (IIDEE) whose content is more focused on the development of tools for the decision support systems. This specialty receives each year a second promotion in evening classes open to professionals.
- Organizations and Business Information Systems Protection (OPSIE) whose program has three facets: Information technology, Management and Legal. This specialty receives also every year a second promotion in evening classes open to professionals.
- Knowledge Discovery from Databases (ECD) that is most focused on our research activities and the most part of our future PhD degrees. This specialty is co-empowered since its creation by école polytechnique of the University of Nantes and was also co-entitled until 2006 by university of Orsay Paris 11. The courses that are insured are given as video-conferencing, which allows some of our students to take lessons from their place of residence in France or abroad, as this is the case for the Romanian students in Bucharest, and Vietnamese students in Cantho.

We are continuing our opening efforts to build partnerships with other universities including European ones and attract good students. In this perspective, and to give a stronger impact, we will submit to the European Commission, a project to create a European master under the Erasmus Mundus program. It will be positioned in the field of data mining and knowledge management and will

be joined with 6 universities including 3 foreign universities (Italy, Spain and Romania). Other actions are also planned in the professional field with engineering schools or business schools or with foreign universities for delocalized training.

Details of these courses are available on the website of the university and particularly the Department of Computer Science and Statistics from the faculty of Economics and Management Science:
<http://dis.univ-lyon2.fr/>

4 RESSOURCES

4.1 Financial balance

The following two tables show the balance over the period 2004-2007.

Annual incomes excluding taxes	2004	2005	2006	2007	Total	Average
Amount allocated by the ministry (Overheads of the university withdrawn)	41 650 €	41 650 €	41 650 €	37 400 €	162 350 €	40 588 €
Other supports from university of Lyon 2		738 €	1 000 €	1 000 €	2 738 €	913 €
Own resources (contracts and benefits from companies)	73 738 €	13 593 €	33 784 €	25 510 €	146 625 €	36 656 €
Own resources from local and public institutions	16 600 €	4 250 €	4 500 €	33 850 €	59 200 €	14 800 €
Own resources from European institutions		20 898 €	30 602 €		51 500 €	25 750 €
National funds for the research	34 078 €	20 700 €	27 700 €		82 478 €	27 493 €
Funds for Research and technology					0 €	
Grand Total	166 066 €	101 829 €	139 236 €	97 760 €	504 891 €	126 223 €

Table 2: Annual incomes excluding taxes for 2004-2007

Annual Expenditures	2004	2005	2006	2007	Total	Average
Operating expenses	108 087 €	81 275 €	83 944 €	62 558 €	335 864 €	83 966 €
Equipement expenses	16 844 €	18 958 €	39 620 €	12 636 €	88 058 €	22 015 €
Staff expenses (only technical staff)	27 495 €	25 204 €	13 993 €	6 512 €	73 203 €	18 301 €
Grand Total	152 426 €	125 437 €	137 557 €	81 705 €	497 126 €	124 281 €

Table 3: Expenditures 2004-2007

4.2 Human Resources at December, 31st 2007

4.2.1 Permanent staff

Surname, Name, birthday date	Position	Discipline	Arrival dates at ERIC
Bentayeb Fadila, May 15th, 1966	Ass. Prof.	Comp. Sces	oct-01
Bousaïd Omar, 2 juin 1954	Ass. Prof.	Comp. Sces	jan-95
Chauchat Jean-Hugues, 6 juillet 1946	Full Prof.	Comp. Sces	jun-97
Darmont Jérôme, 15 janvier 1972	Ass. Prof.	Comp. Sces	oct-99
Harbi Nouria, 27 août 1961	Ass. Prof.	Comp. Sces	oct-05
Lallich Stéphane, 20 septembre 1947	Full Prof.	Comp. Sces	jun-97
Loudcher Rabaséda Sabine, 27 octobre 1969	Ass. Prof.	Comp. Sces	oct-98
Rakotomalala Ricco, 19 juillet 1967	Ass. Prof.	Comp. Sces	oct-98
Velcin Julien, 09 mars 1978	Ass. Prof.	Comp. Sces	nov-07
Viallaneix Jacques, 6 juillet 1963	Ass. Prof.	Comp. Sces	jan-95
Zighed Abdelkader, 12 mars 1955	Full Prof.	Comp. Sces	jan-95

Assistant Professor : MCF ; Computer science discipline =CNU 27 ;

4.2.2 Assistants

Surname - Name	Years
Arigon Anne-Muriel	2006-2008
Favre Cécile	2007-2008
Lefort Virginie	2006-2008
Mahboubi Hadj	2007-2008
Maïz Nora	2007-2008

Assistant = Attaché Temporaire d'Enseignement et de Recherche (ATER)

4.2.3 Theses in progress

Surname and Name	Start	Supervisor	Co-supervisor	Financing
Bahri Emma	2006	S. Lallich		Grant MENRT
Bodin-Niemczuk Anouck	2007	O. Boussaid	S. Loudcher	Grant MENRT
Bouatour Sonia	2007	(Co-supervisor ; Univ. Tunis, Tunisia)	O. Boussaid	Self-Financing
El Sayed Ahmad	2004	D. Zighed		Self-Financing
Gaudin Rémi	2004	D. Zighed		Grant MENRT
Hacid Hakim (done 2008-feb-04)	2004	D. Zighed		Grant from local Gov.
Hachicha Marouane	2007	J. Darmont		Grant MENRT
Julien Charbel	2004	D. Zighed	L. Saitta (Co Supervisor, Univ. Alessandria, Italy))	Self-Financing
Maïz Nora	2005	O. Boussaid	F. Bentayeb	Assistant
Marcellin Simon	2004	D. Zighed		Industrial Grant (CIFRE)
Mavrikas Efthimios	2002	S. Dascalopoulos (Egee, Greece)	D. Zighed	Grant (Greece gov.)
Prudhomme Elie	2005	S. Lallich		Grant MENRT
Qureshi Taimur	2006	D. Zighed		Grant (Pakistan' gov.)
Ralaivao Jean-Christian	2003	J. Darmont	V. Manantsoa, U. of Fianarantsao, Madagascar	Grant from French embassy at Madagascar
Rakotoarivelo Ony	2006	J. Darmont	F. Bentayeb	Grant MENRT
Salem Rashed Kh.	2007	O. Boussaid	J. Darmont	Grant (Egypt gov.)
Stavrianou Anna	2005	JH. Chauchat		Grant MENRT
Thomas Julien	2005	D. Zighed		Industrial Grant (CIFRE)
Wei Zhihua	2006	JH. Chauchat		Grant (Chinese gov.)

MNERT : French Ministry of Education and Research

4.2.4 Theses

Surname, Name	Year	Directors	Co directors	Present situation
Aouiche Kamel	2005	D. Zighed	J. Darmont	Post Doc in Canada
Baume Laurent	2004	N. Nicoloyannis	C. Mirodatos	Post Doc in Spain
Ben Messaoud Riadh	2006	N. Nicoloyannis	O. Boussaid S. Loudcher	Assistant in Tunisia
Clech Jérémy	2004	D. Zighed		Working in private company
Clerc Frédéric	2006	N. Nicoloyannis	R. Rakotomalala	Working in private company
Erray Walid	2006	D. Zighed		Working in private company
Fangseu Badjio Edwige	2006	D. Zighed	F. Poulet	Working in private company
Favre Cécile	2007	O. Boussaid	F. Bentayeb	Assistant at Lyon 2
Legrand Gaëlle	2004	N. Nicoloyannis		Working in private company

4.2.5 Habilitated

Surname, Name	Year	Director	Present Situation
Lenca Philippe	2007	D. Zighed	Assist. Prof
Boussaid Omar	2006	D. Zighed	Assist. Prof
Darmont Jérôme	2006	D. Zighed	Assist. Prof
Poulet François	2004	D. Zighed	Assist. Prof

4.2.6 Administration staff

Surname, Name	Category	Percentage	Arrival date
Gabriele Valérie	Secretary (IATOS)	50	Sept-00
Crevel Julien	Technical staff	50	Sept-07

4.2.7 Summary at 31st December 2007

Category	Number
Professors	11
Assistants	5
Theses in progress	20
Theses	9
Habilitated	4
administration	2

4.2.8 People having completed their contracts or left the lab

4.2.8.1 Professors

Surname, Name	Position	Discipline	Date of arrival	Date of depart.
Viallefont Anne	Assist. Prof	Applied math.	oct-00	sept-06

4.2.8.2 Assistants

Surnam, Name	Year
Ben Messaoud Riadh	2006-2007
Clech Jérémy	2003-2004
Effantin Dit Toussaint Brice	2004-2005
Kouomou Choupo Anicet	2005-2006
Légrand Gaëlle	2004-2005
Muhlenbach Fabrice	2002-2003
Scuturici Marian	2003-2004
Scuturici Michaela	2002-2004
Suchier Maxime	2006-2007
Tweed Tiffany	2002-2004
Walid Erray	2003-2005

4.2.8.3 Post Doc at ERIC's Lab

Surname, Name	Year
Jouve Pierre	2004-2005

4.2.8.4 Administration staff

Surname, name	Financing	Percentage	Date of arrival	Date of departure
Delhomme Lydie	Own funds	100%	oct-02	August-04

5 PUBLICATIONS 2004-2007

In the references to publications, the 1st letter designates the type of publication (eg A for international journal...), the following letters correspond to the original authors and figures to the year of publication.

5.1 International Journals

[ADBB07] J. Darmont, F. Bentayeb, O. Boussaïd, "Benchmarking Data Warehouses", *International Journal of Business Intelligence and Data Mining*, Vol. 2, No. 1, 2007, 79-104.

[ABDFU07] F. Bentayeb, J. Darmont, C. Favre, C. Udréa, "Efficient On-Line Mining of Large Databases", *International Journal of Business Information Systems*, Vol. 2, No. 3, 2007, 328-350.

[ABTBD07] O. Boussaïd, A. Tanasescu, F. Bentayeb, J. Darmont, "Integration and Dimensional Modelling Approaches for Complex Data Warehousing", *Journal of Global Optimization*, Vol. 37, No. 4, April 2007, 571-591.

[ASAN07] A. Stavrianou, P. Andritsos, N. Nicoloyannis, "Overview and Semantic Issues of Text Mining", *SIGMOD Record*, Vol. 36, No. 3, September 2007, 23-34.

[ALLV07] S. Lallich, P. Lenca, B. Vaillant, "Probabilistic framework towards the parametrisation of association rule interestingness measures", *Methodology and Computing in Applied Probability*, Vol. 9, No. 3, 2007, 447-463.

[ACFRNM06] F. Clerc, D. Farrusseng, R. Rakotomalala, N. Nicoloyannis, C. Mirodatos, "Meta Modeling for Combinatorial Catalyst Optimization", *International Journal of Computer Science and Network Security*, Vol. 6, No. 10, 2006, 256-262.

[ARM06] R. Rakotomalala, F. Mhamdi, "Supervised and Unsupervised Feature Reduction for Protein Classification", *WSEAS Transactions on Information Science and Applications*, Vol. 3, No. 12, 2006, 2448-2455.

[AMRE06] F. Mhamdi, R. Rakotomalala, M. Elloumi, "A Compromise Between N-gram Length and Classifier Characteristics for Protein Classification", *International Journal of Computer Science and Network Security*, Vol. 6, No. 4, 2006, 82-87.

[ABBL06] R. BenMessaoud, O. Boussaïd, S. Loudcher-Rabaseda, "A Data Mining-Based OLAP Aggregation of Complex Data: Application on XML Documents", *International Journal of Data Warehousing and Mining*, Vol. 2, No. 4, Oct.-Dec. 2006, 1-26.

[AHD05] Z. He, J. Darmont, "Evaluating the Dynamic Behavior of Database Applications", *Journal of Database Management*, Vol. 16, No. 2, April-June 2005, 21-45.

[AZRES05] D. Zighed, G. Ritschard, W. Erray, V. Scuturici, "Decision tree with optimal join partitioning", *Journal of Intelligent Information Systems*, Vol. 20, 2005, 1-26.

- [AZLM05] D. Zighed, S. Lallich, F. Muhlenbach, "A statistical approach of class separability", *Applied Stochastic Models in Business and Industry*, Vol. 21, No. 2, 2005, 187-197.
- [ASCSZ05] M. Scuturici, J. Clech, V. Scuturici, D. Zighed, "Topological representation model for image databases query", *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 17, No. 1-2, 2005, 145-160.
- [AGVCCM04] O. Gimenez, A. Viallefont, E. Catchpole, R. Choquet, B. Morgan, "Methods for investigating parameter redundancy", *Animal Biodiversity and Conservation*, Vol. 27, No. 1, 2004, 561-572.
- [ABFLM04] L. Baumes, D. Farrusseng, M. Lengliz, C. Mirodatos, "Using Artificial Neural Networks for boosting discovery in High", *QSAR & Combinatorial Science*, 2004.
- [AKFBMS04] C. Klanner, D. Farrusseng, L. Baumes, C. Mirodatos, F. Schüth, "The Development of Descriptors for Solids: Teaching "Catalytic", *Angewandte Chemie International Edition*, Vol. 43, No. 40, 2004, 5347-5349.
- [APCFWMM04] S. Pereira, F. Clerc, D. Farrusseng, J. Waal, T. Maschmeyer, C. Mirodatos, "Effect of the Genetic Algorithm parameters on the optimisation of heterogeneous catalysts", *QSAR & Combinatorial Science*, September 2004.
- [AGVLF04] J. Gaillard, A. Viallefont, A. Loison, M. Festa-Bianchet, "Assessing senescence patterns in populations of large mammals", *Animal Biodiversity and Conservation*, Vol. 27, No. 1, 2004, 47-58.
- [AMLZ04] F. Muhlenbach, S. Lallich, D. Zighed, "Identifying and Handling Mislabeled Instances", *Journal of Intelligent Information Systems*, Vol. 22, No. 1, January 2004, 89-109.

5.2 French journals

- [BR07] R. Rakotomalala, "Data Mining : Spécificités et outils", *Actes de Chimométrie*, Novembre 2007, 108 - 110 (Lyon).
- [BRRMJ06] M. Raimbault, R. Rakotomalala, X. Morandi, P. Jannin, "Mise en évidence d'invariants dans une population de cas chirurgicaux", *Revue des Nouvelles Technologies de l'Information*, Vol. E-5, 2006, 339-348.
- [BLBFCRR06] J. Labarère, J. Bosson, D. Farrusseng, B. Crémilleux, R. Rakotomalala, C. Robert, "Arbres d'Induction : méthodes et exemple d'application", *Journal d'Economie Médicale*, Vol. 24, No. 2, 2006, 115-129.
- [BADBB05] K. Aouiche, J. Darmont, O. Boussaïd, F. Bentayeb, "Auto-administration des entrepôts de données complexes", *Revue des Nouvelles Technologies de l'Information*, Vol. E-4, Septembre 2005, 47-70.
- [BR05] R. Rakotomalala, "TANAGRA, une plate-forme d'expérimentation pour la fouille de données", *MODULAD*, No. 32, 2005, 70-85.
- [BHD04] Z. He, J. Darmont, "Une plate-forme dynamique pour l'évaluation des performances des bases de données à objets", *Ingénierie des Systèmes d'Information (RSTI série ISI)*, Vol. 9, No. 1, 2004, 109-127.

[BLMVPL04] P. Lenca, P. Meyer, B. Vaillant, P. Picouet, S. Lallich, "Evaluation et analyse multicritère des mesures de qualité des règles d'association", *Revue des Nouvelles Technologies de l'Information*, No. 2, 2004, 219-246.

[BLN04] G. Legrand, N. Nicoloyannis, "Sélection de variables et agrégation d'opinions", *Revue des Nouvelles Technologies de l'Information*, Vol. C1, 2004, 89-101 (ISBN 2.85428.667.7.).

[BLT04] S. Lallich, O. Teytaud, "Evaluation et validation de l'intérêt des règles d'association", *Revue des Nouvelles Technologies de l'Information*, No. 2, 2004, 193-217.

5.3 International Conferences

[CCMR07] J. Chauchat, A. Morin, R. Rakotomalala, "Correcting the error rate estimation bias in Data Mining when the dataset comes from a two-stage sampling", *Statistics for Data Mining, Learning and Knowledge Extraction (IAST 07)*, Aveiro, Portugal, August 2007.

[CEHZ07] A. ElSayed, H. Hacid, D. Zighed, "Mining semantic distance between corpus terms", *1st Ph.D. Workshop, 16th ACM Conference on Information and Knowledge Management (PIKM-CIKM 07)*, Lisbon, Portugal, November 2007, 49-54; ACM.

[CRD07] J. Ralaivao, J. Darmont, "Knowledge and Metadata Integration for Warehousing Complex Data", *6th International Conference on Information Systems Technology and its Applications (ISTA 07)*, Kharkiv, Ukraine, May 2007; *Lecture Notes in Informatics (LNI)*, Vol. P-107, GI-Edition, Bonn, Germany, 164-175.

[CTJN07] J. Thomas, P. Jouve, N. Nicoloyannis, "Asymmetric measure for supervised learning models assessment, application to breast cancer detection", *International Conference on Industrial Engineering and Systems Management (IESM 07)*, Beijing, China, May 2007.

[CEHZ07b] A. ElSayed, H. Hacid, D. Zighed, "A Multisource Context-Dependent Semantic Distance Between Concepts", *18th International Conference on Database and Expert Systems Applications (DEXA 07)*, Regensburg, Germany, September 2007; *LNCS*, Vol. 4653, Springer, Heidelberg, Germany, 54-63.

[CEHZ07c] A. ElSayed, H. Hacid, D. Zighed, "Using Semantic Distance in a Content-based Heterogeneous Information Retrieval System", *3rd International Workshop on mining complex data (MCD 07)*, Warsaw, Poland, 2007; *LNAI*, Springer, Heidelberg, Germany.

[CALLV07] J. Azé, P. Lenca, S. Lallich, B. Vaillant, "A Study of the Robustness of Association Rules", *2007 International Conference on Data Mining (DMIN 07)*, Las Vegas, USA, 2007, 163-169; CSREA Press.

[CEHZ07d] A. ElSayed, H. Hacid, D. Zighed, "A New Approach Towards Content-based Heterogeneous Information Retrieval", *ECML/PKDD Workshop on Mining Complex Data*, 2007.

[CHMD07] M. Hachicha, H. Mahboubi, J. Darmont, "Vers une algèbre XML-OLAP : État de l'art", *2ème Atelier Systèmes Décisionnels (ASD 07)*, Sousse, Tunisie, Octobre 2007.

[CSCBM07] A. Silic, J. Chauchat, B. Basic, A. Morin, "N-Grams and Morphological Normalization in Text Classification: A Comparison on a Croatian-English Parallel Corpus", *13th Portuguese Conference on Artificial Intelligence (EPIA 2007)*, Guimaraes, Portugal, December

2007; *LNCS*, Vol. 4874, Springer, Heidelberg, Germany, 671-682.

[CEHZ07e] A. ElSayed, H. Hacid, D. Zighed, "A New Context-Aware Measure for Semantic Distance Using a Taxonomy and a Text Corpus", *IEEE International Conference on Information Reuse and Integration (IRI 07)*, Las Vegas, USA, August 2007, 279-284; IEEE Systems, Man, and Cybernetics Society.

[CBMGNC07] L. Baumes, M. Moliner, R. Gaudin, N. Nicoloyannis, A. Corma, "Genetic Algorithms in Materials Science and Engineering", *2007 E-MRS Fall Meeting*, Warsaw, Poland, September 2007.

[CLLV07] S. Lallich, P. Lenca, B. Vaillant, "Construction of an off-centered entropy for supervised learning", *XIIth International Symposium on Applied Stochastic Models and Data Analysis (AMSDA 07)*, Chania, Crete, Greece, 2007.

[CEHZ07f] A. ElSayed, H. Hacid, D. Zighed, "Combining Text and Image for Content-Based Information Retrieval", *2007 International Conference on Information and Knowledge Engineering (IKE 07)*, 2007; CSREA Press.

[CARGPSG07] E. Antajan, R. Rakotomalala, S. Gasparini, M. Picheral, L. Stemmann, G. Gorsky, "Automatic quantification and recognition of major zooplankton groups in a North Sea time series using the Zooscan imaging system", *4th International Zooplankton Production Symposium*, 2007, 189 - 190 (Hiroshima, Japan).

[CDDFWGBP07] N. Durand, S. Derivaux, G. Forestier, C. Wemmert, P. Gançarski, O. Boussaïd, A. Puissant, "Ontology-based Object Recognition for Remote Sensing Image Interpretation", *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 07)*, Patras, Greece, October 2007.

[CBNM07] E. Bahri, N. Nicoloyannis, M. Maddouri, "Improving boosting by exploiting former assumptions", *3rd International Workshop on mining complex data (MCD 07)*, Warsaw, Poland, 2007.

[CAAD07] S. Azefack, K. Aouiche, J. Darmon, "Dynamic index selection in data warehouses", *4th International Conference on Innovations in Information Technology (Innovations 07)*, Dubai, United Arab Emirates, November 2007; IEEE.

[CFBB07] C. Favre, F. Bentayeb, O. Boussaïd, "Dimension Hierarchy Updates in Data Warehouses: a User-driven Approach", *9th International Conference on Enterprise Information Systems (ICEIS 07)*, Funchal, Madeira, Portugal, June 2007, 206 - 211.

[CEHZ07g] A. ElSayed, H. Hacid, D. Zighed, "A Context-Dependent Semantic Distance Measure", *19th International Conference on Software Engineering and Knowledge Engineering (SEKE 07)*, Boston, USA, July 2007, 432-437; Knowledge Systems Institute Graduate School.

[CMBB07] N. Maiz, F. Bentayeb, O. Boussaïd, "Ontology based mediation system", *18th Information Resource Management Association International Conference (IRMA 07)*, Vancouver, Canada, May 2007; IRMA, Hershey, USA.

[CFBB07b] C. Favre, F. Bentayeb, O. Boussaïd, "Evolution of data warehouses' optimization: a workload perspective", *9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2007)*, 2007; *LNCS*, Vol. 4654, Springer, Heidelberg, Germany, 13 - 22.

- [CRCP06] R. Rakotomalala, J. Chauchat, F. Pellegrino, "Accuracy Estimation with Clustered Dataset", *The Australasian Data Mining Conference (AusDM 06)*, Sidney, Australia, November 2006; *Conferences in Research and Practice in Information Technology*, Vol. 61.
- [CBBL06] R. BenMessaoud, O. Boussaïd, S. Loudcher-Rabaseda, "Efficient Multidimensional Data Representation Based on Multiple Correspondence Analysis", *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 06)*, Philadelphia, USA, August 2006.
- [CLVL06] P. Lenca, B. Vaillant, S. Lallich, "On the Robustness of Association Rules", *IEEE International Conferences on Cybernetics and Intelligent Systems and Robotics, Automation and Mechatronics (CIS-RAM 06)*, Bangkok, Thailand, June 2006, 596-601.
- [CMRE06] F. Mhamdi, R. Rakotomalala, M. Elloumi, "A Hierarchical N-Grams Extraction Approach for Classification Problem", *IEEE International Conference on Signal-Image Technology and Internet-Based Systems (SITIS 06)*, Tunisia, 2006, 310-321.
- [CFBB06] C. Favre, F. Bentayeb, O. Boussaïd, "A Knowledge-driven Data Warehouse Model for Analysis Evolution", *13th ISPE International Conference on Concurrent Engineering: Research and Applications (CE 06)*, Antibes, France, September 2006; *Frontiers in Artificial Intelligence and Applications*, Vol. 143, IOS Press, 271-278.
- [CDO06] J. Darmont, E. Olivier, "A complex data warehouse for personalized, anticipative medicine", *17th Information Resources Management Association International Conference (IRMA 06)*, Washington, USA, May 2006, 685-687; Idea Group Publishing.
- [CRZ06] G. Ritschard, D. Zighed, "Implication Strength of Classification Rules", *Foundations of Intelligent Systems (ISMIS 06)*, Bari, Italy, September 2006; *LNAI*, Vol. 4203, Springer, Heidelberg, Germany, 463-472.
- [CGN06] R. Gaudin, N. Nicoloyannis, "An Adaptable Time Warping Distance for Time Series Learning", *5th International Conference on Machine Learning and Applications (ICMLA 06)*, Orlando, USA, December 2006.
- [CTGLP06] O. Teytaud, S. Gelly, S. Lallich, E. Prudhomme, "Quasi-random bootstrap, with applications to rule extraction and (sub)bagging", *International Workshop on Intelligent Information Access (IIIA 06)*, Helsinki, Finland, July 2006.
- [CMBB06] N. Maiz, O. Boussaïd, F. Bentayeb, "Ontology-Based Mediation System", *13th ISPE International Conference on Concurrent Engineering: Research and Applications (CE 06)*, Antibes, France, September 2006; *Frontiers in Artificial Intelligence and Applications*, Vol. 143, IOS Press, 181-189.
- [CMR06] A. Morineau, R. Rakotomalala, "The TVpercent Criteria to Eliminate Uninformative Models among Association Rules", *Knowledge Extraction and Modeling IASC-INTERFACE-IFCS Workshop (KNEMO 06)*, Anacapri, Italy, 2006.
- [CBBL06b] R. BenMessaoud, O. Boussaïd, S. Loudcher-Rabaseda, "Using a Factorial Approach for Efficient Representation of Relevant OLAP Facts", *Seventh International Baltic Conference on Databases and Information Systems (DB&IS 06)*, Vilnius, Lithuania, July 2006.
- [CMD06] H. Mahboubi, J. Darmont, "Benchmarking XML data warehouses", *Atelier Systèmes*

Décisionnels (ASD 06), 9th Maghrebien Conference on Information Technologies (MCSEAI 06), Agadir, Maroc, December 2006.

[CC06] J. Chauchat, "Microeconomics Forecast: Learning by Doing, A Ten Years Graduate Level Experience", *7th International Conference On Teaching Statistics (ICOTS7), Salvador, Bahia, Brazil, July 2006.*

[CAJD06] K. Aouiche, P. Jouve, J. Darmont, "Clustering-Based Materialized View Selection in Data Warehouses", *10th East-European Conference on Advances in Databases and Information Systems (ADBIS 06), Thessaloniki, Greece, September 2006; LNCS, Vol. 4152, Springer, Heidelberg, Germany, 81-95.*

[CMAD06] H. Mahboubi, K. Aouiche, J. Darmont, "Materialized View Selection by Query Clustering in XML Data Warehouses", *4th International Multiconference on Computer Science and Information Technology (CSIT 06), Amman, Jordan, April 2006, 68-77.*

[CGBNB06] R. Gaudin, S. Barbier, N. Nicoloyannis, M. Banens, "Clustering of Bi-Dimensional and Heterogeneous Time Series: Application to Social Sciences Data", *2006 International Conference on Data Mining (DMIN 06), Las Vegas, USA, June 2006, 10-16.*

[CRM06] R. Rakotomalala, F. Mhamdi, "Combining feature selection and feature reduction for protein classification", *2nd WSEAS International Symposium on Data Mining, Lisbon, Portugal, 2006.*

[CMZR06] S. Marcellin, D. Zighed, G. Ritschard, "Detection of breast cancer using an asymmetric entropy measure", *Computational Statistics (COMPSTAT 06), 2006; Computational Statistics, Vol. XXV, Springer, Heidelberg, Germany, 975-982 (On CD).*

[CBBL06c] R. BenMessaoud, O. Boussaïd, S. Loudcher-Rabaseda, "Mining Association Rules in OLAP Cubes", *International Conference on Innovations in Information Technology (ITT 06), Dubai, November 2006.*

[CLTP06] S. Lallich, O. Teytaud, E. Prudhomme, "Statistical inference and data mining: false discoveries control", *17th COMPSTAT Symposium of the IASC, Rome, Italy, August 2006, 325-336.*

[CMBB06b] N. Maiz, F. Bentayeb, O. Boussaïd, "Un système de médiation basé sur les ontologies pour l'entreposage des données", *Atelier Systèmes Décisionnels (ASD 06), 9th Maghrebien Conference on Information Technologies (MCSEAI 06), Agadir, Maroc, Décembre 2006.*

[CZMR06] D. Zighed, S. Marcellin, G. Ritschard, "An asymmetric entropy measure for decision trees", *Knowledge Extraction and Modeling Workshop (KNEMO 06), Capri, Italy, September 2006.*

[CBBCA06] O. Boussaïd, R. BenMessaoud, R. Choquet, S. Anthoard, "X-Warehousing: an XML-Based Approach for Warehousing Complex Data", *10th East-European Conference on Advances in Databases and Information Systems (ADBIS 06), Thessaloniki, Greece, September 2006; LNCS, Vol. 4152, Springer, Heidelberg, Germany, 39-54.*

[CMZR06b] S. Marcellin, D. Zighed, G. Ritschard, "An asymmetric entropy measure for decision trees", *11th Information Processing and Management of Uncertainty in knowledge-based systems (IPMU 06), Paris, France, July 2006, 1292-1299.*

- [CRM06b] R. Rakotomalala, F. Mhamdi, "Improved Singular Value Decomposition for Supervised Learning in a High Dimensional Dataset", *6th International Workshop on Pattern Recognition in Information Systems (PRIS 06)*, Paphos, Cyprus, May 2006, 38-47.
- [CZH06] D. Zighed, H. Hacid, "Proximity graphs and separability of classes", *11th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 06)*, Paris, France, July 2006, 1488-1495; IPMU.
- [CFBB06b] C. Favre, F. Bentayeb, O. Boussaïd, "WEDriK : une plateforme pour des analyses personnalisées dans les entrepôts de données évolutifs", *Atelier Systèmes Décisionnels (ASD 06)*, *9th Maghrebian Conference on Information Technologies (MCSEAI 06)*, Agadir, Maroc, Décembre 2006.
- [CHZ06] H. Hacid, D. Zighed, "Content-Based Image Retrieval in Large Image Databases", *IEEE International Conference on Granular Computing (GrC 2006)*, Atlanta, USA, May 2006.
- [CVLL06] B. Vaillant, S. Lallich, P. Lenca, "Modelling of the counter-examples and association rules interestingness measures behavior", *2nd International Conference on Data Mining (DMIN 06)*, Las Vegas, USA, June 2006, 132-137.
- [CTJN06] J. Thomas, P. Jouve, N. Nicoloyannis, "Optimisation and evaluation of random forests for imbalanced datasets", *16th International Symposium on Methodologies for Intelligent Systems (ISMIS 06)*, Bari, Italy, September 2006; *LNAI*, Vol. 4203, Springer, Heidelberg, Germany, 642-651.
- [CHZ06b] H. Hacid, D. Zighed, "Content-Based Image Retrieval Using Topological Models", *12th International MultiMedia Modelling Conference (MMM 06)*, Beijing, China, 2006.
- [CBLBM06] R. BenMessaoud, S. Loudcher-Rabaseda, O. Boussaïd, R. Missaoui, "Enhanced Mining of Association Rules from Data Cubes", *9th ACM International Workshop on Data Warehousing and OLAP (DOLAP 06)*, Arlington, USA, November 2006.
- [CPHZ05] V. Pisetta, H. Hacid, D. Zighed, "Automatic Juridical Texts Classification and Relevance Feedback", *First IEEE International Workshop on Mining Complex Data (IEE MCD05)*, Texas, USA, 2005.
- [CHZ05] H. Hacid, D. Zighed, "An Effective Method for Locally Neighborhood Graphs Updating", *16th International Conference on Database and expert Systems Applications (DEXA 05)*, 2005; *LNCS*, Vol. 3588, Springer, Heidelberg, Germany, 930-939.
- [CLN05] G. Legrand, N. Nicoloyannis, "A new feature selection method", *8th International Conference on Pattern Recognition and Information Processing (PRIP05)*, Minsk Belarus, 2005.
- [CPL05] E. Prudhomme, S. Lallich, "Quality measure based on Kohonen maps for supervised learning of large high dimensional data", *International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005)*, Brest, France, 2005, 246-255.
- [CRME05] R. Rakotomalala, F. Mhamdi, M. Elloumi, "Hybrid Feature Ranking for Protein Classification", *1st International Conference on Advanced Data Mining and Applications (ADMA'05)*, 2005; *LNAI*, Vol. 3584, Springer, Heidelberg, Germany, 610-617.
- [CBBL05] R. BenMessaoud, O. Boussaïd, S. Loudcher-Rabaseda, "Evaluation of a MCA-Based

Approach to Organize Data Cubes", *ACM Fourteenth Conference on Information and Knowledge Management (CIKM 05)*, Bremen, Germany, 2005.

[CFB05] C. Favre, F. Bentayeb, "Bitmap index-based decision trees", *15th International Symposium on Methodologies for Intelligent Systems (ISMIS 05)*, New York, USA, May 2005; *LNAI*, Vol. 3488, Springer, Heidelberg, Germany, 65-73.

[CADBB05] K. Aouiche, J. Darmont, O. Boussaïd, F. Bentayeb, "Automatic Selection of Bitmap Join Indexes in Data Warehouses", *7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 05)*, Copenhagen, Denmark, August 2005; *LNCS*, Vol. 3589, Springer, Heidelberg, Germany, 64-73.

[CMKC05] A. Morin, A. Kouomou-Choupo, J. Chauchat, "Dimension reduction and clustering for query-by-example in huge image databases", *3rd IASC World Conference on Computational Statistics and Data Analysis*, Limassol, Cyprus, October 2005.

[CLN05b] G. Legrand, N. Nicoloyannis, "Feature selection and preferences aggregation", *International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005)*, Brest, France, 2005, 305-312.

[CMRE05] F. Mhamdi, R. Rakotomalala, M. Elloumi, "Feature Ranking for Protein Classification", *4th International Conference on Computer Recognition Systems (CORES'05)*, 2005; *Advances in Soft Computing*, Springer, Heidelberg, Germany, 611-617.

[CTBB05] A. Tanasescu, O. Boussaïd, F. Bentayeb, "Preparing Complex Data for Warehousing", *3rd ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 05)*, Cairo, Egypt, January 2005.

[CDBRA05] J. Darmont, O. Boussaïd, J. Ralaivao, K. Aouiche, "An Architecture Framework for Complex Data Warehouses", *7th International Conference on Enterprise Information Systems (ICEIS 05)*, Miami, USA, May 2005, 370-373.

[CDBB05] J. Darmont, F. Bentayeb, O. Boussaïd, "DWEB: A Data Warehouse Engineering Benchmark", *7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 05)*, Copenhagen, Denmark, August 2005; *LNCS*, Vol. 3589, Springer, Heidelberg, Germany, 85-94.

[CLN05c] G. Legrand, N. Nicoloyannis, "Feature selection method using preferences aggregation", *International Conference on Machine Learning and Data Mining (MLDM 05)*, Leipzig Germany, 2005; *LNCS*, Vol. 3587, Springer, Heidelberg, Germany, 9-11.

[CLVL05] S. Lallich, B. Vaillant, P. Lenca, "Parametrised measures for the evaluation of association rule interestingness", *International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005)*, Brest, France, 2005, 220-229.

[CCRF05] F. Clerc, R. Rakotomalala, D. Farrusseng, "Learning Fitness Function in a Combinatorial Optimization Process", *International Symposium on Applied Stochastic Models and Data Analysis*, 2005, 535-543.

[CHZ05b] H. Hacid, D. Zighed, "An Incremental Algorithm for Neighborhood Graphs Construction", *3rd World Conference on Computational Statistics & Data Analysis*, Cyprus, October 2005.

- [CMKC05b] A. Morin, A. Kouomou-Choupo, J. Chauchat, "Dimension reduction and clustering for query-by-example in huge image databases", *3rd World Conference on Computational Statistics and Data Analysis (CSDA 05)*, Cyprus, November 2005.
- [CHZ05c] H. Hacid, D. Zighed, "Neighborhood Graphs for Image Databases Indexing and Content-Based Retrieval", *First IEEE International Workshop on Mining Complex Data (IIE MCD05)*, Texas, USA, 2005.
- [CCPC05] J. Chauchat, M. Pacaut-Troncin, A. Cuerq, "Model Assessment and Selection : a Case Study on Risk Factors for Acute Suicidality in Psychiatric Patients", *Applied Statistics, Ribno (Bled), Slovenia*, 2005.
- [CMNK04] E. Mavrikas, N. Nicoloyannis, E. Kavakli, "Cultural Heritage Information on the Semantic Web", *14th International Conference on Knowledge Engineering and Knowledge Management (EKAW 04)*, Northamptonshire, UK, October 2004; *LNAI*, Vol. 3257, Springer, Heidelberg, Germany, 477-478.
- [CTBB04] A. Tanasescu, O. Boussaïd, F. Bentayeb, "Towards Complex Data Warehousing: A new approach for integrating and modeling Complex data", *5th International Conference on Modelling, Computation and Optimization in Information Systems and Management Sciences (MCO 04)*, Metz, France, July 2004, 619-626.
- [CBBR04] R. BenMessaoud, O. Boussaïd, S. Rabaseda, "A New OLAP Aggregation Based on the AHC Technique", *ACM 7th International Workshop on Data Warehousing and OLAP (DOLAP 04)*, Washington DC, USA, November 2004, 65-72.
- [CJCR04] R. Jalam, J. Clech, R. Rakotomalala, "Un cadre pour la catégorisation de textes multilingues", *7èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT 04)*, Louvain-la-Neuve, Belgique, 2004, 650-660 (A paraître).
- [CBDU04] F. Bentayeb, J. Darmont, C. Udréa, "Efficient Integration of Data Mining Techniques in Database Management Systems", *8th International Database Engineering and Applications Symposium (IDEAS 04)*, Coimbra, Portugal, July 2004, 59-67.
- [CMKN04] E. Mavrikas, E. Kavakli, N. Nicoloyannis, "Ontology-based Narrations from Cultural Heritage Texts", *5th International Symposium on Virtual Reality Archaeology and Cultural Heritage (VAST 2004)*, Ename, Belgium, December 2004 (Submitted for review).
- [CHC04] V. Hopirtean, J. Chauchat, "Knowledge Discovery on Clinical Trials to Explore the Overall Safety of the Medical Products - A case study", *International Workshop on Intelligent Data Analysis and Data Mining, Application in Medecine (SRCE)*, Zagreb, Croatia, June 2004.
- [CMER04] F. Mhamdi, M. Elloumi, R. Rakotomalala, "Textmining, feature selection and datamining for proteins classification", *2nd International Conference on Information and Communication Technologies (ICICT 04)*, Cairo, Egypt, 2004, 457-458.
- [CBRBB04] R. BenMessaoud, S. Rabaseda, O. Boussaïd, F. Bentayeb, "OpAC: A New OLAP Operator Based on a Data Mining Method", *Sixth International Baltic Conference on Databases and Information Systems (DB&IS 04)*, Riga, Latvia, June 2004.
- [CVPPKMB04] N. Vernicos, G. Pavlogeorgatos, D. Papadopoulos, E. Kavakli, E. Mavrikas, S. Bakogianni, "FCS_WORD Project : Wiki-based Ongoing Research Data Management", *32nd*

International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA 2004), Prato, Italy, April 2004.

[CMLZ04] F. Muhlenbach, S. Lallich, D. Zighed, "Outlier Handling in the Neighbourhood-Based Learning of a Continuous Class", *7th International Conference Discovery Science, Padova, Italy*, October 2004; *LNAI*, Vol. 3245, Springer, Heidelberg, Germany, 314-321.

[CMER04b] F. Mhamdi, M. Elloumi, R. Rakotomalala, "Descriptors Extraction for proteins classification", *3rd Conference on Neuro-Computing and Evolving Intelligence (NCEI 04)*, Auckland, New Zealand, December 2004.

[CJCD04] R. Jalam, J. Chauchat, J. Dumais, "Automatic Recognition of Keywords using N-grams", *16th Symposium of IASC (COMPSTAT 04)*, Prague, Czech Republic, August 2004, 1245-1254.

[CVLL04] B. Vaillant, P. Lenca, S. Lallich, "A clustering of interestingness measures", *7th International Conference Discovery Science, Padova, Italy*, October 2004; *LNAI*, Vol. 3245, Springer, Heidelberg, Germany, 290-297.

5.4 National Conferences

[DFBB07] C. Favre, F. Bentayeb, O. Boussaïd, "Intégration des connaissances utilisateurs pour des analyses personnalisées dans les entrepôts de données évolutifs", *7èmes Journées Francophones Extraction et Gestion des Connaissances (EGC 07)*, Namur, Belgique, Janvier 2007; *Revue des Nouvelles Technologies de l'Information*, Cépaduès, Toulouse, 217 - 222.

[DPRZ07] V. Pisetta, G. Ritschard, D. Zighed, "Choix des conclusions et validation des règles issues d'arbres de classification", *7ème Conférence Extraction et Gestion des Connaissances (EGC 07)*, Namur, Belgique, 2007; *Revue des Nouvelles Technologies de l'Information*, Vol. E-9, Cépaduès, Toulouse, 485-496.

[DSRBGM07] M. Studer, G. Ritschard, L. Baccaro, I. Georgiou, N. Muller, "Relations entre types de violation des libertés syndicales garanties par les conventions de l'OIT : Une analyse statistique implicative des résultats d'une fouille de texte", *Nouveaux apports théoriques à l'analyse statistique implicative et applications*, 2007, 111-122; Département de Mathématiques, Université Jaume I.

[DVMMMLL07] B. Vaillant, S. Menou, S. Moga, P. Lenca, S. Lallich, "Qualité des règles d'association : étude de données d'entreprise", *3ème Atelier Qualité des Connaissances à partir des Données (QDC-EGC 07)*, Namur, Belgique, Janvier 2007, 55-64.

[DZPR07] D. Zighed, V. Pisetta, D. Ratsimba, "Separability of Classes in a multidimensional Space", *Classification and Data Analysis*, September 2007; *Meeting of the classification and Data Analysis Group of the Italian Statistical Society*, Eum edizioni università di macerata, 147-150.

[DTJN07] J. Thomas, P. Jouve, N. Nicoloyannis, "Mesure non symétrique pour l'évaluation de modèles, utilisation pour les jeux de modèles, utilisation pour les jeux de données déséquilibrés", *7ème Conférence Extraction et Gestion des Connaissances (EGC 07)*, Namur, Belgique, Janvier 2007; *Revue des Nouvelles Technologies de l'Information*, Vol. E-9, Cépaduès, Toulouse, 509-519.

[DRB07] O. Rakotoarivelo, F. Bentayeb, "Evolution de schéma par classification automatique pour les entrepôts de données", *4ème atelier Fouille de Données Complexes dans un Processus*

d'Extraction des Connaissances (FDC-EGC 07), Namur, Belgique, Janvier 2007.

[DMBB07] N. Maiz, O. Boussaïd, F. Bentayeb, "Clustering method for semi-automatically ontologies fusion", *4ème atelier Fouille de Données Complexes dans un Processus d'Extraction des Connaissances (FDC-EGC 07), Namur, Belgique, Janvier 2007.*

[DMD07] H. Mahboubi, J. Darmont, "Fragmentation des entrepôts de données XML", *3èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 07), Poitiers, Juin 2007; Revue des Nouvelles Technologies de l'Information, Vol. B-3, Cépaduès, Toulouse, 177-190.*

[DRZM07] G. Ritschard, D. Zighed, S. Marcellin, "Données déséquilibrées, entropie décentrée et indice d'implication", *Nouveaux apports théoriques à l'analyse statistique implicative et applications, 2007, 315-327; Departament de Matemàtiques, Universitat Jaume I; ASI4.*

[DBNM07] E. Bahri, N. Nicoloyannis, M. Maddouri, "Amélioration du Boosting par combinaison des hypothèses antérieures", *14èmes Rencontres de la Société Francophone de Classification (SFC 07), Paris, Septembre 2007.*

[DFBB07b] C. Favre, F. Bentayeb, O. Boussaïd, "Evolution de modèle dans les entrepôts de données : existant et perspectives", *3èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 07), Poitiers, Juin 2007; Revue des Nouvelles Technologies de l'Information, Vol. B-3, Cépaduès, Toulouse, 21-36.*

[DZMR07] D. Zighed, S. Marcellin, G. Ritschard, "Mesure d'entropie asymétrique et consistante", *7ème Conférence Extraction et Gestion des Connaissances (EGC 07), Namur, Belgique, 2007; Revue des Nouvelles Technologies de l'Information, Vol. E-9, Cépaduès, Toulouse, 81-86.*

[DPL07] E. Prudhomme, S. Lallich, "Ensemble prédicteur fondé sur les cartes auto-organisatrices adapté aux données volumineuses", *7ème Conférence Extraction et Gestion des Connaissances (EGC 07), Namur, Belgique, Janvier 2007; Revue des Nouvelles Technologies de l'Information, 473-484.*

[DRB07b] O. Rakotoarivelo, F. Bentayeb, "Evolution de schéma par classification automatique pour les entrepôts de données", *3èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 07), Poitiers, Juin 2007; Revue des Nouvelles Technologies de l'Information, Vol. B-3, Cépaduès, Toulouse, 99-112.*

[DEHZ07] A. ElSayed, H. Hacid, D. Zighed, "Recherche d'Information par le Contenu des Données Hétérogènes", *RIAS, 2007; IRIT, Université de Toulouse.*

[DFBB07c] C. Favre, F. Bentayeb, O. Boussaïd, "Evolution et personnalisation des analyses dans les entrepôts de données : une approche orientée utilisateur", *XXVème congrès Informatique des organisations et systèmes d'information et de décision (INFORSID 07), Perros-Guirec, Mai 2007, 308 - 323.*

[DLLV07] S. Lallich, P. Lenca, B. Vaillant, "Construction d'une entropie décentrée pour l'apprentissage supervisé", *3ème Atelier Qualité des Connaissances à partir des Données (QDC-EGC 07), Namur, Belgique, Janvier 2007, 45-54.*

[DBBGP07] R. Brisson, O. Boussaïd, P. Gançarski, A. Puissant, N. Durand, "Navigation et appariement d'objets géographiques dans une ontologie", *7ème Conférence Extraction et Gestion des Connaissances (EGC 07), Namur, Belgique, Janvier 2007; Revue des Nouvelles Technologies*

de l'Information, Cépaduès, Toulouse.

[DFBB06] C. Favre, F. Bentayeb, O. Boussaïd, "Evolution de schémas dans les entrepôts de données : modèle à base de règles", *2ème journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA 06)*, Versailles, Juin 2006; *Revue des Nouvelles Technologies de l'Information*, Vol. B-2, Cépaduès, Toulouse, 175-176.

[DH06] H. Hacid, "Annotation semi-automatique de grandes BD images : Approche par graphes de voisinage", *CONFérence en Recherche d'Informations et Applications (CORIA 06)*, Lyon, Mars 2006.

[DMBB06] N. Maiz, O. Boussaïd, F. Bentayeb, "Un système de médiation basé sur les ontologies", *3ème atelier Fouille de Données Complexes dans un processus d'extraction des connaissances, EGC 06*, Lille, Janvier 2006, 27-38.

[DBJTC06] A. Brémond, P. Jouve, J. Thomas, J. Clech, "Résultats Préliminaires d'une étude comparative de deux CAD", *Innovations Technologiques et Bonnes Pratiques en Sénologie : Dépistage - Diagnostic - Traitement*, Juin 2006, 92-94; Fusium; Sofmis.

[DMER06] F. Mhamdi, M. Elloumi, R. Rakotomalala, "Extraction et Sélection des n-grammes pour le Classement de Protéines", *Atelier Extraction et gestion de connaissances appliquées aux données biologiques (Bio-EGC 06)*, Lille, Janvier 2006, 25-37.

[DBBCA06] O. Boussaïd, R. BenMessaoud, R. Choquet, S. Anthoard, "Conception et construction d'entrepôts XML", *2ème journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA 06)*, Versailles, Juin 2006; *Revue des Nouvelles Technologies de l'Information*, Vol. B-2, Cépaduès, Toulouse, 3-22.

[DMAD06] H. Mahboubi, K. Aouiche, J. Darmont, "Un index de jointure pour les entrepôts de données XML", *6èmes Journées Francophones Extraction et Gestion des Connaissances (EGC 06)*, Lille, Janvier 2006; *Revue des Nouvelles Technologies de l'Information*, Vol. E-6, Cépaduès, Toulouse, 89-94.

[DMR06] A. Morineau, R. Rakotomalala, "Critère VT-100 de sélection des règles d'association", *6èmes Journées Francophones Extraction et Gestion des Connaissances (EGC 06)*, Lille, Janvier 2006; *Revue des Nouvelles Technologies de l'Information*, Vol. E-6, Cépaduès, Toulouse, 581-592.

[DPHZ06] V. Pisetta, H. Hacid, D. Zighed, "Multi-catégorisation de textes juridiques et retour de pertinence", *6èmes Journées Francophones Extraction et Gestion des Connaissances (EGC 06)*, Lille, Janvier 2006; *Revue des Nouvelles Technologies de l'Information*, Vol. E-6, Cépaduès, Toulouse, 235-246.

[DZ06] D. Zighed, "Aspects conceptuels : Différentes Méthodologies des Systèmes Experts pour la Détection ou la Caractérisation", *Innovations Technologiques et Bonnes pratiques en Sénologie : Dépistage - Diagnostic - Traitement*, 2006; Sofmis, Fusium, 76-81.

[DBL06] O. Boussaïd, S. Loudcher-Rabaseda, "Intégration des méta-données dans la fouille de données", *XXIVème Congrès Informatique des organisations et systèmes d'information et de décision (INFORSID 06)*, Hammamet, Tunisie, 2006.

[DFBB06b] C. Favre, F. Bentayeb, O. Boussaïd, "Modèle d'entrepôt de données à base de règles", *3ème atelier Fouille de Données Complexes dans un processus d'extraction des connaissances*,

EGC 06, Lille, 2006, 39-50.

[DFBB06c] C. Favre, F. Bentayeb, O. Boussaïd, "A Rule-based Data Warehouse Model", *23rd British National Conference on Databases (BNCOD 2006)*, Belfast, Northern Ireland, July 2006; LNCS, Vol. 4042, Springer, Heidelberg, Germany, 274-277.

[DGN06] R. Gaudin, N. Nicoloyannis, "Séries temporelles : Vers une mesure de distance optimale", *Fouille de données temporelles, 6èmes Journées d'Extraction et de Gestion des Connaissances (EGC 06)*, Lille, Janvier 2006, 67-75.

[DHZ06] H. Hacid, D. Zighed, "Graphes de Proximité pour l'Indexation et l'Interrogation d'Images par le Contenu", *6èmes Journées Francophones Extraction et Gestion des Connaissances (EGC 06)*, Lille, Janvier 2006; *Revue des Nouvelles Technologies de l'Information*, Vol. E-6, Cépaduès, Toulouse, 11-22.

[DMAD06b] N. Maiz, K. Aouiche, J. Darmont, "Sélection automatique d'index et de vues matérialisées dans les entrepôts de données", *2ème journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA 06)*, Versailles, Juin 2006; *Revue des Nouvelles Technologies de l'Information*, Vol. B-2, Cépaduès, Toulouse, 89-104.

[DPHBR06] V. Pisetta, H. Hacid, F. Bellal, G. Ritschard, "Traitement automatique de textes juridiques", *Semaine de la Connaissance (SdC 06)*, Nantes, Juin 2006 (CDrom).

[DBINZ06] A. Brémond, A. Isnard, N. Nicoloyannis, D. Zighed, "Numérisation secondaire et Lecture sur Ecran : Evaluation des Performances", *Innovations Technologiques et Bonnes Pratiques en Sénologie : Dépistage - Diagnostic - Traitement*, 2006, 22-28; Fusium.

[DRM06] R. Rakotomalala, F. Mhamdi, "Evaluation des Méthodes Supervisées pour le Classement de Protéines", *13èmes Rencontres de la Société Française de Classification (SFC 06)*, Metz, Septembre 2006, 181-184.

[DZI06] D. Zighed, A. Isnard, "Projet NORDOM (Numérisation, Optimisation, Rationalisation du Dépistage Organisé en Mammographie", *Innovations Technologiques et Bonnes Pratiques en Sénologie : Dépistage - Diagnostic - Traitement*, 2006, 29-34; Fusium; Sofmis.

[DFBBN05] C. Favre, F. Bentayeb, O. Boussaïd, N. Nicoloyannis, "Entreposage Virtuel de demandes marketing : de l'acquisition des objets complexes à la capitalisation des connaissances", *2ème atelier Fouille de Données Complexes dans un processus d'extraction des connaissances, EGC 05*, Paris, Janvier 2005, 65-68.

[DBLB05] G. Brunet, S. Lallich, A. Bideau, "Analyse quantitative des réseaux généalogiques ascendants, l'exemple des lignées familiales de la vallée de la Valserine (Jura français)", *XXVe Congrès international de la Population (UIESP)*, Tours, Juillet 2005.

[DPL05] E. Prudhomme, S. Lallich, "Validation statistique des cartes de Kohonen en apprentissage supervisé", *5èmes Journées d'Extraction et de Gestion des Connaissances (EGC 05)*, Paris, Janvier 2005; *Revue des Nouvelles Technologies de l'Information*, Cépaduès, Toulouse, 79-90.

[DGN05] R. Gaudin, N. Nicoloyannis, "Apprentissage non supervisé de séries temporelles à l'aide des k-Means et d'une nouvelle méthode d'agrégation de séries", *5èmes Journées d'Extraction et de Gestion des Connaissances (EGC 05)*, Paris, Janvier 2005; *Revue des Nouvelles Technologies de l'Information*, Cépaduès, Toulouse, 201-212.

- [DFB05] C. Favre, F. Bentayeb, "Intégration efficace des arbres de décision dans les SGBD : utilisation des index bitmap", *5èmes Journées d'Extraction et de Gestion des Connaissances (EGC 05)*, Paris, Janvier 2005; *Revue des Nouvelles Technologies de l'Information*, Cépaduès, Toulouse, 319-330.
- [DLLV05] S. Lallich, P. Lenca, B. Vaillant, "Variations autour de l'intensité d'implication", *Colloque Analyse Statistique Implicative (ASI 2005)*, Palerme, Sicile, Octobre 2005, 237-246.
- [DR05] R. Rakotomalala, "TANAGRA : un logiciel gratuit pour l'enseignement et la recherche", *5èmes Journées d'Extraction et de Gestion des Connaissances (EGC 05)*, Paris, Janvier 2005; *Revue des Nouvelles Technologies de l'Information*, Cépaduès, Toulouse, 697-702.
- [DUB05] C. Udréa, F. Bentayeb, "Fouille de données relationnelles dans les SGBD", *5èmes Journées d'Extraction et de Gestion des Connaissances (EGC 05)*, Paris, Janvier 2005; *Revue des Nouvelles Technologies de l'Information*, Cépaduès, Toulouse, 356.
- [DBRB05] R. BenMessaoud, S. Rabaseda, O. Boussaïd, "L'analyse factorielle pour la construction de cubes de données complexes", *2ème atelier Fouille de Données Complexes dans un processus d'extraction des connaissances*, EGC 05, Paris, Janvier 2005, 53-56.
- [DLN05] G. Legrand, N. Nicoloyannis, "Etat de l'art des méthodes de construction de variables", *12èmes Rencontres de la Société Francophone de Classification (SFC 05)*, Montréal, 2005, 182-185.
- [DRRMJ05] M. Raimbault, R. Rakotomalala, X. Morandi, P. Jannin, "Mise en évidence d'invariants dans une population de cas chirurgicaux", *2ème atelier Fouille de Données Complexes dans un processus d'extraction des connaissances*, EGC 05, Paris, Janvier 2005, 149-158.
- [DR05b] J. Ralaivao, "Améliorer la performance d'un entrepôt de données complexes par l'utilisation de métadonnées et de connaissances du domaine", *2ème atelier Fouille de Données Complexes dans un processus d'extraction des connaissances*, EGC 05, Paris, Janvier 2005, 81-84.
- [DJN05] P. Jouve, N. Nicoloyannis, "Forage distribué des données : une comparaison entre l'agrégation d'échantillons et l'agrégation de règles", *5èmes Journées d'Extraction et de Gestion des Connaissances (EGC 05)*, Paris, Janvier 2005; *Revue des Nouvelles Technologies de l'Information*, Cépaduès, Toulouse, 31-42.
- [DLN05b] G. Legrand, N. Nicoloyannis, "Gestion de la phase de prétraitement des données et coefficient Kappa", *XXXVIIèmes Journées de Statistique*, Pau, 2005, 6-10; SFdS.
- [DBAF05] R. BenMessaoud, K. Aouiche, C. Favre, "Une approche de construction d'espaces de représentation multidimensionnels dédiés à la visualisation", *1ère journée sur les Entrepôts de Données et l'Analyse en ligne (EDA 05)*, Lyon, Juin 2005; *Revue des Nouvelles Technologies de l'Information*, Vol. B-1, Cépaduès, Toulouse, 34-50.
- [DVMPLLB05] B. Vaillant, P. Meyer, E. Prudhomme, S. Lallich, P. Lenca, S. Bigaret, "Mesurer l'intérêt des règles d'association", *Atelier Qualité des Données et des Connaissances (DQK 05)*, EGC 05, Paris, Janvier 2005, 69-78.
- [DJLN04] P. Jouve, G. Legrand, N. Nicoloyannis, "Sélection rapide en apprentissage supervisé", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04)*, Clermont-Ferrand, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Vol. 2, Cépaduès,

Toulouse, 185-196.

[DLN04] G. Legrand, N. Nicoloyannis, "Sélection de variables et agrégation d'opinions", *11èmes Rencontres de la Société Francophone de Classification (SFC 04)*, Bordeaux, 2004.

[DVLL04] B. Vaillant, P. Lenca, S. Lallich, "Etude expérimentale de mesures de qualités de règles d'association", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04)*, Clermont-Ferrand, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Vol. 2, Cépaduès, Toulouse, 341-352.

[DUBDB04] C. Udréa, F. Bentayeb, J. Darmont, O. Boussaïd, "Intégration efficace de méthodes de fouille de données dans les SGBD", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04)*, Clermont-Ferrand, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Vol. 2, Cépaduès, Toulouse, 83-94.

[DLN04b] G. Legrand, N. Nicoloyannis, "Construction de variables et arbres de décision", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04)*, Clermont-Ferrand, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Vol. 2, Cépaduès, Toulouse, 204 (Poster).

[DLN04c] G. Legrand, N. Nicoloyannis, "Sélection de variables et agrégation d'opinions", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04)*, Clermont-Ferrand, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Cépaduès, Toulouse.

[DLM04] S. Lallich, F. Muhlenbach, "Apprentissage à partir de voisinages et fouilles d'images", *Workshop Analyse de données, Statistique et Apprentissage pour la Fouille d'Images, 14e Conference Francophone AFRIF AFIA*, Janvier 2004, 23-28.

[DSSCZ04] M. Scuturici, V. Scuturici, J. Clech, D. Zighed, "Navigation dans une base d'images à l'aide de graphes topologiques", *XXIIème Congrès Informatique des organisations et systèmes d'information et de décision (INFORSID 04)*, Biarritz, Mai 2004.

[DSCSZ04] M. Scuturici, J. Clech, V. Scuturici, D. Zighed, "Modèle topologique pour l'interrogation des bases d'images", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04)*, Clermont-Ferrand, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Vol. 2, Cépaduès, Toulouse, 409-414.

[DE04] W. Erray, "WF : Une méthode de sélection de variables combinant une méthode filtre rapide et une approche enveloppe", *11èmes Rencontres de la Société Francophone de Classification (SFC 04)*, Bordeaux, Septembre 2004.

[DLPT04] S. Lallich, E. Prudhomme, O. Teytaud, "Contrôle du risque multiple en sélection de règles d'association significatives", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04)*, Clermont-Ferrand, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Vol. 2, Cépaduès, Toulouse, 305-316.

[DJC04] R. Jalam, J. Chauchat, "Catégorisation de textes multilingues: quelques solutions", *Atelier Fouille de Textes, EGC 04*, Clermont-Ferrand, Janvier 2004, 27-36.

[DDBLB04] A. Duffoux, O. Boussaïd, S. Lallich, F. Bentayeb, "Fouille dans la structure de documents XML", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04)*, Clermont-Ferrand, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*,

Vol. 2, Cépaduès, Toulouse, 519-524.

[DDBB04] J. Darmont, F. Bentayeb, O. Boussaïd, "Conception d'un banc d'essais décisionnel", *20èmes Journées Bases de Données Avancées (BDA 04)*, Montpellier, Octobre 2004, 493-511.

[DADB04] K. Aouiche, J. Darmont, O. Boussaïd, "Sélection automatique d'index dans les entrepôts de données", *1er atelier Fouille de Données Complexes dans un processus d'extraction des connaissances, EGC 04*, Clermont-Ferrand, Janvier 2004, 91-102.

[DLN04d] G. Legrand, N. Nicoloyannis, "Nouvelle méthode de construction de variables", *11èmes Rencontres de la Société Francophone de Classification (SFC 04)*, Bordeaux, 2004.

[DBRBB04] R. BenMessaoud, S. Rabaseda, O. Boussaïd, F. Bentayeb, "OpAC : Opérateur d'analyse en ligne basé sur une technique de fouille de données", *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 04)*, Clermont-Ferrand, Janvier 2004; *Revue des Nouvelles Technologies de l'Information*, Vol. 2, Cépaduès, Toulouse, 35-46.

5.5 Chapters of book

[EAD07] K. Aouiche, J. Darmont, "Index and Materialized View Selection in Data Warehouses", *Encyclopedia of Database Technologies and Applications, Second Edition*, Idea Group Publishing, 2007.

[ELVML07] P. Lenca, B. Vaillant, P. Meyer, S. Lallich, "Association rule interestingness measures: experimental and theoretical studies", *Quality Measures in Data Mining*, Springer, Heidelberg, Germany, 2007, 51-76.

[EBL07] R. BenMessaoud, S. Loudcher-Rabaseda, "OLEMAR: an On-Line Environment for Mining Association Rules in Multidimensional Data", *Advances in Data Warehousing and Mining*, Vol. 2, Idea Group Publishing, 2007.

[EFBB07] C. Favre, F. Bentayeb, O. Boussaïd, "A Survey of Data Warehouse Model", *Encyclopedia of Database Technologies and Applications, Second Edition*, Idea Group Publishing, 2007.

[EZ07] D. Zighed, "Induction Graphs for Data Mining", *Studies in Classification, Data Analysis and Knowledge Organisation*, Springer, Heidelberg, Germany, 2007, 419-430 (In Selected Contributions in Data Analysis and Classification).

[EBBL07] R. BenMessaoud, O. Boussaïd, S. Loudcher-Rabaseda, "A multiple correspondence analysis to organize data cubes", *Databases and Information Systems IV - Frontiers in Artificial Intelligence and Applications*, Vol. 155(1), IOS Press, 2007, 133-146.

[EMD07] H. Mahboubi, J. Darmont, "Indices in XML databases", *Encyclopedia of Database Technologies and Applications, Second Edition*, Idea Group Publishing, 2007.

[EBAGB07] M. Bouet, M. Aufaure, P. Gançarski, O. Boussaïd, "Pattern Mining and Clustering on Image Databases", *Successes and New Directions in Data Mining*, Idea Group Publishing, 2007, 187-212.

[ERL07] R. Rakotomalala, T. LeNouvel, "Interactive Clustering Tree : Une méthode de classification descendante adaptée aux grands ensembles de données", *Revue des Nouvelles*

Technologies de l'Information, Vol. A1, Cépaduès, Toulouse, 2007, 75-94 (In Data Mining et apprentissage statistique : application en assurance, banque et marketing).

[EHD06] Z. He, J. Darmont, "Evaluating the Performance of Dynamic Database Applications", *Advanced Topics in Database Research*, Vol. 5, Idea Group Publishing, 2006, 294-319.

[EHZ06] H. Hacid, D. Zighed, "A Machine Learning Based Model For Content Based Image Retrieval", , 2006.

[ELTP06] S. Lallich, O. Teytaud, E. Prudhomme, "Association rules interestingness: measure and validation", *Quality Measures in Data Mining*, Springer, Heidelberg, Germany, 2006.

[ED05] J. Darmont, "Object Database Benchmarks", *Encyclopedia of Information Science and Technology*, Vol. 1, Idea Group Publishing, January 2005, 2146-2149.

[EMR05] F. Muhlenbach, R. Rakotomalala, "Discretization of Continuous Attributes", *Encyclopedia of Data Warehousing and Mining, Second Edition*, Idea Group Publishing, 2005, 397-402.

[EBA04] O. Boussaïd, M. Aufaure, "Spatial Data Warehouses: a methodological framework", *Advances in Spatial Analysis and Decision Making*, A.A. Balkema, 2004, 275-282.

[EAD07] K. Aouiche, J. Darmont, "Index and Materialized View Selection in Data Warehouses", *Encyclopedia of Database Technologies and Applications, Second Edition*, Idea Group Publishing, 2007.

Books/Proceedings (Eds.)

[FLP07] S. Lallich, D. Pastor, *Special Issue on the ASMDA International Symposium on Applied Stochastic Models and Data Analysis, Communications in Statistics - Theory and Methods*, Vol. 36(14), Taylor & Francis, January 2007 (Edited special issue).

[FLLG07] S. Lallich, P. Lenca, F. Guillet, *Actes du 3ème Atelier Qualité des Données et des Connaissances (QDC-EGC 07), Namur, Belgique*, EGC, Janvier 2007.

[FDB06] J. Darmont, O. Boussaïd, *Managing and Processing Complex Data for Decision Support*, Idea Group Publishing, April 2006.

[FBBDL05] F. Bentayeb, O. Boussaïd, J. Darmont, S. Loudcher-Rabaseda, *Actes de la 1ère journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA 05)*, *Revue des Nouvelles Technologies de l'Information*, Vol. B-1, Cépaduès, Toulouse, Juin 2005.

[FBGMT05] O. Boussaïd, P. Gançarski, F. Masseglia, B. Trousse, *Fouille de Données Complexes, Revue des Nouvelles Technologies de l'Information*, Vol. 3, Cépaduès, Toulouse, 2005.

5.6 PhD and HDR

[GL07] P. Lenca, "Des données à la décision : apprentissage, validation et exploitation de règles", Université Lumière Lyon 2, Novembre 2007 (Mémoire scientifique d'Habilitation à Diriger des Recherches).

[GF07] C. Favre, "Évolution de schémas dans les entrepôts de données : mise à jour de hiérarchies

de dimension pour la personnalisation des analyses", Université Lumière Lyon 2, Décembre 2007(Thèse de doctorat).

[GD06] J. Darmont, "Optimisation et évaluation de performance pour l'aide à la conception et à l'administration des entrepôts de données complexes", Université Lumière Lyon 2, Novembre 2006 (Mémoire scientifique d'Habilitation à Diriger des Recherches).

[GB06b] O. Boussaïd, "Evolution de l'entrepôtage des données complexes", Université Lumière Lyon 2, Décembre 2006 (Mémoire scientifique d'Habilitation à Diriger des Recherches).

[GB06] R. BenMessaoud, "Couplage de l'analyse en ligne et de la fouille de données pour l'exploration, la classification et l'explication des données complexes", Université Lumière Lyon 2, Novembre 2006 (Thèse de doctorat).

[GE06] W. Erray, "Extensions et nouvelles approches en Extraction des Connaissances à partir des données", Université Lumière Lyon 2, Décembre 2006 (Thèse de doctorat).

[GC06] F. Clerc, "Optimization and datamining for catalysts design", Université Lumière Lyon 2, septembre 2006 (Thèse de doctorat).

[GA05] K. Aouiche, "Techniques de fouille de données pour l'optimisation automatique des performances des entrepôts de données", Université Lumière Lyon 2, Décembre 2005 (Thèse de doctorat).

[GF05] E. P. Fangseu Badjio, "Evaluation qualitative et guidage des utilisateurs en Fouille visuelle de données", Université Lumière Lyon 2, 2005 (Thèse de doctorat).

[GC04] J. Clech, "Contribution Méthodologique à la Fouille de Données Complexes", Université Lumière Lyon 2, 2004 (Thèse de doctorat).

[GL04] G. Legrand, "Approche méthodologique de sélection et construction de variables pour l'amélioration du processus d'extraction de connaissances à partir de grandes bases de données", Université Lumière Lyon 2, 2004 (Thèse de doctorat).

[GB04] L. Baumes, "Combinatorial Stockastic Iterative Algorithms and High Throughput : from discovery to optimisation of heterogeneous catalysts", Université Lumière Lyon 2, 2004 (Thèse de doctorat).

[GP04]F.Poulet, "Visualisation et extraction de connaissances", Université Lumière Lyon 2, Novembre 2004 (Mémoire scientifique d'Habilitation à Diriger des Recherches).

APPENDICES

I. PERSONAL FILES OF ACTIVITIES

ARIGON, 59	LOUDCHER RABASEDA, 93
BAHRI, 61	MAHBOUBI, 95
BENTAYEB, 63	MAIZ, 97
BODIN-NIEMCZUK, 65	MARCELLIN, 99
BOUATTOUT, 67	MAVRIKAS, 101
BOUSSAID, 69	PRUDHOMME, 103
CHAUCHAT, 71	QURESHI, 105
DARMONT, 73	RAKOTOARIVELO, 107
EI SAYED, 75	RAKOTOMALALA, 109
FAVRE, 77	RALAIVAO, 111
GAUDIN, 79	SALEM, 113
HACHICHA, 81	STAVRIANOU, 115
HACID, 83	THOMAS, 117
HARBI, 85	VELCIN, 119
JULIEN, 87	VIALLANEIX, 121
LALLICH, 89	WEI, 123
LEFORT, 91	ZIGHED, 125

Anne-Muriel ARIGON

Current Position : Assistant professor
E-mail : anne-muriel.arigon@univ-lyon2.fr
Web site : <http://eric.univ-lyon2.fr/~amarigon>
Birth Date : 26/11/1980
Arrival Date : 01/10/2007
Administrative Charges :



Research topics

The first theme of my research topics is in bioinformatics area. The number of available biological sequences is growing very fast, due to the development of massive sequencing techniques. Sequence classification is needed and contributes to the assessment of gene and species evolutionary relationships. Classification methods are thus necessary to carry out these identification operations in an accurate and fast way. I develop a classification method dedicated to homologous sequence family databases, allowing to attribute a new sequence to a cluster using similarity measures. I used this classification method to implement two applications, HoSeqI (Homologous Sequence Identification) and MultiHoSeqI. They allow to automatically identify biological sequences and to rapidly add several sequences to a database. HoSeqI is accessible through a Web interface (<http://pbil.univ-lyon1.fr/software/HoSeqI/>) allowing to identify one or several sequences and to visualize resulting alignments and phylogenetic trees. MultiHoSeqI makes it possible to quickly add a large set of sequences to a family database in order to identify them, to update the database, or to help automatic genome annotation. Lately, I developed a chimera detection method and implement an application, ChiSeqI (Chimeric Sequence Identification), to automate the processes of classification of specific biological data, the bacterial 16S ribosomal RNA sequences, and of detection of chimeric sequences. The second theme of my research topics is in information system area and, more precisely, the multimedia data warehouse. Data warehouses are dedicated to collecting heterogeneous and distributed data in order to perform decision analysis. In numerous fields, like in medical or bioinformatics, multimedia data are used as valuable information in the decisional process. One of the problems when integrating multimedia data as facts in a multidimensional model is to deal with dimensions built on descriptors that can be obtained by various computation modes on raw multimedia data. I propose a new multidimensional model that integrates functional dimension versions allowing the descriptors of the multidimensional data to be computed by different functions. With this approach, the user is able to obtain and choose multiple points of view on the data he analyses. This model is used to develop an OLAP application for navigation into a hypercube integrating various functional dimension versions for the calculus of descriptors in a medical use case.

Publications

Arigon A.-M., Perrière G. and Gouy M., Automatic identification of large collections of protein-coding or rRNA sequences, A paraître dans Biochimie (2007), doi:10.1016/j.biochi.2007.08.006
Arigon A.M., Miquel M. and Tchounikine A. Multimedia data warehouses: a multiversion model and a medical application. Multimedia Tools Appl. 2007 October; 35(1): 91-108
Arigon A.M., Tchounikine A. and Miquel M. Handling multiple points of views in a multimedia data warehouse. ACM Transactions on Multimedia Computing, Communications and Applications. 2006 August; 2(3):199-218
Arigon A.M., Perrière G., Gouy M. (2006) HoSeqI: automated homologous sequence identification in gene family databases. Bioinformatics. 2006 Jul 15; 22(14):1786-7

Emna BAHRI

Current Position : PhD student
E-mail : Emna.bahri@univ-lyon2.fr
Web site :
Birth Date : 15/04/1981
Arrival Date : 17/10/2006



Research supervisor : Stéphane Lallich

Research topics

The recent advances in Communication and Information Technologies led to huge amounts of data, which exceeds the human processing and understanding capabilities. These data repositories contain an enormous amount of information, but require the development of intelligent tools in order to transform this information to knowledge. Those needs gave rise to data mining, which is an active area of research today.

In spite of great theoretical and practical achievements, data mining still lacks from low-scalability to large and real-world datasets. Two major problems are thus, the treatment of large data volumes, and the intolerance to the presence of noisy data. Even if these two problems seem classic, they still constitute major challenges in the area of machine learning.

My PhD goal is to design more powerful prediction systems, able to reach better success rates (seldom but not perfect), while being insensitive to noisy data. We can divide our prospects for this thesis into two parts. The first, which will be investigated this year, consists of providing a general and an exact definition of noise in order to handle it. The second part, which will be carried out later during my PhD, consists in finding new approaches and new algorithms to detect and manage the noise already modeled.

Publications

E.bahri, N.Nicoloyannis, M..Maddouri, « Amélioration du Boosting par combinaison des hypothèses antérieures », 14èmes Rencontres de la Société Francophone de Classification (SFC07), Paris, Septembre 2007.

E.bahri, N.Nicoloyannis, M..Maddouri « improving boosting by exploiting former assumptions », Third International Workshop on mining complex data (MCD07),warsaw,Poland.

FADILA BENTAYEB

Current Position : Associate professor since 2001
E-mail : bentayeb@eric.univ-lyon2.fr
Web site : <http://eric.univ-lyon2.fr/~bentayeb>
Birth Date : 15/05/1966
Arrival Date : 01/09/1999



Administrative Charges : Head of the bachelor of science in computer science and statistics (Informatique Décisionnelle et Statistique –IDS-) Member of the recruitment commission for mathematic-informatics and automatic at the university Lyon 2

Research topics

My current research interests regard complex data warehousing, integration of data mining techniques into data warehouses that we call on-line data mining and schema evolution in data warehouses. The special nature of complex data poses different and new requirements to data warehousing technologies, over those posed by conventional data warehouse applications. Indeed, current multidimensional data models fail to model the complex data found in some real-world application domains. Our main contribution is, then, the definition of a general framework to warehouse complex data. We used XML as the canonical standard to transform and store complex data from original data sources and we used the XML Schema to define the global ODS (Operating Data Storage) schema. Our approach presents several advantages. We can mention the unified format of complex data with XML and the use of data mining techniques for extracting relevant information necessary for building dimensional models.

On-line data mining: Data mining research has made many efforts to apply various mining algorithms efficiently on large databases. However, a serious problem in their practical application is the long processing time of such algorithms since they operate in main memory. We propose then a complete integrated solution for mining large databases into DBMSs without size limit in acceptable processing times. We think that data mining and databases should not remain separate components of the decision support. Indeed, data mining tools need integrated, consistent, and clean data. A database is constructed exactly by such preprocessing steps. Our first contribution consists in reducing the size of the learning database by building its contingency table, and our second contribution consists in reducing the number of database accesses by using bitmap indices. As a perspective of this work, we intend to extend our integrated approach to deal with multi-relational tables.

Data warehouse evolution : Due to the role of data warehouses in the daily business work of an enterprise, the requirements for the design and the implementation are dynamic and subjective. Therefore, data warehouse design is a continuous process which has to reflect the changing environment of a data warehouse, in other words, the data warehouse schema must evolve in reaction to the enterprise's evolution. My research focuses in integrating user's new analysis needs in the data warehouse process. We propose, then, a global approach composed by (1) the user's knowledge acquisition, (2) the user's needs integration, (3) a data warehouse schema update, and (4) an on-line analysis. Our main contribution consists in defining a user-driven approach that enables a data warehouse schema update. We integrate the specific user's knowledge representing new aggregated data under the form of If-Then rules into the data warehouse model. These rules are used to dynamically and automatically create new granularity levels in dimension hierarchies.

Publications

1. F. Bentayeb, J. Darmont, C. Favre, C. Udréa, "Efficient On-Line Mining of Large Databases", <i>International Journal of Business Information Systems</i> , Vol. 2, No. 3, 2007, 328-350. 2. J. Darmont, F. Bentayeb, O. Boussaïd, "Benchmarking Data Warehouses", <i>International Journal of Business Intelligence and Data Mining</i> , Vol. 2, No. 1, 2007, 79-104. 3. O. Boussaïd, J. Darmont, F. Bentayeb, S. Loudcher-Rabaseda, "Warehousing complex data from the Web", <i>International Journal of Web Engineering and Technology</i> , 2007. 4. C. Favre, F. Bentayeb, O. Boussaïd, "A Survey of Data Warehouse Model", <i>Encyclopedia of Database Technologies and Applications, Second Edition</i> , Idea Group Publishing , 2007. 5. C. Favre, F. Bentayeb, O. Boussaïd, "Evolution of data warehouses' optimization: a workload perspective", <i>9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2007)</i> , Regensburg, Germany, September 2007; LNCS .	
Scientific activities and valorisation	
Scientific programs and/or industrial collaborations	Phd Program Cécile Favre, « Data Warehouse evolutions », 2004-2007; Nora Maiz, « Integration by Mediation for data warehousing », 2005-2008; Ony Rakoarivélo, «On-line data mining for schema evolution in data warehouses», 2006-2009 Industrial collaborations LCL-Le Crédit Lyonnais (Rhône-Alpes Auvergne), 2004-2007 (Cécile Favre's thesis) Scientific programs ACI FodoMust (Fouille de données Multi-stratégie) 2005-2007
Editorial boards and program committees	- International Journal of Information Technology and Web Engineering Idea Group Publishing, 2007 - International Workshop « Ateliers sur les Systèmes Décisionnels », 2006-2007 Processing and Managing Complex data for Decision Support, Idea Group Publishing, 2005 - Journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 2006, EDA 2007) Editorial board and Committee steering member - International Journal of Biomedical Engineering and Technology (IJBET). SPECIAL EDITION "Warehousing and Mining Complex Data: Applications to Biology, Medicine, Behavior, Health and Environment", 2007 - French conference « Journées francophones sur les Entrepôts de Données et l'Analyse en ligne » (EDA), since 2005 International Multiconference on Computer Science and Information Technology (CSIT 06), Amman, Jordan , 2006 (Chair of session) Organizing Committee member French Conference EDA, Lyon, 2005 International Conference on Flexible Query Answering Systems (FQAS), Lyon 2004

Anouck BODIN-NIEMCZUK

Current Position : PhD student
E-mail : anouck.bodin-niemczuk@eric.univ-lyon2.fr
Web site :
Birth Date : 28/07/1984
Arrival Date : 01/09/2007



Research supervisor : Omar Boussaid and Sabine Loudcher Rabaséda

Research topics

The on-line analysis OLAP (On-line Analytical Processing) is a technology which comes to supplement data warehouses by proposing tools for visualization, exploration and navigation in data cubes in order to discover interesting information.

The user finds manually potential knowledge contained in data cubes. Indeed, OLAP technology makes it possible to visualize facts, to structure them according to analysis axes and to explore them but does not allow classification, explanation and prediction.

On the other hand, data mining employs machine learning techniques for visualization and description, for the structuring and classification, and for explanation and prediction.

During the last years, several works showed that it was possible to enrich the decision-making process using the coupling of on-line analysis and data mining [Imieliński 1996], [Han 1997], [Messaoud 2006].

Our approach consists in defining a new concept of on-line analysis by integrating data mining methods into OLAP data cubes.

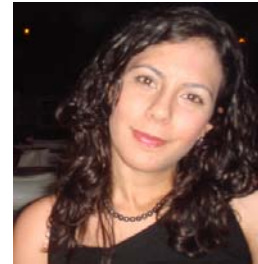
Han J., OLAP Mining: An Integration of OLAP with Data Mining, Proceedings of the 7th IFIP Conference on Data Semantics, 1997, Leysin, Switzerland

Imieliński T. and Mannila H., A Database Perspective on Knowledge Discovery, Communications of the ACM, vol. 39, n°11, 1996, pages 58-64.

Messaoud R.B., Couplage de l'analyse en ligne et de la fouille de données pour l'exploration, l'agrégation et l'explication des données complexes, Thèse de doctorat informatique, Université Lumière Lyon 2, novembre 2006.

Sonia BOUATTOUT

Current Position : PhD student
E-mail : bouattoursonya@yahoo.fr
Web site :
Birth Date : 03/07/1983
Arrival Date : 09/10/2006



Research supervisor : Omar Boussaid

Research topics

In the space domain, the construction of an operandi information and its availability to different types of users including mobile clients (embedded systems, PDAs, mobile phones, etc.) requires a change in traditional architectures of data warehouses. It is necessary to take charge, through computerization may leave some with interactivity, analysis Scenario by integrating them into the same process of storage. This results in the form of shares that may be triggered under given conditions including on the same sources OLTP, allowing access and act on detailed data. These treatments will be expressed in the form of analysis rules. They can make an effective contribution to improving the performance of these new architectures as did the pre-aggregated data in a classic OLAP cube. By integrating analysis rules in warehouses, they become active. The active data warehouses are new architectures, which create a dynamic where the OLAP cube is no longer an end in itself but on the contrary an intermediate step to design and produce information decision at the request with a return on decision-making sources, ETL, the sources OLTP ...

There are several approaches for designing such an architecture data warehouse of spatial data:

In the first approach, it is a traditional configuration centered on a warehouse or a datamart with a device of ETL from OLTP sources. The multidimensional model (star diagrams or snowflakes or flakes facts (Fact flake)) may have one or more spatial dimensions and / or measures space. A number of cubes can be constructed from complaints decision already identified. To give a dynamic to this setup, it must be complemented by a set of analysis rules corresponding to decision-making queries well established.

In the second approach, in contrast to the first one, there is no centralized multidimensional source (warehouse or datamart). The ETL device consists of a system of mediation to build at the request of cubic spatial data from space or non-space OLTP sources. The goals of analysis are supported by a mediator who identifies relevant sources, selects and extracts the data and propose a cube (or a set of cubes). The analysis scenarios have been identified and defined as analysis rules.

The third approach presents a solution that combines the first and the second approaches to take into account a set of multidimensional or OLTP sources, which are assumed to exist. The demand of a spatial information regarded as making a request may be met by a multidimensional structure (warehouse datamart, or cube) already existed. Otherwise, we have to build this cube from existing multidimensional sources or even from OLTP sources. This approach requires, of course, a mediation device which must support the request of spatial information. The analysis rules will complete of this configuration to provide an active character to this solution.

Publications

S. Bouattour, R.Ben messaoud, O. Boussaïd, "Modélisation de règles d'analyse dédiées aux entrepôts de données actifs", 2^{ème} édition de l'atelier des systèmes décisionnels (ASD 07), Sousse, 2007.

Omar BOUSSAID

Current Position : Assistant professor
E-mail : Omar.boussaid@univ-lyon2.fr
Web site : <http://eric.univ-lyon2.fr/~boussaid/>



Birth Date : 02/06/1954
Arrival Date : 01/09/1990

Administrative Charges : In charge of the IIDEE (Informatic Engineering and Economic Evaluation of Decision Support Systems) in Master IDS (Business Intelligence and Statistic)

Research topics

My research tasks relate to the complex data warehousing and on-line analyzing. The decision-making processes are based on the technology of the data warehouses and the OLAP. This technology is considered as mature in particular when the data are simple data. The challenge of today is to make evolve this technology by applying it to the complex data. To achieve this objective, I organized my work according to three research orientations:

1°) Integration of the complex data. After proposing an approach for describing complex data with the aid of UML and XML languages in order to store them into a target database, nowadays, we are working, as part of a thesis, on an approach of data integration based on a mediation system using ontologies for each of the data sources. The aim is to provide analysis contexts (datacubes built on the fly) and achieve on-line analysis.

2°) Modeling of complex data. We have chosen to use XML as a language of complex data modeling. We are currently working on methods of dimensional modeling to build XML-based complex data warehouse. As part of a thesis, we develop some works on the conceptual and dimensional modeling of complex data. We proposed a dimensional conceptual model of complex objects -representing complex data- which we describe at logic level with XML schemas. This model is being validated. The optimization of the physical models in XML warehouses is another objective of this thesis. Furthermore, we have developed an approach, which starting from a conceptual and multidimensional mode, to generate an XML complex data cube automatically. On the other hand, we are focused on further work to address the problem of performances in XML warehouses. To do that, we currently experienced a new method of fragmentation of the complex data warehouses. This work is being developed as part of another thesis.

3°) On-line analysis of complex data. In order to reinforce the capabilities of OLAP and expand its capacities to the explanation and the prediction, we work on the coupling of OLAP with data mining. As part of a thesis, we tried out different methods of coupling allowing to aggregate data, to improve the data representation in an OLAP cube and to apply the association rules as an analytical tool in an OLAP cube. We have proposed the theoretical foundations of these OLAP and data mining coupling. We are generalizing this formal framework to define any approach of coupling. We continue this work to extend this coupling in order to achieve the prediction analysis in OLAP cubes.

Publications

O. Boussaïd, Adrian Tanasescu , Fadila Bentayeb, Jerome Darmont, "Integration and Dimensional Modelling Approaches for Complex Data Warehousing", in Journal of Global Optimization, Vol. 37, No. 4, pp 571-591, Springer Netherlands, 2007

O. Boussaïd, R. Ben Messaoud, R. Choquet, S. Anthoard, "X-Warehousing : an XMLBased

<p>Approach for Warehousing Complex Data", 10th East-European Conference on Advances in Databases and Information Systems (ADBIS 06), in LNCS Vol. 4152, 39-54, Thessaloniki, Greece, September 2006</p> <p>R. Ben Messaoud, S. Loudcher Rabaséda, O. Boussaïd, R. Missaoui, "Enhanced Mining of Association Rules from Data Cubes", Proceedings of the 9th ACM International Workshop on Data Warehousing and OLAP (DOLAP'06), Arlington, VA, USA, ACM Press, November 2006, pp 11-18</p> <p>O. Boussaïd, J. Darmont, F. Bentayeb, S. Loudcher-Rabaseda, "Warehousing complex data from the Web", International Journal of Web Engineering and Technology, 2007.</p> <p>J. Darmont, O. Boussaïd, Eds., "Processing and Managing Complex Data for Decision Support", Idea Group Publishing, April 2006</p>	
Scientific activities and valorization	
Scientific programs and/or industrial collaborations	<p>2004-2007: FoDoMuSt (multistrategy data mining). Project with the LSIIT computer science and LIV geography labs (Strasbourg) for automatically identifying vegetation from satellite images. Funding from the Ministry of Research (ACI project)</p> <p>2002-2005: <i>CLAPI (spoken language corpus)</i>. Project with the ICAR linguistics lab for building, managing and exploiting a complex database of spoken language corpora. Funding from the Ministry of Research (ACI project).</p>
Editorial boards and program committees	<p><i>Editorial boards:</i> International Journal of Biomedical Engineering and Technology, Advances in Data Warehousing and Mining book series; EDA conferences steering committee</p> <p><i>Journal and book paper reviewing:</i> Journal of Intelligent Information Systems, International Journal of Foundations of Computers Science, International Journal of Software and Systems Modeling, The International Journal of Computers and Applications, "Multimedia Systems and Applications" book, Kluwer Academic Publishers, Ingénierie des Systèmes d'Information, numéro spécial : "Elaboration des entrepôts de données", Encyclopedia of Data Warehousing and Mining 2nd Edition.</p> <p><i>Conference program committees:</i> CE 06, EDA 05-07, ASD 06-07, MDDE, 01-02, SFdS, 03, FDC 04-08, SimSem 08, EGC 08, CSIT 06</p> <p><i>Conference organizing committees:</i> EDA 05, SFdS 03, ISMIS 02, ReTIS 01</p>
International activities	<p>Scientific stay as invited professor to the university Laval (Québec-Canada) at Laboratory CRG (Research center in Geomatic) of Pr. Yvan Bédard March 2006 ;</p> <p>Scientific Stay as invited professor to the university of Quebec in Outaouais (Canada) at Laboratory LARIM (Research Laboratory on Multimedia Information) of Pr. Rokia Missaoui in June-July 2006</p>

Jean-Hugues CHAUCHAT

Current Position : Full professor
E-mail : jean-hugues.chauchat@univ-lyon2.fr
Web site : <http://eric.univ-lyon2.fr/%7Echauchat/>
Birth Date : 06 July 1946
Arrival Date :



Administrative Charges : In charge of the strand SISE (Statistics & Informatics) in Master IDS (Business Intelligence and Statistic)
In charge of the double diploma Master/Magister (Statistics & Informatics) of University Lyon2 and the National University of Economics in Kharkov, Ukraine

Research topics

Statistics and Data Mining: models and validation
validation methods when the dataset is not collected using a two-stage, or a clustered, or a strata sampling design,
sampling in the whole dataset,
visualization.

Text mining.
Complex surveys analysis.
Applied statistics for managers.
Teaching statistics.

Publications

CHAUCHAT J.H., A. MORIN & R. RAKOTOMALALA, 2007. "Correcting the error rate estimation bias in Data Mining when the dataset comes from a two-stage sampling", *Statistics for Data Mining, Learning and Knowledge Extraction (IAST'07)*, Aveiro, Portugal.

RAKOTOMALALA, R., JH CHAUCHAT & F. PELLEGRINO, 2006. Accuracy Estimation With Clustered Dataset. In Proc. *Fifth Australasian Data Mining Conference (AusDM2006)*, Sydney, Australia. *CRPIT*, 61. Peter, C., Kennedy, P. J., Li, J., Simoff, S. J. and Williams, G. J., Eds., ACS. 17-22.

MORIN A, A. KOUOMOU-CHOUPPO, JH CHAUCHAT, 2005, Dimension reduction and clustering for query-by-example in huge image databases. Proc. *3rd world conference on Computational Statistics and Data Analysis, Limassol, Cyprus*, October 2005.

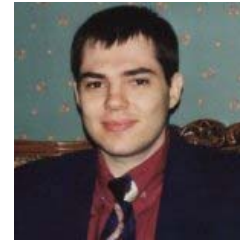
RADWAN J., CHAUCHAT J.-H. and DUMAIS J. 2004 "Automatic Recognition of Keywords using N-grams". In Jaromir A., editor, *Compstat'04 - Proceedings in Computational Statistics*, 1245-1254. Physica Verlag, Heidelberg, Germany.

PELLEGRINO F., CHAUCHAT J.H. & R. RAKOTOMALALA, 2002, "Can Automatically Extracted Rhythmic Units Discriminate among Languages?", *Proceedings of Speech Prosody 2002*, pp.562-565.

Scientific activities and valorisation		
Scientific and/or collaborations	programs and/or industrial	<p>Scientific programs</p> <p>2004-2005 Program EGIDE Econet (France – Croatia – Slovenia) « Fouille de données intelligente pour l'aide à la décision avec applications en médecine » - « Intelligent Data Mining in order to help decision making with applications in the medical field».</p> <p>2007-2008 Program EGIDE COGITO (France – Croatia) and PROTEUS (France – Slovenia) “Knowledge discovery and visualization for textual data”</p> <p>Industrial collaborations</p> <p>2006 Institut Fournier statistical and computer techniques for the analysis of large files that contain financial data received by insurance companies.</p> <p>2004 Commissariat Général au Plan Analysis and implementation of a national survey regarding the changes in the public sector.</p> <p>2004 Laboratoire SERVIER. Data Mining Advisor for the research of undesirable effects of new drugs..</p> <p>2002-2003 Région Rhône-Alpes. Computer-based modelization for the estimation of the total number of commuters between the towns of the Rhone-Alpes region.</p> <p>2000-2001 Crédit Agricole Centre-Est. Data mining for marketing : update of an online banking tool. Design and analysis of the clients' satisfaction in different market segments.</p>
Editorial boards and program committees		<p>Referee for the conference IASC 07</p> <p>Referee for the conference IASE'06</p>
International activities		<p>1997-98. Visiting Professor, University of Delaware, USA, College of Economics and Business Administration, Course taught : Data Analysis (Master in Economics)</p> <p>2005-2006-2007 Scientific Expert pour the research funds of Quebec</p> <p>Elected member of the International Statistical Institute</p> <p>Member of the International Association of Computing Statistics (IASC),</p> <p>Member of the International Association of Surveys Statisticians (IASS),</p> <p>Member of the International Association For Statistical Education (IASE).</p> <p>Member of the French Statistical Society (SFdS),</p> <p>Member of the French Classification Society (SFC),</p>

Jérôme DARMONT

Current Position Associate professor (HDR)
E-mail jerome.darmont@univ-lyon2.fr
Web site <http://eric.univ-lyon2.fr/~jdarmont/>
Birth date 15/01/1972
Arrival date 01/09/1999



Administrative charges Since 2003: Director, Computer Science and Statistics Department (DIS), School of Economics and Business Administration; U. Lyon 2
Since 2000: Head, Decision Support Databases group, ERIC lab

Research topics

Since my arrival at ERIC, I have been working on the border of databases and data mining. More precisely, I have lead my research following two complementary axes: data warehouse performance optimization and evaluation. The mix between databases and data mining is particularly obvious in the performance optimization part, since the automatic indexing and view materialization approach we proposed in K. Aouiche's PhD thesis (defended in 2005) is based on data mining techniques. Moreover, this research has lead to the design of generic benchmarks for data warehouse performance evaluation. Both these research topics allowed me to pass my "HDR" (qualification for supervising research) in 2006. Three PhD theses follow up this work. The first one (J.C. Ralaivao, started in 2003) aims at identifying performance factors in complex data warehouses. We have also proposed an XML-based reference architecture for complex data warehouses.

The second thesis' subject (H. Mahboubi, started in 2005) is dedicated to XML-native database management systems' performance optimization, and especially addresses two critical issues: response time and data volume. To help solve them, we have proposed XML data warehouse indexing, view materialization, fragmentation and distribution (over a grid) techniques.

The third thesis' objective (M. Hachicha, started in 2007) is to allow On-Line Analytical Processing over complex data stored in an XML warehouse. In this context, we have already proposed to formulate OLAP operators in an XML algebra, which helps execute classical OLAP queries over XML-native data (XML-OLAP or XOLAP).

On a longer term, my research project lies on the idea that XML must definitely become a pivot language for complex data warehousing, and I envisage three research axes: new, Web-based data warehouses architectures (Web 2.0, Web services); analytical extensions of the XQuery language for decision support; and exploiting semantic information about complex data for analysis. To restrict the scope of these research axes, I shall keep on addressing them from a performance point of view.

Publications

J. Darmont, F. Bentayeb, O. Boussaïd, "Benchmarking Data Warehouses", *International Journal of Business Intelligence and Data Mining*, Vol. 2, No. 1, 2007, 79-104

F. Bentayeb, J. Darmont, C. Favre, C. Udréa, "Efficient On-Line Mining of Large Databases", *International Journal of Business Information Systems*, Vol. 2, No. 3, 2007, 328-350

J. Darmont, O. Boussaïd, Eds., *Processing and Managing Complex Data for Decision Support*, Idea Group Publishing, April 2006

K. Aouiche, P. Jouve, J. Darmont, "Clustering-Based Materialized View Selection in Data Warehouses", *10th East-European Conference on Advances in Databases and Information Systems (ADBIS 06)*, Thessaloniki, Greece, September 2006; *LNCS*, Vol. 4152, 81-95

Z. He, J. Darmont, "Evaluating the Dynamic Behavior of Database Applications", *Journal of Database*

Scientific activities and valorization

Scientific programs and/or industrial collaborations	<p>2007-2008: <i>TAPEO</i>. Project with a young company for expressing a collaborative, collective intelligence from a Web site managing virtual stock exchange portfolios. Funding from the Rhône-Alpes Region: 29,500 €.</p> <p>2004-2007: <i>FoDoMuSt (multistrategy data mining)</i>. Project with the LSIT computer science and LIV geography labs (Strasbourg) for automatically identifying vegetation from satellite images. Funding from the Ministry of Research (ACI project): 69,000 €.</p> <p>2002-2005: <i>CLAPI (spoken language corpus)</i>. Project with the ICAR linguistics lab (Lyon 2) for building, managing and exploiting a complex database of spoken language corpora. Funding from the Ministry of Research (ACI project): 36,000 €.</p> <p>2003-2004: <i>MAP (anticipative, personalized medicine)</i>. Project with Dr Ferret, former physician of the French national soccer team, for storing, managing and analyzing complex medical data to optimize the health capital of high-level athletes. Funding from the Rhône-Alpes Region: 29,000 €.</p>
Editorial boards and program committees	<p><i>Editorial boards</i>: International Journal of Biomedical Engineering and Technology, Advances in Data Warehousing and Mining book series, IGI Editorial Advisory Review Board; EDA conferences steering committee</p> <p><i>Journal and book paper reviewing</i>: Data & Knowledge Engineering, Journal of Intelligent Information Systems, Ingénierie des Systèmes d'Information – Special Issue: Information retrieval and information mining; Encyclopedia of Database Technologies and Applications 2nd Edition, Encyclopedia of Information Science and Technology 1st and 2nd Editions</p> <p><i>Conference program committees</i>: IRMA 2005-2007, ASD 2006-2007, SAC-WT 2007, PICCIT 2007, CSIT 2006, ISWC 2004, FQAS 2004; EGC 2001-2008, EDA 2006-2007, INFORSID 2007, FDC-EGC 2006-2007, BDA 2003, SFdS 2003</p> <p>Conference organizing committees: EDA 2005, SFdS 2003</p>
International activities	<p><i>Since 2006</i>: Pedagogical director, French-Ukrainian double diploma (MSc in Computer Science and Statistics), in collaboration with the Kharkiv National University of Economics, Ukraine. PhD co-supervisions are planned.</p> <p>2003-2005: Collaboration with Dr. Zhen He, La Trobe University, Australia: joint research project and publications about database dynamic performance evaluation.</p> <p>2003: Collaboration with Pr. Le Gruenwald, University of Oklahoma, USA: joint research project and publications about database auto-indexing, student exchanges.</p>

Ahmad El SAYED

Current Position : PhD student
E-mail : asayed@eric.univ-lyon2.fr
Web site : <http://eric.univ-lyon2.fr/~asayed>
Birth Date : 11/03/1982
Arrival Date : 1/11/2004
Research supervisor : Djamel Abdelkader Zighed



Research topics

My PhD's goal essentially consists on developing intelligent methods and tools for allowing a semantic content-based information retrieval on heterogeneous documents (including texts and images). At a first stage, "semantics" were acquired using hand-crafted resources like Wordnet or domain ontologies in order to allow a query/document matching on the highest semantic level. We explored an approach where image and text contents in a document are analyzed automatically to represent each part by a set of terms. The deduced terms will be redirected into a fuzzy ontology enabling a conceptual representation of the whole document content. At a second stage, "semantics" were acquired automatically by means of a developed technique for knowledge acquisition from text. Furthermore, we designed a framework for learning taxonomy from a target text corpus. To achieve this, semantic relations between terms are first mined from text using a hybrid approach combining pattern-based and text mining techniques. Then, the entire set of relations is used for clustering terms into sense-bearing units that will be regarded to some extent as concepts. Following this, taxonomic relations will be deduced between the obtained concepts in order to build the hierarchy. To improve accuracy, the learned taxonomy is finally involved in our information retrieval environment where users interactions with the system are taken into account in order to launch a relevance feedback mechanism able to adapt the taxonomy to the user vision over text. As for future works, we're intending to use the acquired knowledge to perform a semantic parsing of text in order to represent it in predicate-argument structure rather than a bag of words. This can lead to the development of more sophisticated text-based applications for many areas like Question-Answering and Text Summarization.

Publications

- A. El Sayed, H. Hacid, A.D. Zighed. "Mining Semantic Distance Between Corpus Terms", In Proceedings of the ACM CIKM 1st Ph.D. Workshop in Information and Knowledge Management, PIKM 07, November 2007, *Lisboa, Portugal*.
- A. El Sayed, H. Hacid, A. D. Zighed "A Multisource Context-Dependent Semantic Distance Between Concepts", In *Proceedings of the 18th International Conference on Database and Expert Systems Applications DEXA'07*, September 2007- Regensburg, Germany
- A. El Sayed, H. Hacid, A. D. Zighed "Combining Text and Image for Content-Based Information Retrieval", In *Proceedings of the 2007 International Conference on Information and Knowledge Engineering, IKE 2007*, June 2007, Las Vegas, Nevada, USA.
- A El Sayed, H. Hacid, A. D. Zighed "A New Context-Aware Measure for Semantic Distance Using a Taxonomy and a Text Corpus", In *Proceedings of the 2007 IEEE International Conference on Information Reuse and Integration, IEEE IRI'07*, August 2007, Las Vegas, USA.
- A. El-Sayed, H. Hacid, D.A. Zighed, "Recherche d'Information par le Contenu des Données Hétérogènes", in *Actes des 3èmes Rencontres Inter-Associations RIA's 07*, March 2007, Toulouse.

Cécile FAVRE

Current Position : PhD student
E-mail : cecile.favre@univ-lyon2.fr
Web site : <http://eric.univ-lyon2.fr/~cfavre>
Birth Date : 19/08/1980
Arrival Date : 15/01/2004
Research supervisor : Fadila Bentayeb, Omar Boussaid



Research topics

After working on data mining integration into DBMSs, my current research works focus on data warehouse evolution.

Data warehouses store aggregated data issued from different sources to meet users' analysis needs in terms of decision support. As a matter of fact, user's requirements change over time and never reach a final state. Therefore, a data warehouse model cannot be designed in one step, it usually has to evolve progressively. We are thus interested in data warehouse model evolution. More specifically, we aim at involving users in the evolution process in order to supply them with personalized answers to their analysis needs.

Data warehouse evolution usually means evolution of its model. Meanwhile, a decision support system is composed of the data warehouse along with several other components, such as optimization structures, e.g. indices or materialized views. Thus, dealing with the data warehouse evolution also implies dealing with the maintenance of these structures. However, propagating evolution to these structures thereby maintaining the coherence with the evolutions on the data warehouse is not always enough. In some cases, redeployment of optimization strategies is required. Thus, we are interested in finding efficient solutions to ensure good performances, taking into account the model evolution. Since selection of optimization strategies is mainly based on workload according to user queries, one perspective is to lead the workload to evolve in order to test performances without waiting for a new workload for taking decisions on the optimization strategy.

Publications

C. Favre, F. Bentayeb, O. Boussaid, Evolution of Data Warehouses' Optimization: a Workload Perspective, 9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 07), Regensburg, Germany, September 2007 ; LNCS, Vol. 4654, 13-22.

C. Favre, F. Bentayeb, O. Boussaid, Dimension Hierarchies Updates in Data Warehouses: a User-driven Approach, 9th International Conference on Enterprise Information Systems (ICEIS 07): Databases and Information Systems Integration, Funchal, Madeira, Portugal, June 2007 ; 206-211.

F. Bentayeb, J. Darmont, C. Favre, C. Udréa, Efficient On-Line Mining of Large Databases, International Journal of Business Information Systems, Vol.2, N°3, 2007, 328-350.

C. Favre, F. Bentayeb, O. Boussaid, A Knowledge-driven Data Warehouse Model for Analysis Evolution, 13th ISPE International Conference on Concurrent Engineering: Research and Applications (CE 06), Antibes, France, September 2006 ; Frontiers in Artificial Intelligence and Applications, Vol. 143, 271-278.

C. Favre, F. Bentayeb, Bitmap Index-based Decision Trees, 15th International Symposium on Methodologies for Intelligent Systems (ISMIS 05), New York, USA, May 2005 ; LNAI, Vol. 3488, 65-73.

Rémi GAUDIN

Current Position : PhD student
E-mail : remi.gaudin@univ-lyon2.fr
Web site :
Birth Date : 01/06/1981
Arrival Date : 01/09/2004
Research supervisor : Djamel A. Zighed



Research topics

Most machine learning algorithms for classification problems are similarity/dissimilarity-based approaches. The similarity between instances is often explicitly expressed by a distance. In addition to the classical p-norm distance, other measures have been studied for the special case of time series, and among them the Dynamic Time Warping. Due to the observed performances variability of the previously proposed solutions considering various applications and benchmarks, a new distance called Adaptable Time Warping (ATW) is investigated. ATW is a form of generalization of both the classical Euclidian distance and DTW. A learning process which uses a genetic algorithm allows ATW to reach optimal solutions. We can prove that ATW efficiency is always at least equivalent to other distances use whatever the classification problem to be handled. We also demonstrate empirically the efficiency of ATW through different applications. Some are classical benchmarks for allowing comparative tests with previous studies, and two others are dealing with material science and more precisely zeolite crystalline structure. Effectiveness and stability are the two key advantages of ATW, which made it a promising methodology within our young research area.

Publications

R. Gaudin, L. A. Baumes, S. Jimenez, N. Nicoloyannis and A. Corma. "Improving Time Series Classification Using an Adaptable Distance and a Genetic Algorithm: Application to Automatic Classification of Zeolite Structures from X-Ray Diffraction". *The Third International Conference on Advanced Data Mining and Applications (ADMA'07)*, Harbin, China, August 6-8, 2007, LNCS series, Springer Press.

L. A. Baumes, M. Moliner, R. Gaudin, N. Nicoloyannis, A. Corma. "A robust methodology for high throughput identification of mixture of crystallographic phases from powder diffraction data". *Invited at E-MRS Fall Meeting 2007, Symposium on Genetic algorithms in materials science and engineering*, Warsaw, Poland, September 17-21, 2007.

R. Gaudin and N. Nicoloyannis. "An Adaptable Time Warping Distance for Time Series Learning". *Fifth International Conference on Machine Learning and Applications (ICMLA'06)*, Orlando, USA, December 14-16, 2006. IEEE Press, Pages 213—218.

R. Gaudin, S. Barbier, N. Nicoloyannis and M. Banens. "Clustering of Bi-Dimensional and Heterogeneous Time Series: Application to Social Sciences Data". *2006 International Conference on Data Mining (DMIN'06)*, Las Vegas, USA, June 26-29, 2006. CSREA Press, pages 10—16.

R. Gaudin et N. Nicoloyannis. "Apprentissage non supervisé de séries temporelles à l'aide des k-means et d'une nouvelle méthode d'agrégation de séries". *5èmes Journées d'Extraction et de Gestion des Connaissances (EGC'05)*, Paris, Janvier 2005. Presse RNTI, pages 201—212.

Marouane HACHICHA

Current Position : PhD student
E-mail : Marouane.Hachicha@univ-lyon2.fr
Web site : <http://eric.univ-lyon2.fr/~mhachicha/>
Birth Date : 7th June, 1983
Arrival Date : 3st September, 2007



Research supervisor : Jérôme Darmont

Research topics

The objective of this thesis is to allow OLAP analysis of complex data structured in XML data warehouses. This work consists on the design of a XML-OLAP (or XOLAP) algebra in the order to carry out traditional OLAP queries on native XML data.

This new formal framework represents the first step of our work. Then, it will be necessary to enrich this XOLAP algebra with new specific operators to the XML context. Then, it appears necessary to be able to carry out some operations like roll up and drill down on complex hierarchies of dimensions such as the ragged hierarchies [1]. Operators coupling the principles of OLAP and Data Mining could also allow the treatment (aggregation) of multimedia data resulting from the Web [2]. This work also aims at supporting the efforts of the extension of the XQuery language for the decisional applications.

In addition, to have a XOLAP algebra for the decisional complex data processing must allow the optimization of OLAP queries expressed in XQuery. Native XML Data Bases Management Systems (DBMS), although in a constant progress, present some limitations in term of performance and would profit largely from an automatic queries' optimization, particularly the decisional queries which are, generally, very expensive.

Finally, an implementation of this work on XOLAP is envisaged within the framework of a platform of complex XML data storage, under development at the ERIC laboratory [3]. The objective is to allow, through a simple and accessible interface since the Web, the construction and the handling of XML cubes of complex data.

Hakim HACID

Current Position : PhD student
E-mail : hhacid@eric.univ-lyon2.fr
Web site : <http://eric.univ-lyon2.fr/~hhacid/>
Birth Date : 05/03/1979
Arrival Date : 01/10/2004



Research supervisor : Pr. Abdelkader Djamel Zighed

Research topics

These last five years, particularly, due to the emerging new data acquisition technologies: scanner, satellite, video, web, etc. the available data related to the same problematic become in the same time larger, richer, and more heterogeneous, in one word, more complex. This situation concerns all the human activities such as medicine, astronomy, marketing, etc. The challenge of the next decade is the valorisation of these collected data. Access to the hidden knowledge in these large, heterogeneous, and unstructured databases is the most important task from a scientific and a technological point of view. Combining techniques issued from different domains, databases and data mining in our context, is a crucial question to face the new challenges in analysis, interrogation, and efficient access. Proceeding like that (combination of different techniques), databases can benefit from data mining to improve the quality of the answers. The data mining as for it can benefit from the optimisation strategies offered by the domain of databases to access larger datasets. This is very interesting since data mining models become more efficient when they learn on larger datasets.

Thus, we propose in this thesis to adapt a data mining, an instance-based learning particularly, structure to index large databases. This is done in order to incorporate certain intelligence in the indexing process and extending the functionalities of the indexing structures. Indeed, by doing that, the index can be useful not only to answer queries quickly but to offer also other possibilities like classification. From a data mining point of view, since neighbourhood graphs are hardly scalable to large datasets, we propose optimisations, issued from the databases domain, in order to make them operational on large datasets.

Another major problem in data mining and databases is that data collection doesn't hold in the memory. The databases techniques offer an efficient data access, sorting techniques, grouping techniques, and query optimisation techniques which are the basis of system's scalability. The majority of the methods issued from statistics, automatic learning, etc. consider that data hold completely in memory and do not consider the case where they do not satisfy this condition. We address also this problem in this thesis and we propose some solutions to manage it in the context of our study.

Publications

Hakim Hacid, Tetsuya Yoshida: Incremental Neighborhood Graphs Construction for Multidimensional Databases Indexing. Canadian Conference on AI 2007: 405-416

Ahmad El Sayed, Hakim Hacid, Djamel A. Zighed: A Multisource Context-Dependent Semantic Distance Between Concepts. DEXA 2007: 54-63

Ahmad El Sayed, Hakim Hacid, Djamel A. Zighed: A New Context-Aware Measure for Semantic Distance Using a Taxonomy and a Text Corpus. IRI 2007: 279-284

Hakim Hacid: Neighborhood Graphs for Semi-automatic Annotation of Large Image Databases. MMM (1) 2007: 586-595

Ahmad El Sayed, Hakim Hacid, Djamel A. Zighed: A Context-Dependent Semantic Distance Measure. SEKE 2007: 432-437

Nouria HARBI

Current Position : Assistant professor
E-mail : Nouria.harbi@univ-lyon2.fr
Birth Date : 27/08/1961
Arrival Date : 01/06/2006



Administrative Charges : In charge of the OPSI in Master IDS (Business Intelligence and Statistic)

Research topics

My research interests are about the decisional Information systems which integrate different categories of information processing: data warehouses, data marts, multidimensional databases. The main dimensions are:

Methodologies of the decisional Information Systems design and Conception: the objective is to build methods of conception and development for decisional information systems. These methods should allow conceiving, establishing, feeding and updating the different areas of storing data for decisional Information Systems. As a result, these methods will offer concepts, formalisms and steps adapted to decision systems, oriented towards the decision-makers. The proposed methodology will be validated by the respective tools.

Decisional data systems modelling: definition of data models based on a multidimensional data representation. The models should allow for a dependable, uniform and secured presentation of decisional data derived from various sources (Databases, Files, HTML, XML). These models should also permit the representation and archiving of all or part of the data warehouse.

Publications

Nouria Harbi, Omar Boussaid, Fadila Bentayeb, Propriétés d'un modèle conceptuel multidimensionnel pour les données complexes, Communication, EGC 2008, Nice Sophia Antipolis, Janvier 2008, 12 pages

Nouria HARBI, Henri SAVALL, Véronique ZARDET, , Spectral analysis of socio-economic diagnoses: qualimetric treatment of qualitative data, Communication, AOM HONOLULU, Août 2005, 20 p.

HARBI Nouria, SAVALL Henri, ZARDET Véronique, Analyse spectrale de diagnostics socio-économiques : traitement qualimétrique de données qualitatives, Communication, Colloque International AOM-RMD, Mars 2004, 26 p.

Nouria HARBI, Henri SAVALL, Véronique ZARDET, , Spectral analysis of socioeconomic diagnoses : qualimetric treatment of qualitative data, Communication, Colloque international AOM-RMD, Mars 2004, 17 p.

Editorial boards and program committees

ASD 2007
(IJBET) International Journal of Biomedical Engineering and Technology

Charbel JULIEN

Current Position : PhD student

E-mail : charbeljulien@hotmail.com

Web site : eric.univ-lyon2.fr/~jcharbel

Birth Date : 31/08/1978

Arrival Date : 01/10/2004



Research supervisor : Prof. Djamel Zighed university Lyon2 and Prof. Lorenza Saitta university of Turin

Research topics

I work on image modeling, Unsupervised and Semi-supervised learning. My research activity is usually related to statistical learning. My current interests are unsupervised classification of digital images.

Images may include different kinds of content descriptors from different levels. Until now no direct way has been found to extract high level semantic descriptors from images. Many low-level visual descriptor schemes have been proposed in the literature to extract visual content from images. Using these low-level visual descriptors, we can get high semantic level by inference.

Mixture distributions such as signatures or Gaussian Mixture Model (GMM) of color and texture are very interesting to describe the global composition of image. Mixture distributions, unlike histograms, try to abstract the content of image, color and texture, by a number of classes depending on the complexity of a particular image and this by using clustering techniques. To compute the distance between signatures linear optimization techniques are needed such as Mallows distance or Earth Mover's distance.

Unlike fixed size vector feature, we are interested of using a set of signatures to represent the image low-level visual content. The clusters in signatures representation mode are defined for each image individually. Simple images have a short signatures while complex images have long ones. Two clustering algorithms were tested to extract signatures from image: k-means algorithm and Expectation Maximization (EM) using Gaussian Mixture Model (GMM) with minimum description length to find the optimal number of clusters

This set of signatures that abstract the color and the texture of images is used afterward to compute the distance between pairs of images. The Earth Mover's Distance (EMD) is applied to each pairs of signatures of color and texture independently. The distance between two centers of clusters is computed using the Euclidean distance, this distance is used internally by the EMD algorithm. Afterward, the distance between two images is worked out using a linear combination of individual distances.

While signatures are very interesting to abstract the color and texture of images, continuous distribution like GMM offer a powerful way to abstract the color with correlation to spatial coordinates (x, y) . We appended the (x, y) to the color features and we compute a GMM of color plus the spatial correlation.

Image modeling can be extended to image-set modeling using mixture distributions. By image-set, we mean a collection of images that exhibit visual similarity in color content and/or in spatial relationships between colored regions. Image-sets are generated either by supervised categorization or by unsupervised clustering of image collection into groups. Modeling an image-set can be done by computing a mixture distribution that minimizes the distance to all mixture distributions of images within the image-set, as can be modeled by a mixture of mixture distributions; in this case the image-

set is partitioned into homogenous subsets, and for every subset a prototype is computed. Unlike fixed-size feature vector, where the centroid that minimizes the distance to a set of vectors can be computed by averaging the values in the feature vectors, mixture distribution's centroid needs a more complex technique to be computed. We use linear optimization algorithm, to compute a mixture distribution that minimizes the distance to all distributions in the image-set.

Publications

C. Julien, L. Saitta, "Image Database Browsing by Unsupervised Learning", 17th International Symposium on Methodologies for Intelligent Systems (ISMIS 08), Toronto, Canada, 2008; LNAI, Springer, Heidelberg, Germany.

C. Julien, L. Saitta, "Automatic Handling of Digital Image Repositories: A Brief Survey", 17th International Symposium on Methodologies for Intelligent Systems (ISMIS 08), Toronto, Canada, May 2008; LNAI, Springer, Heidelberg, Germany.

Stéphane LALLICH

Current Position : Full professor
E-mail : Stephane.lallich@univ-lyon2.fr
Web site : <http://eric.univ-lyon2.fr/~lallich/>
Birth Date : 20/09/1947
Arrival Date : 01/09/1997



Administrative Charges : Head of the master IDS (Business Intelligence and Statistic)
Head of the PhD program in Computer Sciences of University Lyon 2 (since 2007)

Research topics

In the past four years, I have developed my research around two main topics, the measures in data mining and the ensemble methods.

Concerning the measures, I was first interested by the measures which allow evaluating the quality of association rules, mainly in collaboration with Philippe Lenca, ENST Bretagne. We have identified various criteria for classifying usual objective measures, which allowed us to propose an automated procedure for assistance in choosing the most appropriate to the needs of a user. On the basis of these same criteria, we have also built a formal typology of the usual measures resulting from their properties according to the different criteria. This typology was compared to an experimental typology associated with the experimental behavior of these measures on different sets of test. We have also developed a presentation of usual measures parameterized according to the reference value of the confidence (independence value in case of targeting or 0.5 in case of prediction). This presentation allows to at the same time to emphasize the links between the measures which differ only by the value of the parameter and generate new control measures corresponding to a desired reference value of the confidence, which is particularly useful in case of targeting.

As a consequence of this latter work, we proposed a method to off-center the various entropies used in supervised learning to select at each step the best predictive attribute, for example Shannon entropy in C4.5 or Gini quadratic entropy in CART algorithm. In fact, at each node of a decision tree, we off-center the entropy in order that the off-centered entropy takes its maximum value for the distribution of the class in the node and not for the. This strategy improves systematically the precision on the class minority without degrading the results on the class majority.

Several of my works deals with ensemble methods : In the case of large high dimensional databases, with Elie Prudhomme (PhD ongoing), we proposed to replace neighborhood graphs by self organized maps to represent information from predictors. This substitution is moving from a complexity $O(n^3)$ to linear complexity depending on the number of individuals and variables, while retaining the capability of representation and navigation that is the interest of neighborhood graphs and putting in before a statistical cross-product basis of the map and taking into account the class of predictive performance generalization. To escape the dimensionality of the data, we propose to use a combination of Kohonen maps compiled from a limited number of predictors, thus viewing while improving the accuracy generalization. With Bissan Audeh (master thesis), we have developed a strategy that combines ensemble approach and sampling approach, which makes its complexity independent of the number of individuals and of the number of dimensions. With Romain Billot (master thesis), we have undertaken to adapt the boosting to clustering, and we propose UBLA, a method which leads to improve the value of the clustering quality coefficients.

Publications	
<p>Lenca P., Meyer P., Vaillant B., Lallich S. (2008), On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid, <i>EJOR, European Journal of Operational Research</i>, 184(2), 610–626.</p> <p>Lallich S., Lenca P., Vaillant B. (2007), Probabilistic framework towards the parametrisation of association rule interestingness measures, <i>MCAP, Methodology and Computing in Applied Probability</i>, 9(3), 447–463.</p> <p>Lallich S., Teytaud O. Prudhomme E. (2007), Association rules interestingness: measure and validation. In <i>Quality Measures in Data Mining</i>, pp. 251-275, Springer.</p> <p>Zighed D.A., Lallich S., Muhlenbach F. (2005), A statistical approach of classes separability, <i>Applied Stochastic Models in Business and Industry</i>. Vol. 21, No. 2, 2005, pp. 187-197.</p> <p>Lallich S., Muhlenbach F., Jolion J.-M.(2003), A test to control a region growing process within a hierarchical graph, <i>Pattern Recognition</i>, Pergamon, 36 (10), pp. 2001-2011.</p>	
Scientific activities and valorization	
Scientific programs and/or industrial collaborations	<p>Collaboration with <i>Hospices Civils de Lyon</i>, devoted to data mining and hospital acquired infections, 2007</p> <p>Collaboration with <i>Banque Populaire Rhône et Loire</i> to initiate the staff of the bank to data mining methods, 2005</p>
Editorial boards and program committees	<p>Editorial activity</p> <p>Lallich S., Pastor D. (2007), Special Issue on the ASMDA International Symposium on Applied Stochastic Models and Data Analysis, <i>Communications in Statistics - Theory and Methods</i>, Volume 36, Issue 14 January 2007 , pages 2475 – 2671</p> <p>S. Lallich, P. Lenca et F. Guillet (2007, 2008), Proceedings of the workshop <i>Qualité des Données et des Connaissances</i>, QDC 07, in association with Conference <i>Extraction et Gestion des Connaissances</i>, EGC 2007 and 2008.</p> <p>Program Committee</p> <p><i>International Conference on Data Mining</i>, DMIN, Las Vegas, USA : DMIN 2006, DMIN 2007 ;</p> <p>Conference <i>International Symposium on Applied Stochastic Models and Data Analysis</i> : ASMDA 2005 (Brest, France), ASMDA 2007 (Chania, Crète, Grèce)</p> <p>Conference <i>Extraction et Gestion des Connaissances</i>, EGC 2005 Paris, EGC 2006 Lille, EGC 2007 Namur, EGC 2008 Nice</p> <p>Workshop <i>Qualité des Données et des Connaissances</i> , associated with Conference EGC (2005 Paris, 2006 Lille, 2007 Paris, 2008 Nice)</p>
International activities	<p>Collaboration with Dragan Gamberger (Chercheur Rudger Boskovic Institute, Zagreb), as part of Program Egide, with Jean-Hugues Chauchat, ERIC Lyon 2 et Annie Morin, IRISA, Rennes ; one week in Zagreb, sept. 05 (overfitting in machine learning).</p> <p>Collaboration with <i>Faculté des Sciences Economiques et de Gestion de Jendouba</i>, to welcome during 4 months a master research student (Nejmeddine Ben Ouarred, 2007).</p> <p>Expert's valuation for the Fonds québécois de recherche sur la nature et les technologies technologies à Québec (2007)</p>

Virginie LEFORT

Current Position : Assistant professor
E-mail : virginie.lefort@univ-lyon2.fr
Web site : web-lefort.net
Birth Date : 24/09/1980
Arrival Date : 01/09/2007
Administrative Charges :



Research topics

Second order evolution (or indirect selection) corresponds to a situation where the individuals are not only selected on their fitness to an environment, but also on their ability to evolve “better”. Even if such a mechanism seems, *a priori*, very interesting in artificial evolution, it is not permitted by the structure of evolutionary algorithms because the evolutionary processes are fixed. Therefore, we propose a new evolutionary algorithm, RBF-Gene. It includes an intermediate level, the proteom (made of “proteins”), between the phenotype of an individual and its genotype, that allows for changes in the structure of the genome without changing the phenotype. We show the existence of an indirect selection in our algorithm, acting on genomes by changing the size of non coding sequences or the order of the genes.

Publications

Simultaneous optimization of weights and structure of an RBF Neural Network, V. Lefort, C. Knibbe, G. Beslon, J. Favrel, Talbi et al., *Artificial Evolution, proceedings of the 7th International Conference, EA 2005, Revised and selected papers, Lille, October 2005*, LNC3 3871, Springer

A bio-inspired genetic algorithm with a self-organizing genome: The RBF-Gene model, V. Lefort, C. Knibbe, G. Beslon, J. Favrel, Kalyanmoy Deb et al., *Genetic and Evolutionary Computation – GECCO 2004, Part II*, LNC3 3103, Springer, p. 406-407

Introducing « proteins » into genetic algorithms, V. Lefort, C. Knibbe, G. Beslon, J. Favrel, dans les actes de la conférence *Complex Systems, Intelligence and Modern Technology Applications (CSIMTA'04, Cherbourg)*, p. 181-186

Scientific activities and valorisation

International activities

Targeted Thematic Action (TTA) week, organized by François Kepès (Génopole d'Evry), on « New ideas for Genetic and Evolutionary Computation inspired by Recent Developments in Biology ». We were 8 : François Kepès (Genopole d'Evry), Jeremy Ramsden (Cranfield University, UK), James Foster (University of Idaho, Moscow, USA), Julian Miller (University of York, Heslington, UK), Wolfgang Banzhaf (University of Newfoundland, Canada), Steffen Christensen (Carleton University, Ottawa, Canada), Guillaume Beslon and me (INSA Lyon). After this week, we have written a paper published in *Nature Reviews Genetics*.

Sabine LOUDCHER RABASEDA

Current Position : Associate professor
E-mail : Sabine.Loudcher@univ-lyon2.fr
Web site : <http://eric.univ-lyon2.fr/~sabine/>
Birth Date : 27/10/1969
Arrival Date : 01/10/1998

Administrative Charges : Since 2003 : Assistant director, ERIC laboratory
1998-2002 : Head of the Computer Science and Statistics department, Institute of Technology, University of Lyon 2



Research topics

My research area is based on combining online analytical processing and data mining in order to improve the decision-making process, especially in the case of complex data. OLAP and data mining could be two complementary fields that interact together within a unique analysis process. The aim of this research is to propose new approaches based on coupling online analytical processing and data mining for exploration, aggregation, explanation and prediction of complex data in OLAP cubes.

In order to do so, we have established four main proposals :

The visualization of sparse data. According to the multiple correspondence analysis, we have reduced the negative effect of sparsity by reorganizing the cells of a data cube.

A new aggregation of facts in a data cube by using agglomerative hierarchical clustering. The obtained aggregates are semantically richer than those provided by traditional multidimensional structures.

An explanation of the possible relationships within multidimensional data by using association rules. We have designed a new algorithm for a guided-mining of association rules in data cubes.

An extension to prediction capacities. Our approach is based on the regression trees and consists in predicting the value measure of new data aggregates. By exploiting the decision rules, the user can anticipate the realization of future events. Moreover, the model makes it possible to improve the knowledge of the relations existing in the data.

Publications

R. Ben Messaoud, S. Loudcher Rabaséda, R. Missaoui, O. Boussaid. OLEMAR: an On-Line Environment for Mining Association Rules in Multidimensional Data. *Advances in Data Warehousing and Mining*, vol. 2. Idea Group Inc., 2007.

O. Boussaïd, J. Darmont, F. Bentayeb, S. Loudcher-Rabaseda, "Warehousing complex data from the Web", *International Journal of Web Engineering and Technology*, 2007.

Riadh Ben Messaoud, Omar Boussaid, Sabine Loudcher Rabaséda, "A Data Mining-Based OLAP Aggregation of Complex Data: Application on XML Documents", *International Journal of Data Warehousing and Mining*, 2(4) : 1-26. Idea Group Inc., 2006.

Riadh Ben Messaoud, Sabine Loudcher Rabaséda, Omar Boussaid, Rokia Missaoui, "Enhanced Mining of Association Rules from Data Cubes", *In Proceedings of the 9th ACM International Workshop on Data Warehousing and OLAP (DOLAP'2006)*, pp. 11-18, Arlington, VA, USA : ACM Press. November, 2006.

Riadh Ben Messaoud, Omar Boussaid, Sabine Loudcher Rabaséda, "Efficient Multidimensional Data Representation Based on Multiple Correspondence Analysis", *In Proceedings of the 12th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining (KDD'2006), pp. 662-667, Philadelphia, PA, USA : ACM Press. August, 2006.

Scientific activities and valorisation	
Scientific programs and/or industrial collaborations	<p><i>Since 2003:</i> Person in charge for FORMASUP RA for the annual inquire of the becoming to training students in the Rhone-Alpes area: 3,800 € per year.</p> <p><i>2004-2007: FoDoMuSt (multistrategy data mining).</i> Project with the LSIIT computer science and LIV geography labs (Strasbourg) for automatically identifying vegetation from satellite images. Funding from the Ministry of Research (ACI project): 69,000 €.</p>
Editorial boards and program committees	<p><i>Editorial boards:</i> International Journal of Biomedical Engineering and Technology ; EDA conferences steering committee</p> <p><i>Journal and book paper reviewing:</i> Revue des Nouvelles Technologies de l'Information (RNTI), Processing and Managing Complex Data for Decision Support, 2005</p> <p>Conference program committees : ASD 2006-2007, EDA 2006-2007, PKDD 2004, ISWC 2004, JDS-SFdS 2003</p> <p>Conference organizing committees: EDA 2005, SFdS 2003</p>

Hadj MAHBOUBI

Current Position : PhD student
E-mail : hadj.mahboubi@eric.univ-lyon2.fr
Web site : <http://eric.univ-lyon2.fr/~hmahboubi>
Birth Date : 17/03/1981
Arrival Date : 01/09/2005



Research supervisor : Jérôme Darmont

Research topics

Decision-support applications currently exploit more and more heterogeneous data from various sources. In this context, XML can greatly help in their integration and warehousing. However, decision-support queries, exploiting XML data warehouses, are generally complex because they involve several join and aggregation operations. In addition, XML-native database management systems present poor performances when data volume is very large and queries are complex.

Several studies address the issue of designing and building XML data warehouses. These works propose different architectures and they differ on the way they represent facts and dimensions. Hence, we define a unified XML data warehouse model. Entirely based on XML formalism, this model is actually the translation of a classical snowflake schema. It presents better performance compared to the existing models.

In order to guarantee the best performance when accessing warehouse data, we propose a new index that is specifically adapted to the multidimensional architecture of XML warehouses [5]. It eliminates join operations. We also design and implement an automatic strategy for the selection of XML materialized views that exploit a data mining technique (clustering of the query workload) [2].

We actually focus on designing a distributed XML data warehouse system to reduce warehouse storage cost and to perform parallel execution of queries. Traditionally, this process involves data fragmentation and fragments repartition. So, we propose to adapt existing fragmentation techniques (as defined in the relational context) to partition XML data warehouses [4]. After that, a repartition architecture (distributed system, peer to peer network or data grid) must be defined. The choice of the architecture is based on the query performance evaluation over these architectures. Hence, a distributed decision-support query processing mechanism is also defined.

Publications

- [1] Hadj Mahboubi and Jérôme Darmont. *Indices in XML databases* . Encyclopedia of Database Technologies and Applications, Second Edition. Idea Group Publishing. 2007.
- [2] Hadj Mahboubi, Kamel Aouiche, Jérôme Darmont, *Materialized View Selection by Query Clustering in XML Data Warehouses* , 4th International Multiconference on Computer Science and Information Technology (CSIT'06), Amman, Jordan , 2006, pages 68-77
- [3] Hadj Mahboubi, Jérôme Darmont, *Benchmarking XML data warehouses* , Atelier Systèmes Décisionnels (ASD'06), 9th Maghrebien Conference on Information Technologies (MCSEAI'06), Agadir, Maroc , December 2006
- [4] Hadj Mahboubi, Jérôme Darmont, *Fragmentation des entrepôts de données XML*, 3èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA'07), Poitiers , 2007, pages 177-190
- [5] Hadj Mahboubi, Kamel Aouiche, Jérôme Darmont, *Un Index de Jointure pour les Entrepôts de données XML*, 6èmes Journées Francophones Extraction et Gestion des Connaissances (EGC'06), Lille , 2006, pages 89-94

Nora MAIZ

Current Position : PhD student
E-mail : nmaiz@eric.univ-lyon2.fr
Web site :

Birth Date : 08/04/1979
Arrival Date : 01/10/2005



Research supervisor : Omar Boussaid and Fafila Bentayeb

Research topics

In a data warehousing process, data integration is one of the most important phases. Centralized data warehouse is a solution for companies that handle static data. However, when data change, this solution is not practical because of the refreshment cost. We believe that data integration by mediation can solve this problem by allowing the construction of a mediation system for building an analysis context on-the-fly using data from their real sources.

The use of ontologies in the mediation process allows semantic and structural integration. In our work, we try to propose a new mediation system based on a hybrid architecture of ontologies modelled according to GLAV (Generalized Local As View) model. The hybrid architecture defines a local ontology for each data source and a global ontology viewed as a mediator. The integration model defines how sources, local and global ontologies are mapped. So we propose an ascending method for building ontologies, which starts by building local ontologies. After that, we use data mining technics to merge local ontologies. This method facilitates the semantic reconciliation between data sources. We use OWL (Ontology Web Language) for defining ontologies and mappings between data sources and ontologies. Moreover, user queries are expressed in the specific language that we propose which handles global ontology concepts and local ontology properties because we assume that the user is expert in his domain. User queries are decomposed by the rewriting algorithm in order to obtain a set of equivalent subqueries that are sent to the corresponding data sources to be executed. After that, the subqueries are recomposed to obtain the final result.

Publications

N. Maiz, O. Boussaid and F. Bentayeb, "Ontology-based mediation system", 13th ISPE International Conference on Concurrent Engineering: Research and Applications (CE06), Antibes, France, September, 2006.

N. Maiz, O. Boussaid and F. Bentayeb, "Ontology-based data integration in datawarehouses", 18th Information Ressources Management Association (IRMA) International Conference, Vancouver, Canada. 2007.

N. Maiz, O. Boussaid and F. Bentayeb, " Un système de médiation basé sur les ontologies pour l'entrepasage de données". Atelier Systèmes Décisionnels (ASD 06), 9th Maghrebien Conference on Information Technologies (MCSEAI 06), Agadir, Maroc, December, 2006.

N. Maiz, O. Boussaid and F. Bentayeb, " Fusion automatique des ontologies-OWL par classification hiérarchique pour la conception d'un entrepôt de données", 4ème atelier Fouille de Données Complexes dans un Processus d'Extraction des connaissances (FDC-EGC 07), Namur, Belgique, January 2007.

N. Maiz, K. Aouiche, J. Darmont, " Sélection automatique d'index et de vues matérialisées dans les entrepôts de données". 2ème journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA 06), Versailles, Juin 2006; Revue des Nouvelles Technologies de l'Information, Vol. B-2, 89-104.

Simon MARCELLIN

Current Position : PhD student
E-mail : smarcellin@eric.univ-lyon2.fr
Web site : -
Birth Date : 08/07/1981
Arrival Date : 15/09/2004
Research supervisor : Djamel A. Zighed



Research topics

Our main research topic is machine learning on imbalanced datasets (when an important class is weakly represented). We propose some methods to deal with this kind of data, especially using decision trees-based algorithms :

An asymmetric entropy measure for decision trees: a new splitting criterion taking into account the class imbalance. We also propose a framework for asymmetric entropy measures.

An adaptation of Laplace estimation of probabilities, adapted to imbalanced datasets.

Decision rules adapted to imbalanced data: to obtain a prediction model from a decision tree, a decision rule must be applied on each leaf. We propose different decision rules.

Performance measures of prediction models adapted to imbalanced data, and empirical comparison of asymmetric splitting criteria using ROC curves.

This thesis is financed by the French Ministry of Research and Industry (*CIFRE* financing)

Publications

D.A. Zighed, S. Marcellin, G. Ritschard « Mesure d'entropie asymétrique et consistante », *Revue des Nouvelles Technologies de l'Information*, E-9 (Vol. I), EGC'2007, 81-86.

S. Marcellin, D. Zighed, G. Ritschard, "Une mesure d'entropie asymétrique pour les arbres de décision", *38^{ème} journées des statistique (JDS 06)*, Clamart, France, Mai - Juin 2006

S. Marcellin, D. Zighed, G. Ritschard, "An asymmetric entropy measure for decision trees", *11th Information Processing and Management of Uncertainty in knowledge-based systems (IPMU 06)*, Paris, France, July 2006, 1292-1299.

S. Marcellin, D. Zighed, G. Ritschard, "Detection of breast cancer using an asymmetric entropy measure", *17th Computational Statistics (COMPSTAT 2006)*, Rome, Italy, August - September 2006, 975 - 984.

D. Zighed, S. Marcellin, G. Ritschard, "An asymmetric entropy measure for decision trees", *Knowledge Extraction and Modeling, Island of Capri, Italy*, September 2006.

Efthimios MAVRIKAS

Current Position : PhD student
E-mail : efthimios.mavrikas@eric.univ-lyon2.fr
Web site : -
Birth Date : 04/03/1979
Arrival Date : 06/01/2003
Research supervisor : Djamel. ZIGHED



Research topics

Cultural Heritage documents deal with objects/artifacts and the people that created, owned, used, or (re)discovered them. Their fates are intertwined in unique and complex stories forming a cumulative body of knowledge, often fragmented across large online document collections. While our collective memory has explicitly documented these stories, the heterogeneity of the available sources creates islands of information that can only be implicitly connected by a limited, expert audience.

My current research work aims to define a semantically consistent framework for the online presence of Cultural Heritage document collections, set upon a participatory centre stage and supported by a shared knowledge model. In this framework, Cultural Heritage document contributors benefit from knowledge-rich document processing modules which analyse and classify each contribution, capture the notion of time and the unfolding of events spanning single or multiple documents, and establish meaning connectivity over the entire collection. Overall, this framework assists a scholarly audience with the exploration of online Cultural Heritage document collections, and offers an informed tap into the collective memory scattered therein.

Keywords: Discourse Analysis, Ideology, CIDOC CRM, WorldNet, PLSA, Scripts.

Publications

Efthimios C. Mavrikas, Evangelia Kavakli and Nicolas Nicoloyannis (2005) The Story between the Lines: Exploring Online Cultural Heritage Document Collections using Ontology-based Methods, *Annual Conference of the International Committee for Documentation of the International Council of Museums (CIDOC 2005)*, Zagreb, Croatia, May 2005.

Efthimios C. Mavrikas, Vagelis Stournaras and Christis Konnaris (2005) Historical Memory Preservation on the Semantic Web: the Case of the Historical Archive of the Aegean - Ergani, *33rd International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA 2005)*, Tomar, Portugal, March 2005.

Efthimios C. Mavrikas, Evangelia Kavakli and Nicolas Nicoloyannis (2004) Ontology-based Narrations from Cultural Heritage Texts, *5th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST 2004)*, Ename, Belgium, December 2004.

Efthimios C. Mavrikas, Nicolas Nicoloyannis and Evangelia Kavakli (2004) Cultural Heritage Information on the Semantic Web, *14th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2004)*, Northamptonshire, UK, October 2004, Springer LNAI, vol. 3257, pp. 477-478.

Dimitris C. Papadopoulos and Efthimios C. Mavrikas (2003) Peer-to-Peer Ways to Cultural Heritage, *31st International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA 2003)*, Vienna, Austria, April 2004.

Elie PRUDHOMME

Current Position : PhD student

E-mail : eprudhomme@eric.univ-lyon2.fr
Web site : eric.univ-lyon2.fr/~eprudhomme/
Birth Date : 23/01/1979
Arrival Date : 01/10/2005
Research supervisor : Stéphane Lallich



Research topics

The learning process encounters many difficulties to analyze large amount of data. Indeed, algorithms must be of linear complexity to handle these datasets and some theoretical problems, related to high dimensional spaces, appear and degrade their predictive capacity. Furthermore, end-user needs to understand and interact with the prediction.

The selection of data “features” - variables or association rules that can be derived from them - is a simple response to this problem, applied at the pre-processing stage. In high-dimensional space, this selection requires a large number of tests from which arise a number of false discoveries. We have proposed an original non-parametric control method. A new criterion, UAFWER, defined as the risk of exceeding a pre-set number of false discoveries, is controlled by BS-FD, a bootstrap based algorithm that can be used on one- or two-sided problems. We have illustrated the usefulness of that procedure by the selection of differentially interesting association rules on genetic data.

High-dimensional space prevents algorithms from doing a data representation. Nevertheless, in some applications, this representation can help the user to make good use of the learning model. For that purpose, we propose an ensemble approach to overcome problems related to high dimensional spaces. Self-organized map, which allows both a fast learning and a navigation through the data is used like base classifiers to learn several features subspaces. Further, a genetic algorithm is used to optimize diversity of the ensemble by relying on an adapted error measure. This approach offers similar representation capabilities and competitive prediction performance with boosting and random forests.

Publications

Lallich S., Teytaud O., Prudhomme E. (2006), Statistical inference and data mining: false discoveries control. *Proc. of 17th COMPSTAT Symposium of the IASC*, La Sapienza, Rome, août 2006, pp. 325-336.
Prudhomme E. and Lallich S. (2005), Quality measure based on Kohonen maps for supervised learning of large high dimensional data, *Proc. of ASMDA 2005*, pp. 246-255, Brest.
Prudhomme E. et Lallich S. (2007), Ensemble prédicteur fondé sur les cartes auto-organisatrices adaptées aux données volumineuses, *Actes EGC'07, RNTI-E-9*, vol. 2, pp. 473-484, Namur, Belgique.
Prudhomme E. et Lallich S. (2005), Validation statistique des cartes de Kohonen en apprentissage supervisé, *Actes EGC 2005, RNTI-E-3*, vol. 1, pp. 79-90, Paris.
Lallich S., Prudhomme E. et Teytaud O. (2004), Contrôle du risque multiple en sélection de règles d'association significatives, *Actes EGC'04, RNTI-E-2*, vol. 2, pp. 305-316, Clermont-Ferrand.

Taimur QURESHI

Current Position : PhD student
E-mail : taimur.q80@gmail.com
Web site : NA
Birth Date : 10/09/1980
Arrival Date : 01/10/2006



Research supervisor : D.A.Zighed

Research topics

The goal of this PhD can be divided into two parts:

Practical Part:

Design and implementation of a generalized algorithm for decision trees and induction graphs.

Development of a software that incorporates the above algorithm and thus, creation of a test bed for experimentation using vast data and understanding of various algorithms and techniques.

Development of an internet based collaborative tool that implements a huge data store for decision trees and induction graph based resources e.g. articles with their summaries, tools, books etc.

Theoretical Part: The theoretical part concerns with the development of new techniques and methods in the area of decision trees and experimentation with huge data on the created test bed and obtaining applicable results.

1) Practical Part:

In this portion a generalized decision tree and induction graph algorithm has been conceived and designed using flowcharts and object oriented designing techniques. The concept of the generalized algorithm is to create such an algorithm which is generic and can be used to implement anyone of the existing decision tree or induction graph techniques. We have implemented various discretization algorithms such as Chi merge, FUSBIN, FUSINTER, MDLPC, CONTRAST and also various decision trees as ID3, C4.5, CART and Arbogodai. We have implemented our algorithm in R and once the object oriented implementation is completed, we will transfer it into our software which is implemented in Java. The software has a user friendly interface that converts many types of data e.g. text, xml, db etc into a table structure. After that the user can select the type of technique to use on that data and the results shall be given as a graphical output. The third phase of the implementation is development of an internet based collaborative platform for decision tree and induction graph based resource sharing. We have developed a “Wikipedia” like tool for information sharing and editing. It shall contain resumes and sources of many articles, tools and platforms and a test bed; thus forming a complete resource for decision tree and induction graphs.

2) Theoretical Part:

From various studies done earlier, we know that the learning sample is an approximation of the whole population, so the optimal discretization built on a single sample set is not necessarily the global optimal one. Whereas, we proved that resampling gives a better estimate of the distretization point distribution in terms of acheiving a well-defined distribution. We have created a discretization point selection protocol which selects cut points from a certain frequency distribution achieved by resampling. This protocol selects the discretization points from a given frequency point distribution

having higher probability of occurrences and splits on those points if a certain criterion (e.g. entropy) is met. When we apply that protocol, it significantly improves the quality of discretization and prediction rate and thus, nearing to a global optimal solution. Moreover, the same protocol when applied to the frequency point distribution of random samples, achieved much lesser improvements in the prediction rate as compared to bootstrap. We applied the discretization point selection protocol (after resampling) to various methods on the breiman waveform dataset. Except for Chi-Merge, all the other methods provide small variations in terms of prediction rates. MDLPC performs the best and FUSBIN achieves the best time complexity, which is a key point when dealing with a lot of examples.

We applied the above mentioned resampling methodology in the context of fuzzy or soft discretization in decision trees. Our soft discretization technique gives better prediction rates than the hard discretization based methods. As ongoing work, we are applying this soft discretization in building soft decision trees and thus, will try to show that this method will also improve the classification accuracy of decision trees.

Publications

IEEE ICSEA-2004: Integration of Mobile IP and Adhoc Networks with Multi-homing and Smooth Handoff capabilities.

IEEE 16th IST Mobile Summit, Budapest, Hungary: A Network Layer based Hard Real Time Protocol for Wireless Sensor Networks.

Ony RAKOTOARIVELO

Current Position : PhD student
E-mail : orakoto@eric.univ-lyon2.fr
Web site :
Birth Date : 27/08/1981
Arrival Date : 07/11/2006
Research supervisor : Fadila Bentayeb



Research topics

Data warehouses provide an integrated and consistent view on all enterprise data which are relevant for the OLAP analysis. This analysis requires time-variant and non-volatile data. Thus, dimension updates and schema evolutions on the data warehouse are prohibited because they can induce data loss or erroneous results. However, needs and data change constantly. As a result, requirements are not, then, satisfied and some trends are not explored. This is the reason why data warehouse schema evolution becomes an important research topic. In our research, we are interested in the following issue: how can this problem be treated by using data mining techniques? We have proposed a schema evolution operator based on the k-means clustering algorithm. This leads us to the very interesting research topic of online data mining which is how to integrate effectively data mining methods in a RDBMS (Relational DataBase Management System).

Publications

O. Rakotoarivelo, F. Bentayeb, "Evolution de schéma dans les entrepôts de données: utilisation de la méthode des k-means", 4ème atelier Fouille de Données Complexes dans un Processus d'Extraction des Connaissances (FDC-EGC 07), Namur, Belgique, Janvier 2007.

O. Rakotoarivelo, F. Bentayeb, "Evolution de schéma par classification automatique pour les entrepôts de données", 3èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 07), Poitiers, Juin 2007; Revue des Nouvelles Technologies de l'Information, Vol. B-3, 99-112

Ricco RAKOTOMALALA

Current Position : Assistant professor
E-mail : Ricco.rakotomalala@univ-lyon2.fr
Web site : <http://eric.univ-lyon2.fr/~ricco/>
Birth Date : 19 July 1967
Arrival Date : 01 Sept 1995



Administrative charges : Joint manager of the strand SISE (Statistics & Informatics) in Master IDS (Business Intelligence and Statistic)

Research topics

My research's activities are mainly the applications of data mining. We try to characterize the outline of a successful data mining process, in the area, needed to be precisely defined. One of our goal, but not restricted to, is the determination of the most effective strategies in this context.

One of my highlighted project, with many publications, is the automatic classification of protein from their primary structure. Carried out in collaboration with Mr. Elloumi of the Faculty of Science of Tunis (Tunisia), this work is an important step in the defense of PhD thesis of Mr. Mhamdi at the beginning of 2008. The principal task is the comprehension of data comprising a few observations but a very large number of descriptors, that are being automatically generated from very rough techniques such as the n-grams. We developed rapid approaches for dimensionality reduction by carrying out a very aggressive feature selection without reducing the accuracy of the classifiers. The experiments show, without surprise, that the margin maximization methods such as SVM (Support Vector Machines) are powerful. Surprisingly, other strongly regularized approaches such as PLS regression, not well known in the machine learning community, are also very accurate.

Another project is the automatic classification of planktons from scanned images. Carried out in collaboration with the team of Mr. Gorsky of the laboratory of Oceanography of Villefranche-sur-Mer (France), this project aims to industrialize the automatic categorization of planktons (Plankton Identifier Project -- <http://www.obs-vlfr.fr/>). We also handle unstructured datasets here, with the original data description being the image. Beyond the research of the most effective strategy, the question of validation of performances arise. Indeed, the composition of the marine environment is very dynamic, according to the location, and the period. We must take a new view of the validation problem, as the traditional indicators (accuracy rate in particular) are not really adequate. We must produce results which are transposable in various contexts.

This project is accompanied by important software development activity software which is placed freely at the disposal of the scientific community (http://www.obs-vlfr.fr/~gaspari/Plankton_Identifier/index.php).

Then, the last highlighted project, more personal, is the development of the TANAGRA data mining software, freely available with source code on the web (<http://chirouble.univ-lyon2.fr/~ricco/tanagra/>). The project, started in 2003, comprises of more than 200.000 lines of source code today. With about 100 implemented methods, it covers a very broad field of the data mining techniques, starting from the statistical approaches (parametric and not-parametric tests), to

the machine learning algorithms (supervised and unsupervised, association rule mining), while passing by the traditional techniques of the exploratory data analysis (factorial methods, etc.).

The diffusion of the software is accompanied by about sixty tutorials in English and in French. Our Web site has a rather good frequentation. On average, we count 130 visitors daily since the beginning of the year 2007.

It is a very important project for me. To give a larger base to the dissemination of the knowledge, I started to put on-line detailed course notes. The described techniques can be applied directly via the free software, via TANAGRA of course, but also using tools such as the R-project software or a spreadsheet. We count nearly 50 visitors per day since the starting of the website (January, 2007). This value is all the more interesting since all the documents are in French (http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html). In addition, we developed a website which directs on the most interesting courses notes that one can find on the Web about the various subjects which are related to the data mining process (<http://eric.univ-lyon2.fr/~ricco/data-mining/>).

Publications

E. Antajan, R. Rakotomalala, S. Gasparini, M. Picheral, L. Stemmann, G. Gorsky, « Automatic quantification and recognition of major zooplankton groups in a North Sea time series using the Zooscan imaging system», in the Proceedings of the 4th International Zooplankton Production Symposium, pp. 189-190, Hiroshima, Japan, 2007.

Chauchat J.H., A. Morin & R. Rakotomalala, 2007. "Correcting the error rate estimation bias in Data Mining when the dataset comes from a two-stage sampling", Statistics for Data Mining, Learning and Knowledge Extraction (IAST'07), Aveiro, Portugal.

F. Clerc, D. Farrusseng, R. Rakotomalala, N. Nicoloyannis, C. Mirodatos, "Meta Modeling for Combinatorial Catalyst Optimization", International Journal of Computer Science and Network Security, vol. 6, n°10, pp.256-262, 2006.

R. Rakotomalala, F. Mhamdi, "Supervised and Unsupervised Feature Reduction for Protein Classification", WSEAS International Journal -- WSEAS Transactions on Information Science and Applications, vol. 3, n°12, pp. 2448-2455, 2006.

A. Morineau, Rakotomalala R. "The TVpercent Criteria to Eliminate Uninformative Models among Association Rules", in Electronic Proceedings of Knowledge Extraction and Modeling, IASC-INTERFACE-IFCS Workshop, KNEMO'06, Anacapri, Italy, 2006.

Jean-Christian RALAIVAO

Current Position : PhD student
E-mail : jean-christian.ralaivao@eric.univ-lyon2.fr
Web site : <http://eric.univ-lyon2.fr/~jcralaivao/>
Birth Date : 03/07/1966
Arrival Date : 01/11/2003
Research supervisor : Jérôme Darmont

Research topics

Within decision processes, data warehousing technologies are now mature to handle simple, numerical or symbolic data. However, various sources including the Web store, contain very heterogeneous data: texts, images, sounds, videos, databases, temporal or geographical data; which may be expressed in several languages, stored in various formats, located in different places and frameworks, etc. These so-called complex data carry a lot of information and are thus interesting to include within a decision process. However, numerous issues relate to structuring, storing and querying complex data.

The aim of my PhD thesis is to address the issue of complex data warehouse performance. Several techniques do exist to optimize simple data access and storage in a warehouse. However, they cannot be applied very efficiently onto complex data. Thus, we have to define complex data warehouse models that are adapted to the nature of stored data, and to design custom performance optimization tools for these warehouses: indexing, view materialization, partitioning, clustering, buffering, etc.

Aside, using the XML language for managing data warehouses has several advantages, especially when integrating heterogeneous data. XML indeed helps represent both structure and contents. Hence, we have proposed an XML-based complex data warehouse architecture [2] that helps benefit from optimization techniques developed in the database and XML communities.

Performance optimization always needs well-defined measures and metrics. In order to identify them, we listed performance indicators for complex data warehouses. This list helped identify performance factors that are used to determine metrics.

Finally, integrating metadata and domain-related knowledge in the complex data warehouse has a positive impact on managing data complexity, especially in the process of performance optimization [1, 3].

Publications

1. J.C. Ralaivao, J. Darmont, "Knowledge and Metadata Integration for Warehousing Complex Data", *6th International Conference on Information Systems Technology and its Applications (ISTA 07)*, Kharkiv, Ukraine, May 2007; *Lecture Notes in Informatics (LNI)*, Vol. P-107, 164-175.
2. J. Darmont, O. Boussaïd, J.C. Ralaivao, K. Aouiche, "An Architecture Framework for Complex Data Warehouses", *7th International Conference on Enterprise Information Systems (ICEIS 05)*, Miami, USA, May 2005, 370-373.
3. J.C. Ralaivao, "Améliorer la performance d'un entrepôt de données complexes par l'utilisation de métadonnées et de connaissances du domaine", *2ème atelier Fouille de Données Complexes dans un processus d'extraction des connaissances, EGC 05*, Paris, Janvier 2005, 81-84.

Rashed Khalil SALEM

Current Position : PhD student
E-mail : rashed.salem@eric.univ-lyon2.fr
Web site : <http://eric.univ-lyon2.fr/~rsalem/>
Birth Date : 20/10/1981
Arrival Date : 01/11/2007



Research supervisor : Jérôme Darmont & Omar Boussaïd

Research topics

Decision-support technologies, including data warehouses and OLAP (On-Line Analytical Processing), are nowadays technologically mature. However, their complexity makes them unattractive to many companies; hence, some vendors develop simple Web-based interfaces (Lawton, 2006). Furthermore, many decision-support applications necessitate external data sources. For instance, performing competitive monitoring for a given company requires the analysis of data available only from its competitors. In this context, the Web is a tremendous source of data, and may be considered as a farming system (Hackathorn, 2000).

There is indeed a clear trend toward on-line data warehousing, which gives way to new approaches such as virtual warehousing (Belanger et al., 1999) or XML warehousing (Pokorny, 2002; Hümmel, 2003; Park et al., 2005; Boussaïd, Darmont et al., 2007). This research is backed up by new technologies such as Web services, a set of protocols and norms that help exchange data between applications over the Web (Eckert, 2005), or Active XML, a declarative framework that harnesses Web services for data integration in a peer-to-peer architecture (Abiteboul et al., 2002).

The ERIC lab is currently designing and developing a whole XML warehousing platform. In this context, the objective of this thesis is to design and include into this platform active features (Thalhammer et al., 2001) to turn it into an active XML warehouse. This work includes integrating analysis scenarios into the warehouse, automatically. Such scenarios may be based on ECA (Event, Condition, Action) rules similar to that used active databases (Dayal et al., 1995). To devise ECA rules within the OLAP framework, they may be coupled with analysis graphs. This technique shall help break up an OLAP cube with classical OLAP operators to express the targeted on-line analysis scenario. Eventually, an active XML warehouse may be viewed as a set of distributed data sources over a peer-to-peer architecture. Deploying on-line operations for this kind of warehouse shall be based on Web services (Bonifati et al., 2000).

The objective of my PhD thesis are to:

- propose an approach to integrate ECA rules into the warehousing process,
- design an algebra for analysis graphs,
- define a framework to handle problems linked to different analysis scenarios,
- propose a Web service-based architecture for automatic, Web-based, on-line analyses.

Publications

- Rashed Khalil, Wail Elkilani, Nabil Ismail, Mohie Hadhoud, "A Cost Efficient Location Management Technique for Mobile Users with Frequently Visited Locations ", Proceedings of the 4th Annual Communication Networks and Services Research Conference (CNSR'06) - Volume 00, p.p. 259 - 266
- Rashed Khalil, Wail Elkilani, Nabil Ismail, Mohie Hadhoud, "A Cost Efficient Location Management Technique using Replicated Database ", INFOS2006, March 2006, Cairo, Egypt.

Anna STAVRIANOU

Current Position : PhD student
E-mail : Anna.Stavrianou@univ-lyon2.fr
Web site : -
Birth Date : 23/02/1977
Arrival Date : 01/10/2005
Research supervisor : Jean-Hugues CHAUCHAT



Research topics

Text mining is an interdisciplinary field that combines techniques and methodologies from various areas such as information extraction, information retrieval, computational linguistics and categorization. In our work, we concentrate on the semantic rather than the statistical techniques since it seems that the statistics alone are not sufficient for the mining of a text. More specifically, our objective is to make an initial step in combining text mining and database methodologies for the purpose of categorizing and retrieving knowledge from text.

We base our work on the LIMBO [1] algorithm which is a hierarchical clustering algorithm for categorical data, based on the Information Bottleneck framework. It has been used to cluster both tuples and categorical attribute values, discover duplication in a set of tuples and identify structures in data that may contain erroneous information or duplicates. Within LIMBO, the similarity between the values of the same attribute is measured on the basis of the distribution they induce on the remaining attributes. The semantics of the attribute values are not taken into account. In this work, our goal is to identify the similarities between the values of each tuple attribute and feed this semantic information into LIMBO in order to perform clustering of the tuples. A comparison between the clustering results while using or not the semantic information provided, will allow us to identify whether semantics can benefit the clustering task or not.

For the purpose of incorporating semantic knowledge into the tuple representation, our objective becomes two-fold: a) find the semantic relations among the values of a particular attribute and b) use these relations in order to re-distribute the weights in each tuple. A Java application has been implemented called “SemanticLIMBO” in order to allow for the application of various semantic measures on the values of an attribute. The available measures include those proposed by Seco et al. [2] and the measures that appear in the WordNet-Similarity package [3].

References

Andritsos, P., Tsaparas, P., Miller R.J., Sevcik K.C. 2004. LIMBO: Scalable Clustering of Categorical Data. In *9th International Conference on Extending Database Technology (EDBT)*, March 2004.
Seco, N., Veale, T., and Hayes, J. 2004. An intrinsic information content metric for semantic similarity in WordNet. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*, Valencia, Spain, 1089-1090.
WordNet-Similarity. <http://www.d.umn.edu/~tpederse/similarity.html>

Publications

Stavrianou, A., Andritsos, P., and Nicoloyannis, N. Overview and Semantic Issues of Text Mining. In *SIGMOD Record*, 36(3), September 2007, 23-34.

Julien THOMAS

Current Position : PhD student
E-mail : jthomas@fenics-sas.com
Web site :
Birth Date : 02/23/1982
Arrival Date : 09/19/2005
Research supervisor : D. Zighed (N. Nicoloyannis)



Research topics

Measure for supervised learning models assessment, taking into account user needs specificities and working well on imbalanced datasets. (PRAGMA : Precision and RecAll Guided Model Assessment)

Adaptive sampling strategy for imbalanced datasets and random forest. (FUNSS : Fitting User Needs Sampling Strategy)

Association rules base fuzzification.

Features construction and reduction of high dimensional space using association rules fuzzyfication.

Evolutionary features space for random forest. (G2S : Gradual Shaping Space)

Supervised visual and interactive clustering.

Search of similarity between objects using random forest.

Publications

J. Thomas, S. Marcellin "Fouille de bases d'images mammographiques", Groupe de Travail sur la Fouille de Données Complexes, Lyon, France, Septembre 2005.

A. Brémond, P.E. Jouve, J. Thomas, J. Clech, D.A. Zighed "Résultats préliminaires d'une étude comparative de deux CAD", Innovations Technologiques et bonnes pratiques en sénologie, Congrès de la SOciété Française de Mastologie et d'Imagerie du Sein (SOFMIS 06), Clermont-Ferrand, France, Mai 2006; pp 92-94.

J. Thomas, P.E. Jouve, N. Nicoloyannis "Optimisation and evaluation of random forests for imbalanced datasets", 16th International Symposium on Methodologies for Intelligent Systems (ISMIS 06), Bari, Italy, September 2006; Springer LNAI, Vol. 4203, pp 642-651.

J. Thomas, P.E. Jouve, N. Nicoloyannis "Mesure non symétrique pour l'évaluation de modèles, utilisation pour les jeux de données déséquilibrés", Extraction et Gestion des Connaissances (EGC 07), Namur, Belgique, Janvier 2007; Cepadues RNTI, Vol E-9.

J. Thomas, P.E. Jouve, N. Nicoloyannis "Asymmetric measure for supervised learning models assessment, application to breast cancer detection", International Conference on Industrial Engineering and Systems Management (IESM 07), Beijing, China, May 2007.

Julien VELCIN

Current Position : Assistant professor
E-mail : Julien.Velcin@univ-lyon2.fr
Web site : <http://eric.univ-lyon2.fr/~jvelcin/>
Birth Date : 09/03/1978
Arrival Date : 01/11/2007



Administrative Charges : Joint manager of the strand ECD in Master IDS (Business Intelligence and Statistic)

Research topics

As a whole, my researches deal with artificial intelligence, machine learning and data mining. More precisely, I'm working on concept extraction from symbolic and sparse datasets. This task is done in a non-supervised way, task which is known as *conceptual clustering*, as defined by Michalski, Diday et al. A lot of applications can be addressed like online news analysis, technological survey and database summary. I also take into account the relationship of my work with other areas, such as psychology and sociology.


My current work is on topic extraction from binary, sparse and high-dimensional datasets. I use an optimization approach and especially the meta-heuristic of tabu search defined by Glover and Laguna in order to go through this very combinatorial search space. The preliminary results I obtained, both on artificial datasets and on real news sources, are really promising and lead to publications at the MLDM and ADMA international conferences. This work is done in collaboration with sociologists from the EHESS in Paris who are studying press content and controversies through the media. I'm also working on non-supervised learning evaluation, both with a theoretical point of view and considering a pragmatic approach. Hence, a clustering evaluation software was implemented and presented at EGC in 2007.

Publications

VELCIN, J. and GANASCIA, J.-G.. Default Clustering with Conceptual Structures. In *Journal on Data Semantics VIII*, LNCS 4380, Springer-verlag (2007), p. 1-25.
VELCIN, J. and GANASCIA, J.-G.. Topic Extraction with AGAPE. In: *Proceedings of the International Conference on Advanced Data Mining and Applications* (ADMA 2007). GANASCIA, J.-G. and VELCIN, J.. Modeling Stereotype Construction with Artificial Intelligence. *Annual scientific meeting of the International Society of Political Psychology (ISPP)*. Portland, Oregon, USA (2007).
VELCIN, J. and GANASCIA, J.-G.. Stereotype Extraction with Default Clustering. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. Edinburgh, Scotland (2005).
VELCIN, J. and GANASCIA, J.-G.. Modeling default induction with conceptual structures. In *ER 2004 Conference Proceedings*. Lu, Atzeni, Chu, Zhou, and Ling editors. Springer-Verlag. Shanghai, China (2004).

Scientific activities and valorisation	
Scientific programs and/or industrial collaborations	Project “metadata extraction from textual data using machine learning techniques”, in collaboration with the LIP6 (Paris 6 University) and Alcatel-Lucent.

Jacques VIALLANEIX

Current Position :	Associate Professor	
E-mail :	jacques.viallaneix@univ-lyon2.fr	
Web site :	http://eric.univ-lyon2.fr	
Birth Date :	06/07/1963	
Arrival Date :	01/09/1993	
Administrative and Pedagogic Charges :	<p>Within the Faculty of Sociology and Anthropology of the University Lyon 2 :</p> <p>In charge of teaching in data processing for the 1st years of Bachelor of Sociology and of Bachelor of Anthropology until in 2004 (representing on average 450 hours TD per year) ;</p> <p>In charge of teaching in Data processing for the 2nd year of Bachelor of Sociology and of Bachelor of Anthropology until in 2005 (representing on average 400 hours TD per year) ; Co-responsible since 2005 ;</p> <p>In charge of the 2nd year of the Course Bachelor of MISASHS (Mathematics, Computer Science and Applied Statistics for Humanities and Social Sciences) (representing on average 470 hours per year) ;</p> <p>In charge of the 3rd year of the Course Bachelor of MISASHS (representing on average 280 hours per year) ;</p> <p>Being added to on average 300 hours per year of personal teaching, these responsibilities are heavy : ranging from the development (and update) of teaching contents until recruitment of professors or lecturers, adding the management of computer rooms, organizing schedules, and so on ;</p> <p>Member of the recruitment committee in mathematics and computer science of the university Lyon 2 (CSE 26-27-61) ;</p> <p>Member of the recruitment committee in computer science of INSA de Lyon (CSE 27-61).</p>	
Research topics	Because of my teaching and administrative charges, I unfortunately cannot currently lead a substantial research activity.	

Zhihua WEI

Current Position : PhD student
E-mail : zhihua.wei@univ-lyon2.fr
Web site :
Birth Date : 22/01/1979
Arrival Date : 01/12/2006



Research supervisor : Jean-Hugues CHAUCHAT

Research topics

My research area is Chinese text mining, including Chinese text categorization and extracting knowledge from texts based on statistical learning and natural language understanding.

My research objectives include:

1. Text representation methods which are based on bag-of-words, n-grams, keywords or noun phrases and verb phrases. Besides n-grams, the other methods are all based on natural language analysis. Different from most Latin languages, there is no delimiter between two characters in Chinese texts. As a result, most of researches based on the text content need the process of word segmentation as prerequisite. Disambiguation and recognition of unknown words are the most difficult in this process.
2. Text feature selection methods which include improving the traditional selection algorithm and exploring better methods for measuring the dissimilarities among texts in different classes.
3. Multi-class and multi-label classification methods which mainly aim to solve the classification problem in some large corpora. My work is exploring the methods to improve the performance of classifier in the complex multi-class and multi-label conditions and decrease the effects from unbalanced distribution among real corpora.

Publications

1. Book chapter: "*Chinese Language Understanding Algorithms and Applications*" Duoqian Miao, Zhihua WEI, 2007 by Tsinghua University Press (in China).
2. *A New Structure-based Bill Location Technique*, Zhihua WEI, Duoqian MIAO, Fuchun XIA, Hongyun ZHANG, Computer Application.No.10.2006.

Djamel Abdelkader ZIGHED

Current Position : Full professor,
E-mail : Abdelkader.zighed@univ-lyon2.fr
Web site : <http://morgon.univ-lyon2.fr>
Birth Date : 12/March/1955
Arrival Date : 1/Oct./1987



Administrative Charges : Head of the ERIC's Lab (1995-2002; since 2007)
President of the recruitment commission for mathematic-informatics and automatic at the university Lyon 2
Member of the scientific council of the university Lyon 2 (2007-...)
Head of the Master "Extraction des Connaissances à partir des Données (ECD)" on Data Mining (Univ. Of Lyon 2 and Univ. Of Nantes) (since 1999)
Head of the PhD program in Computer Sciences of University Lyon 2 (1995-2007)
Member of the steering committee at the faculty of economics

Research topics :

My research interests focus mainly on data mining problems, involving particularly complex data (heterogeneous, semi or unstructured, large data sets). The objective is to study the different representation spaces of the data and how the domain knowledge can be incorporated to better manage data mining tasks. This is done by proposing new machine learning algorithms that better take into account the real world applications constraints. More particularly, the current research work focuses on the following problems:

In the area of machine learning, two directions are being explored. The first one is part of the PhD thesis of Simon Marcellin. It aims to identify approaches to directly tackle the problems of mining datasets with imbalanced classes. This led us to review the properties of entropy measures and to define a new more appropriate one. The second direction is related to continuous attributes discretization. Here, we introduce re-sampling based approaches, such as the Bootstrap. These works, under investigation in Taimur Qureshi's PhD thesis, led to new algorithms for fuzzy trees. In the area of the mining complex data, two directions are also followed. The first exploits topological approaches. It uses the neighborhood graphs based models for navigating, in a more appropriate and natural way, in multimedia databases. This work, done during Hakim Hacid's PhD thesis, will be presented for defense in early 2008. The second direction, followed in Ahmad El Sayed's PhD thesis, aims to capture domain knowledge (taxonomies, ontologies) automatically from text corpora. That is, new techniques were proposed for more effective clustering on textual data, that will be used in a general framework for taxonomy learning from text. As a continuity of the work that I have conducted on the use of neighborhood graphs in machine learning, and, more generally, in data mining, a new project is being launched. The finality is to apply our results to semi supervised learning. This will help us to connect, at the same time, to other emerging works in the field of topological learning.

Publications

Berka, P., Rauch, J. and Zighed, D. A., (eds.) *Case studies in medical data mining*, Idea Group, 2008 .
Ciampi, O., Zighed, D. A. and Ritschard, G. *Prediction Trees*, Wiley, 2008 (To appear).
Zighed, D. A. "Induction Graphs for Data Mining", in Brito, P., Bertrand, P., Cucumel, G. and de Carvalho, F., ed., 'Selected Contributions in Data Analysis and Classification', Springer, 2007, pp. 419-430.
Sayed, A. E., Hacid, H. and Zighed, D. A. "A New Context-Aware Measure for Semantic Distance Using a Taxonomy and a Text Corpus" Proceedings of the IEEE International Conference on Information Reuse and

Integration, IRI 2007, 13-15 August 2007, Las Vegas, Nevada, USA', IEEE Systems, Man, and Cybernetics Society, 2007, pp. 279-284. Zighed, D. A. and Hacid, H. "Proximity graphs and separability of classes""Proceedings of the 11th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2006, Paris', IPMU, Paris, 2006, pp. 1488-1495.																																																																																																																																																																																																																																																																																																													
Scientific activities and valorisation																																																																																																																																																																																																																																																																																																													
Scientific programs and/or industrial collaborations	International Labour Organisation (BIT) and University of Geneva: Mining Expert Comments on the Application of ILO Conventions on Freedom of Association and Collective Bargaining. Breast Cancer Center Leon Berard Lyon : Computer Aided Diagnosis on mammograms Sanofi-pasteur : Improving the production of vaccines INTERREG IIIA France-Suisse INTERREG IIIA France-Suisse INTERREG IIIA France-Suisse : Study of the interdependence of markets residential property on Lake Geneva																																																																																																																																																																																																																																																																																																												
Editorial boards and program committees	<table><tr><td></td><td colspan="10">since (Year)</td></tr><tr><td></td><td>00</td><td>01</td><td>02</td><td>03</td><td>04</td><td>05</td><td>06</td><td>07</td><td>08</td></tr><tr><td>International Conference on Advanced Data Mining and Applications (ADMA)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>International Society devoted to the advancement of the theory and practice of stochastic models and data analysis techniques (ASMDA)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>joint meeting of the Société Francophone de Classification and the Classification and Data Analysis Group of the Italian Society of Statistics (CLADG-SFO)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>International Conference on Computational Statistics (COMPSTAT)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Data Warehousing and Knowledge Discovery (DaWak)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>International Conference on Discovery Science (DS)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>European Conference on Machine Learning (ECML)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>European Semantic Web Conference (ESWC)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Atelier "Fouille de Données Complexes" associé à EGC</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Flexible Query Answering Systems</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications (GfKL)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>International Association for Statistical Computing (IASC); Statistics for Data Mining, Learning and Knowledge Extraction, Satellite meeting of International Statistic Institute (IS)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>International Conference on Natural Computation (ICNC)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>International Conference on NonConvex Programming (ICN)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>International Conference Intelligent Information Systems (IIS)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>International Symposium on Methodologies for Intelligent Systems (ISMIS)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>International Semantic Web Conference</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Journées Francophones sur les Réseaux Bayésiens (JFRB)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Mining Complex Data Workshops (IEEE ICDM & PKDD/ECML)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>International Workshop on Multimedia Data Mining "Mining Integrated Media and Complex Data"</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Atelier « Mesures de similarité sémantique » associé à EGC</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Rencontre sur la Statistique Implicative et ses Applications (SA)</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Workshop on Visual Data Mining VDM@ICDM</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table>											since (Year)											00	01	02	03	04	05	06	07	08	International Conference on Advanced Data Mining and Applications (ADMA)										International Society devoted to the advancement of the theory and practice of stochastic models and data analysis techniques (ASMDA)										joint meeting of the Société Francophone de Classification and the Classification and Data Analysis Group of the Italian Society of Statistics (CLADG-SFO)										International Conference on Computational Statistics (COMPSTAT)										Data Warehousing and Knowledge Discovery (DaWak)										International Conference on Discovery Science (DS)										European Conference on Machine Learning (ECML)										Journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA)										European Semantic Web Conference (ESWC)										Atelier "Fouille de Données Complexes" associé à EGC										Flexible Query Answering Systems										International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)										Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications (GfKL)										International Association for Statistical Computing (IASC); Statistics for Data Mining, Learning and Knowledge Extraction, Satellite meeting of International Statistic Institute (IS)										International Conference on Natural Computation (ICNC)										International Conference on NonConvex Programming (ICN)										International Conference Intelligent Information Systems (IIS)										International Symposium on Methodologies for Intelligent Systems (ISMIS)										International Semantic Web Conference										Journées Francophones sur les Réseaux Bayésiens (JFRB)										Mining Complex Data Workshops (IEEE ICDM & PKDD/ECML)										International Workshop on Multimedia Data Mining "Mining Integrated Media and Complex Data"										Atelier « Mesures de similarité sémantique » associé à EGC										Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)										European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)										Rencontre sur la Statistique Implicative et ses Applications (SA)										Workshop on Visual Data Mining VDM@ICDM									
	since (Year)																																																																																																																																																																																																																																																																																																												
	00	01	02	03	04	05	06	07	08																																																																																																																																																																																																																																																																																																				
International Conference on Advanced Data Mining and Applications (ADMA)																																																																																																																																																																																																																																																																																																													
International Society devoted to the advancement of the theory and practice of stochastic models and data analysis techniques (ASMDA)																																																																																																																																																																																																																																																																																																													
joint meeting of the Société Francophone de Classification and the Classification and Data Analysis Group of the Italian Society of Statistics (CLADG-SFO)																																																																																																																																																																																																																																																																																																													
International Conference on Computational Statistics (COMPSTAT)																																																																																																																																																																																																																																																																																																													
Data Warehousing and Knowledge Discovery (DaWak)																																																																																																																																																																																																																																																																																																													
International Conference on Discovery Science (DS)																																																																																																																																																																																																																																																																																																													
European Conference on Machine Learning (ECML)																																																																																																																																																																																																																																																																																																													
Journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA)																																																																																																																																																																																																																																																																																																													
European Semantic Web Conference (ESWC)																																																																																																																																																																																																																																																																																																													
Atelier "Fouille de Données Complexes" associé à EGC																																																																																																																																																																																																																																																																																																													
Flexible Query Answering Systems																																																																																																																																																																																																																																																																																																													
International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)																																																																																																																																																																																																																																																																																																													
Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications (GfKL)																																																																																																																																																																																																																																																																																																													
International Association for Statistical Computing (IASC); Statistics for Data Mining, Learning and Knowledge Extraction, Satellite meeting of International Statistic Institute (IS)																																																																																																																																																																																																																																																																																																													
International Conference on Natural Computation (ICNC)																																																																																																																																																																																																																																																																																																													
International Conference on NonConvex Programming (ICN)																																																																																																																																																																																																																																																																																																													
International Conference Intelligent Information Systems (IIS)																																																																																																																																																																																																																																																																																																													
International Symposium on Methodologies for Intelligent Systems (ISMIS)																																																																																																																																																																																																																																																																																																													
International Semantic Web Conference																																																																																																																																																																																																																																																																																																													
Journées Francophones sur les Réseaux Bayésiens (JFRB)																																																																																																																																																																																																																																																																																																													
Mining Complex Data Workshops (IEEE ICDM & PKDD/ECML)																																																																																																																																																																																																																																																																																																													
International Workshop on Multimedia Data Mining "Mining Integrated Media and Complex Data"																																																																																																																																																																																																																																																																																																													
Atelier « Mesures de similarité sémantique » associé à EGC																																																																																																																																																																																																																																																																																																													
Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)																																																																																																																																																																																																																																																																																																													
European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)																																																																																																																																																																																																																																																																																																													
Rencontre sur la Statistique Implicative et ses Applications (SA)																																																																																																																																																																																																																																																																																																													
Workshop on Visual Data Mining VDM@ICDM																																																																																																																																																																																																																																																																																																													
Other activities	Co-founder and co-director of the journal RNTI (since 2001) President of the Association EGC “Extraction et Gestion des Connaissances” (since 2006) ; co-founder and VP of EGC (since 2001). Vice-President of SFC « Société Francophone de Classification » Member elected of the International Statistical Institute Member of the board of the European Section of (IASC) “International Association for Statistical Computational” in charge of the relationship with machine learning community.																																																																																																																																																																																																																																																																																																												

II. EDITORIAL ACTIVITIES

Here is the full list of the RNTI journals published by Cépaduès since 2004 Co-Directed by D.A. Zighed:

- Guillet, F. & Trousse, B. (*ed.*)
Extraction et gestion des connaissances (EGC'2007),
Actes des huitièmes journées Extraction et Gestion des
Connaissances, Nice, France, 29 janvier - 1er février
2008, 2 Volumes
Cépaduès-Éditions, **2008**, RNTI-E-10
- Noirhomme-Fraiture, M. & Venturini, G. (*ed.*)
Extraction et gestion des connaissances (EGC'2007),
Actes des septièmes journées Extraction et Gestion des
Connaissances, Namur, Belgique, 23-26 janvier 2007, 2
Volumes
Cépaduès-Éditions, **2007**, RNTI-E-9
- Bénani, F.; Béra, M.; Patrat, C. & Saporta, G. (*ed.*)
Data Mining et Apprentissage Statistique :
Application en Assurance, Banque et Marketing
Cépaduès-Éditions, **2007**, A-1
- Bénani, Y. & Viennet, E. (*ed.*)
Apprentissage Artificiel et Fouille de Données
Cépaduès-Éditions, **2007**, A-2
- Bellatreche, L.; Giacometti, A. & Marcel, P. (*ed.*)
Entrepôt de Données et Analyse en Ligne (3)
Cépaduès-Éditions, **2007**, B-3
- Prince, V.; Kodratoff, Y.; Azé, Jé. & Roche, M. (*ed.*)
Défi Fouille de Textes
Cépaduès-Éditions, **2007**, E-10
- Aït-Ameur, Y.; Boniol, F. & Wiels, V. (*ed.*)
Isola 2007 Workshop on Leveraging Applications of
Formal Methods, Verification and Validation
Cépaduès-Éditions, **2007**, SM-1
- Reynaud, C. & Venturini, G. (*ed.*)
Fouille du Web
Cépaduès-Éditions, **2007**, W-1
- Ritschard, G. & Djeraba, C. (*ed.*)
Extraction et gestion des connaissances (EGC'2006),
Actes des sixièmes journées Extraction et Gestion des
Connaissances, Lille, France, 17-20 janvier 2006, 2
Volumes
Cépaduès-Éditions, **2006**, RNTI-E-6
- Grigori, D.; Lopes, S.; Nguyen, B. & Zeitouni, K. (*ed.*)
Entrepôt de Données et Analyse en Ligne (2)
Cépaduès-Éditions, **2006**, B-2
- Poulet, F. & Kuntz, P. (*ed.*)
Visualisation en Extraction de Connaissances
Cépaduès-Éditions, **2006**, E-7
- Khenchaf, A. (*ed.*)
Systèmes d'Information pour l'Aide à la Décision en
Ingénierie des Systèmes
Cépaduès-Éditions, **2006**, E-8
- Pinson, S. & Vincent, N. (*ed.*)
Extraction et gestion des connaissances (EGC'2005),
Actes des cinquièmes journées Extraction et Gestion
des Connaissances, Paris, France, 18-21 janvier 2005, 2
Volumes
Cépaduès-Éditions, **2005**, RNTI-E-3
- Bentayeb, F.; Boussaid, O.; Darmont, Jé. & Rabaséda,
S. (*ed.*)
Entrepôts de Données en Ligne
Cépaduès-Éditions, **2005**, B-1
- Boussaid, O.; Gançarski, P.; Maseglier, F. & Trousse,
B. (*ed.*)
Fouille de Données complexes
Cépaduès-Éditions, **2005**, E-4
- Cloppet, F.; Pettit, J. & Vincent, N. (*ed.*)
Extraction des Connaissances : État et Perspectives
Cépaduès-Éditions, **2005**, E-5
- Hébrail, G.; Lebart, L. & Petit, J. (*ed.*)
Extraction et gestion des connaissances (EGC'2004),
Actes des quatrième journées Extraction et Gestion des
Connaissances, Clermont Ferrand, France, 20-23 janvier
2004, 2 Volumes
Cépaduès-Éditions, **2004**, RNTI-E-2
- Chavent, M. & Langlais, M. (*ed.*)
Classification et Fouille de Données
Cépaduès-Éditions, **2004**, C-1
- Briand, H. & Sebag, M. (*ed.*)
Mesures de Qualité pour la Fouille de données
Cépaduès-Éditions, **2004**, E-1
- Hacid, M.; Kodratoff, Y. & Boulanger, D. (*ed.*)
Extraction et gestion des connaissances (EGC'2003),
Actes des troisièmes journées Extraction et Gestion des
Connaissances, Lyon, France, 22-24 janvier 2003
Hermes Science Publications, **2003**, 17

Boussaid, O. & Lallich, S. (*ed.*)
Entreposage et Fouille de données
Cépaduès-Éditions, **2003**, 1

Hérin, D. & Zighed, D. A. (*ed.*)
Extraction et gestion des connaissances (EGC'2002),
Actes des deuxièmes journées Extraction et Gestion des
Connaissances, Montpellier, France, 21-23 janvier 2002

Hermes Science Publications, **2002**, 1

Briand, H. & Guillet, F. (*ed.*)
Extraction et gestion des connaissances (EGC'2001),
Actes des premières journées Extraction et Gestion des
Connaissances, Nantes, France, 17-19 janvier 2001
Hermes Science Publications, **2001**, 1

III. ORGANISATION OF SCIENTIFIC EVENTS

a. Conferences, Workshops and working research groups

EGC : The conference "Extraction and Knowledge Management (EGC)" aims to bring together researchers from disciplines of information technology such as Knowledge Discovery from Databases, Knowledge Management, Business Intelligence, etc. More specifically, it promotes exchanges between multidisciplinary communities (machine learning, statistic, Data Analysis, classification, pattern recognition, information retrieval, web semantic, knowledge engineering...). It creates synergies between academic world and companies for developing partnerships. It helps to the formation of a scientific community in the Francophone world around this dual theme of the extraction and knowledge management. Researchers at the laboratory ERIC were at the origin of the creation of the conference in 2000 and they continue, today, to lead EGC's association and the events around.

<http://www.polytech.univ-nantes.fr/associationEGC/>

EDA : The National Conference on Data Warehouse and Analysis Online has been created by researchers in the laboratory of ERIC. The aim of the francophone meetings on data Warehouses and Data Analysis online is to create and sustain a forum exclusively dedicated to this work, in order to foster interaction between researchers and users and to discuss progress of research and development experiences in this area. 3 events have already taken place. The next is scheduled in Toulouse on 5 and 6 June 2008.

EDA 2008: <http://www.irit.fr/EDA08/contact.html>

EDA 2007: <http://eda2007.sir.blois.univ-tours.fr/>

EDA 2006: <http://www.prism.uvsq.fr/~eda06/>

EDA 2005: <http://eric.univ-lyon2.fr/~eda05/>

Workshop on « Qualité des Données et des Connaissances », In association with EGC conferences

<http://conferences.enst-bretagne.fr/qdc2007/> and <http://conferences.enst-bretagne.fr/qdc2008/>.

Workshops on « Fouille de Données Complexes » in association with EGC. Five meetings have taken place.

29 Janvier 2008 : Nice, Sophia-Antipolis, EGC'08

23 Jan. 2007 : Namur, Belgique, EGC' 07

16 Jan. 2006 : Lille, EGC'06

18 Jan. 2005 : Paris, EGC'05

20 Jan. 2004 : Clermont-Ferrand, EGC'04

Workshop on « Mesure de similarité sémantique » (SimSem 2008) :

<http://www-rocq.inria.fr/axis/SimSem/AtelierEGC2008/AtelierEGC.html>

<http://eric.univ-lyon2.fr/%7Enmaiz/cmss07/>

Workshop on « les Systèmes Décisionnels » (ASD) :

<http://eric.univ-lyon2.fr/~asd/asd2008/>

It is a Franco-Maghreb workshop on decision-making systems. Its aim is to forge links between the North African researchers working in their own countries or in research laboratories abroad and french researchers. It is also an opportunity to encourage all North African doctoral students, involved in this theme, to participate to this event and make themselves known in order to create a genuine community working in the field of decision-making systems.

Co-organisation de la 6th International Conference on Flexible Query Answering Systems (FQAS 2004), 24-26 June, 2004 (Lyon).

b. Seminars of the Master ECD

2003-2004

- Jean-Paul Rasson, LIGSAT, Facultés Universitaires N-D de la Paix, Namur, Belgique, De deux méthodes de stratification avant discrimination, Jeudi 11 décembre 2003
- Alain Dussauchoy, Laboratoire PRISMA, Université Lyon 1, Un siècle de modèles de processus stochastiques appliqués aux phénomènes boursiers et autres, Jeudi 18 décembre 2003
- Amedeo Napoli, Équipe Orpailleur, LORIA Nancy, Extraction de/et connaissances, Jeudi 8 Janvier 2004
- Michel Verleysen, Université catholique de Louvain, Engineering Faculty, DICE - Microelectronics laboratory Apprentissage par réseaux de neurones: Le problème des données en grande dimension, Jeudi 29 Janvier 2004
- Jean Pierre Barthélémy, ENST Bretagne, Groupe des Ecoles de Télécommunications, Classifications binaires : une introduction, Jeudi 5 février 2004
- Christine Guinot, Unité de Biométrie et Epidémiologie, C.E.R.I.E.S., Neuilly sur Seine, France, Statistique exploratoire multidimensionnelle : Application à la recherche d'une typologie de la peau humaine saine du visage, Jeudi 26 février 2004

2004-2005

- Sylvie Philipp-Foliguet, ETIS (Equipes Traitement des Images et du Signal), CNRS UMR 8051, ENSEA, Cergy-Pontoise, France, Recherche d'images dans des bases à partir de signatures visuelles, Jeudi 15 octobre 2004

- Dragan Gamberger, Rudjer Boskovic Institute, Division of Electronics, Laboratory for Information Systems, Zagreb, Croatie, Avoiding data overfitting in scientific discovery : Experiments in functional genomics, Jeudi 25 novembre 2004
- Georges Hébrail, LTCI-UMR 5141 CNRS, Département Informatique et Réseaux, ENST Paris, Transformation de longues séries temporelles en descriptions symboliques, Jeudi 13 janvier 2005
- Gilles Venturini, Laboratoire d'Informatique, Université François Rabelais, Tours, Un survol des algorithmes biomimétiques pour la classification, Jeudi 27 janvier 05
- Christian Derquenne, R&D EDF, Clamart, France, Méthodes de fusion mises en oeuvre dans le cadre de l'enrichissement de base de données clientèle EDF, Jeudi 03 février 2005
- Lorenza Saitta, Università del Piemonte Orientale Amedeo Avogadro Dipartimento di Informatica, Complexity of Learning and Phase Transitions, Jeudi 10 février 2005

2005-2006

- Jean-Marc Petit, Laboratoire LIRIS, INSA Lyon, Recherche adaptative de bordures, Jeudi 9 février 06
- Marc Sebban, Laboratoire EURISE, Faculté des Sciences, Université de Saint-Etienne, Apprentissage non biaisé d'une distance d'édition stochastique sous la forme d'un transducteur déterministe, Jeudi 17 novembre 2005
- Jean-Michel POGGI, Laboratoire de Mathématiques, Equipe de Probabilités, Statistique et Modélisation, Université Paris-Sud Orsay, Détection de Données Aberrantes par Boosting, Jeudi 9 mars 2006

2006-2007

- Alain Morineau, Directeur de la Revue MODULAD, Préhistoire, histoire et perspectives du DM – le point-de-vue d'un statisticien, Jeudi 5 octobre 2006
- Grégoire de Lassence, Consultant Expert Académique SAS Institute, Exemples d'application de Data Mining et retour d'expérience, Jeudi 12 octobre 2006
- Michel Tenenhaus, HEC School of Management (GRECHEC), Approche PLS et analyse de tableaux multiples, Jeudi 6 octobre 2006
- Abdelaziz Faraj, Ingénieur de recherche, Institut Français du Pétrole, Sélection de modèle en régression PLS, Jeudi 9 novembre 2006
- Marc Boullé, France Telecom R&D, Spécificités du Data Mining dans les Télécoms, Jeudi 16 novembre 2006
- Abderrafih Lehman, PERTINENCE MINING Sarl, Solution de text Mining et Linguistique, Jeudi 23 novembre 2006
- Gilbert Saporta, CEDRIC, CNAM, Paris, Classification supervisée et credit scoring, Jeudi 7 décembre 2006
- Malick Paye, Biomathématicien, bioMérieux, Grenoble, Using Data Mining for Biomarker Identification, Jeudi 14 décembre 2006
- Francois Wahl, Institut Francais du Petrole, Analyse d'incertitude et de sensibilité des modèles, Jeudi 11 janvier 2007
- Christophe Roche, ERT Condillac, LISTIC, Université de Savoie, Introduction aux problématiques des ontologies : état et perspectives en recherche et en applications, Jeudi 18 janvier 2007
- Serge Muller, Ingénieur Principal General Electric, Healthcare, Technologies Applications avancées en mammographie numérique, Jeudi 1er février 2007

- Roland Marion-Gallois, Expert Consultant, Statelis, La biostatistique dans les essais cliniques, Jeudi 8 février 2007
- Alexandre Aussem, Laboratoire PRISMA, Lyon 1, Apprentissage sous contraintes de la structure des réseaux bayésiens : Applications au cancer du Nasopharynx, Jeudi 15 février 2007
- Attilio Giordana, Università del Piemonte Orientale, Dipartimento di Informatica, Modeling Complex events by means of Structured Hidden Markov Models, Jeudi 8 mars 2007
- Jean-Gabriel Ganascia, LIP6 - Université Pierre et Marie Curie (Paris VI), Apprentissage non supervisé sur des données très partiellement décrites, Jeudi 15 mars 2007
- Bertrand Chabbat, CNAF-CNEDI Lyon, L'entreprise informationnelle - Exemple : la Branche Famille de la Sécurité Sociale et les documents réglementaires, Jeudi 5 avril 2007
- Yves Lechevallier, INRIA Paris – Rocquencourt, Autour des données d'intervalles, Jeudi 26 avril 2007

2007-2008

- Jean Riondet, Directeur de l'Institut International de Formation des Cadres de Santé, IFSCS, HCL, La statistique administrative et les questionnements sociaux, de Vauban à l'INSEE, Jeudi 13 décembre 2007
- Pablo Jensen, Laboratoire IXXI, ENS Lyon, Analyser la répartition des commerces en ville, 20 décembre 2007
- Gilles Bisson, Laboratoire TIMC-IMAG, Equipe Apprentissage Modèle et Algorithmes, Grenoble, Clustering d'objets structurés, application au traitement des molécules et à celui des données de criblage haut débit, Jeudi 10 janvier 2007
- Christian Derquenne, EDF R&D, Clamart, Méthodes de fusion mises en oeuvre dans le cadre de l'enrichissement de base de données clientèle EDF, Jeudi 17 janvier 2007
- Stefan Trausan-Matu, Equipe RACAI, "POLITEHNICA" University of Bucharest, Extraction de connaissances à partir de conversation chat, Jeudi 23 janvier 2007

c. Seminars of the ERIC Lab

2003-2004

- Pierre-Alain LAUR, Recherche de structures typiques au sein d'une collection de données semi-structurées ; 13/06/2004
- Djamel Zighed, Arbogodaï : Decision tree with optimal joint partitioning, 22/03/2004
- Dan J. Smith; Construction of domain-specific digital libraries, 26/01/2004
- Ricco Rakotomalala, TANAGRA : un logiciel libre pour l'enseignement et la recherche, 15/12/2003
- Florent Massegla, Fouille de données : algorithmes et applications pour l'extraction de motifs séquentiels, 01/12/2003
- Kamel Aouiche, Utilisation des Index Bitmap pour la Fouille de Données, 17/11/2003
- Amandine Duffoux, Fouille de données à partir de la structure de documents XML, 17/11/2003
- Pierre Gañarski, L'approche multi-stratégies pour la classification non supervisée; la sélection automatique non-supervisée d'attributs pour la classification d'objets hétérogènes, 27/10/2003 à 10h00
- Didier PUZENAT, Visualisation et fouille de données, 13/10/2003

2004-2005

- Jerzy Korczak, Fouille interactive de séquences d'images IRMf, 30/05/2005
- Brice Effantin, Extraction de communautés dans le graphe du Web, 14/03/2005
- Chantal Reynaud, Comprendre le Web sémantique, 07/03/2005
- Nicole Vincent, La loi de Zipf en analyse d'images, 14/02/2005
- Sébastien Lefèvre, Introduction à la Morphologie Mathématique : principaux outils et applications 31/01/2005
- Karine Zeitouni, Entreposage et fouille de données spatiales et spatio-temporelles, 29/11/2004
- Zdenko Sonicki, Intelligent Data Analysis and Data Mining – Application in Medicine, 29/11/2004
- Edwige Fangseu Badjio, Qualité des IHM pour la fouille visuelle des données, 27/09/2004

2005-2006

- Michel Simonet, Ontologies, bases de connaissances et bases de données, 10/04/2006
- Ricco Rakotomalala, Les logiciels gratuits de DATA MINING pour l'enseignement, 12/12/2005
- Kamel Aouiche, Techniques de fouille de données pour l'optimisation automatique de performance des entrepôts de données, 28/11/2005
- Silvia Biffignandi, Shift-Share Analysis, 17/10/2005

2006-2007

- Miriam Alvariez, Plans d'expériences pour un modèle de simulation, 18/06/2007
- Rokia Missaoui, Opérateurs algébriques pour la manipulation des treillis de concepts, 23/04/2007
- Anne-Muriel Arigon, Développements d'applications pour l'identification de séquences génomiques, 12/03/2007
- Henri-Maxime Suchier, Nouvelles contributions du boosting en apprentissage automatique, 12/02/2007
- Frédéric Château, Inférence pour la Statistique Structurale, 15/01/2007
- Omar Boussaid, Evolution de l'entreposage des données complexes, 27/11/2006
- Yvan Bédard, Complexité des données géospatiales et peuplement de cubes de données : problématique, besoins et solutions, 27/11/2006
- Riadh Ben Messaoud, Couplage de l'analyse en ligne et de la fouille de données pour l'exploration, l'agrégation et l'explication des données complexes, 24/11/2006
- Jérôme Darmont, Optimisation et évaluation de performance pour l'aide à la conception et à l'administration des entrepôts de données complexes, 20/11/2006
- Djamel Zighed, Variation autour des mesures d'entropie, 16/10/2006

IV. APPLIED RESEARCH PROJECTS

Survey become apprentices of higher education in Rhône-Alpes

<i>Identifying Partners</i>	Regional Council of Rhône-Alpes Formasup Rhône-Alpes (IPRA) Rectorats of the Academy of Lyon and Grenoble
<i>Objectives of the study</i>	Design, production, use and presentation of a survey of integration of all apprentices of higher education in Rhone-Alpes.
<i>Duration and financing</i>	Financed by par Formasup and Rhône-Alpes : - 15 000 € in 2002-2003 - 3 600 € since 2004

Methods of data mining for the operation of databases CV

<i>Identifying Partners</i>	Foundation Védior Bis
<i>Objectives of the study</i>	VédiorBis Research Foundation (RVF) aims to help research laboratories working on topics that can help better characterize the supply and demand of employment. In this context, two projects were funded in the form of scholarship thesis over a period of two years each. Both projects, the second as an extension of the first, are designed to develop methods of data mining for the operation of databases CV. The work of Jérémy Clech, presented in his Ph.D. in March 2004, was dedicated to discrimination automated resume business executives. The work of Riadh Benmessaoud sought deepening of these issues for all categories of CV.
<i>Duration and financing</i>	Financed by Védior Bis foundation : Grant of € 1500 per month over two years (Jérémy Clech) 2001-2003 Grant of € 1500 per month for two years (Riadh Ben Messaoud) 2003-2005

Corpus Language Spoken in Interaction (CLAPI)

<i>Identifying Partners</i>	Call for national offer Concerted Incitative Action: Applications, Techniques, Theories (Action Concertée Incitative : Terrain, Techniques et Théories) Laboratory "ICAR" from the University of Lyon 2 and the "Ecole Supérieure Normale: Lettres et Sciences Humaines" Lyon's Laboratory "RIM (Networking, Information, Multimedia)" of the "Ecole Nationale Supérieure des Mines de Saint-Etienne".
<i>Objectives of the study</i>	Ensuring within three years the creation, management, recovery and putting online multimedia database (audio, video) gathering a corpus of oral language.
<i>Duration and financing</i>	Duration : 2002-2005 Financed by the state (ACI) : - 36 000 € - Grant to Kamel Aouiche (3 years)

Personalized Medicine anticipatory (MAP)

<i>Identifying Partners</i>	Dr Ferret, doctor of sport and holder of a project to establish firm, was hosted at the laboratory in partnership with CREALYS (incubator).
<i>Objectives of the study</i>	Extending the results and progress of empirical studies, developed for high-level athletes, to other people. This is to ensure that athletes become managers of their capital health. The work consisted in Structuring, managing and analyzing a set of complex medical data (qualitative digital texts, images ...) for a wide range of sports.
<i>Duration and financing</i>	Duration : 2003-2004 Financed by Région Rhône-Alpes and University Lyon 2 : 29000 €

Virtual Data Warehouse bank

<i>Identifying Partners</i>	Crédit Lyonnais ; Direction d'Exploitation Rhône-Alpes-Auvergne
<i>Objectives of the study</i>	The objective of this project is to provide, within a period of three years, the development of methodological tools for the management and the analysis of heterogeneous data bank. From the viewpoint of the "Credit Lyonnais", the aim is to develop a system for decision support in the area of target customers. From the point of view of the laboratory ERIC, it is to gain expertise in the field of virtual storage of heterogeneous data. This consist at on line building of data cubes and carrying out data mining analysis. All this needs an efficient process for data integration.
<i>Duration and financing</i>	Duration : 2004-2007 Financed Crédit Lyonnais : - Grant (CIFRE) for 3 years to Cécile Favre.

Citizens and users respond to changes in utilities

<i>Identifying Partners</i>	Commissariat Général au Plan, the prime Minister Service
<i>Objectives of the study</i>	Design, construction and operation of a sample survey of 1000 interviews and analysis of results. We aim understanding the expectations of French regarding the utilities (rail, posts, urban transport, electricity, gas,...). This, in a context of bouleversement of their organization, according to their experiences (user services; personal relationships with the companies, etc.).
<i>Duration and financing</i>	Duration : 2003 – 2004 Financed by Commissariat Général au Plan (French state): 43325 €

DataMining for research in pharma

<i>Identifying Partners</i>	Laboratories SERVIER
<i>Objectives of the study</i>	Analysis of data collected during the testing of drugs in Phase IV, just prior to the request for permission to placing on the market. The aim is to discover the potential side effects that could be dangerous and their causes. Do they come from the molecule tested or from its combination with other drugs or are they related to the specific pathological history of the patient?
<i>Duration and financing</i>	Duration : (one year) 2004 Financed by SERVIER' labs : 7 200 €

Multistrategy data mining

<i>Identifying Partners</i>	Under a tender National l'Agence Nationale de la Recherche "Concerted Action Incitative Masses Data." LSIIT Laboratory (Laboratoire des Sciences de l'Image, la technologie de l'information et de télédétection), University of Strasbourg I LIV Laboratory (Laboratoire Image et Ville), University of Strasbourg I.
<i>Objectives of the study</i>	The objectives of the project, associated with spatial imagery, are: first, to propose a method of interpretation assistance from a mass of data and images on the other hand, define a comprehensive Data Mining process (structuring, construction of the "objects", classification and interpretation of information) for a joint and complementary use of different sources. The latter is rarely considered in the current methods of Data Mining. The main lock is the need to use a multi-formalization at several levels of abstraction using a multi-strategy approach in the process of data mining.
<i>Duration and financing</i>	Duration : 2004-2007 Financed by the stat : 69 000 €

Intelligent System for Information Retrieval at the Use of Health (SIRIUS)

<i>Identifying Partners</i>	Council of Rhône-Alpes Region Hospital Léon Bérard Lyon
<i>Objectives of the study</i>	System for Intelligent Information Retrieval at the Use of Health (SIRIUS) will be developed and tested with users (Centre Leon Berard). The choice in the field of oncology result of both the long partnership we have with the Centre Léon Bérard and the interest shown by the Rhône-Alpes region in this area.
<i>Duration and financing</i>	Duration : 2004-2007 Financed by Council Rhône-Alpes Region : - 3700 € working grant - grant of 30444 € for 3 years allocated to Hakim HACID

Interdependence of residential real estate markets (INTERREG)

<i>Identifying Partners</i>	University of Genva
<i>Objectives of the study</i>	<p>Study on the interdependence of the residential real estate markets on Lake Geneva in the framework of the European programme INTERREG.</p> <p>This project involves the analysis of land markets, private residential rental markets and the markets for the sale of residential flats and individual houses simultaneously on different areas of the basin.</p> <p>Its objective is to improve the understanding of how the private residential property markets:</p> <ul style="list-style-type: none"> -- By visualizing the evolution of prices of goods and services over a period of thirty years, -- Observing the dynamics of these real estate markets simultaneously in the four areas of the Basin, -- Highlighting the interdependence between the real estate markets different areas, -- Creating econometric models for a prospective analysis.
<i>Duration and financing</i>	Duration : 2004-2006 Financed by the European fund INTEREG : 55000 €

Positioning on labour law

<i>Identifying Partners</i>	International Labour Office University of Geneva GIAN Foundation
<i>Objectives of the study</i>	This project aims to develop methods for searching text to study and position the labour laws of different countries. The International Labour Office (ILO) want and then draw up maps allowing representatives of various countries to position themselves relative to each other. The laboratory provides ERIC part of the draft text mining to extract the descriptive parameters of the legal corpus. It doing so, operated several hundred texts relating to labour legislation.

<i>Duration and financing</i>	Duration : 2005-2007 Financed by ILO : 12000 €
-------------------------------	---

Methods and software for the extraction of association rules

<i>Identifying Partners</i>	Laboratory Knowledge engineering at the University of Prague, Czech Republic
<i>Objectives of the study</i>	We have begun a scientific collaboration with the team of knowledge engineering from the University of Prague. The aim is to develop common platforms for data mining. ERIC bringing his experience and know-how through the platform SIPINA, the Czech team has developed a platform for extracting association rules called LispMiner.
<i>Duration and financing</i>	Duration : 2004-2006 Financed by the exchange program Franco-Czech Barrande : 6000 €

Automatic analysis of stock market prices (Tradingbots)

<i>Identifying Partners</i>	Nicolas Macherey, holder of a project to create a company, was hosted at the laboratory ERIC in partnership with the incubator CREALYS
<i>Objectives of the study</i>	Design and development of software for finance to analyze the stock market or exchange in order to make decisions automatic.
<i>Duration and financing</i>	Duration : 2007-2008 Financed by the Council of Rhône-Alpes Region : 29000 €

Generating association rules

<i>Identifying Partners</i>	SPAD
<i>Objectives of the study</i>	Implementing a module creation of association rules in the latest version 7.0 software.
<i>Duration and financing</i>	Duration : 2005 Financed by the company SPAD : 4000 €

Private PMSI

<i>Identifying Partners</i>	UMR LIRIS, University Claude Bernard Lyon 1 PRISMA, INSA of Lyon and University Claude Bernard Lyon 1
<i>Objectives of the study</i>	Methodology Analysis of the decisions referred to large databases medico-economic: the private PMSI
<i>Duration and financing</i>	Duration : 2005 Financed by the council of Rhône-Alpes Region: 25000 €

Management and interactive visualization of association rules

<i>Identifying Partners</i>	DEENOV
<i>Objectives of the study</i>	Realization of a module in a data mining software management and interactive visualization of association rules.
<i>Duration and financing</i>	Duration : 2006-2007 Financed by DEENOV : 8000 €

Marketing studies

<i>Identifying Partners</i>	DATAEXPRESSO
<i>Objectives of the study</i>	Assistance and expertise for studies in the field of marketing
<i>Duration and financing</i>	Duration : 2005-2007 Financed by DATAEXPRESSO : 25000 €

Modeling the process of vaccine manufacture of acellular pertussis

<i>Identifying Partners</i>	SANOPHI-PASTEUR
<i>Objectives of the study</i>	Development of methodologies to formalize knowledge of the processes in order to define the requirements for optimum production. Using methods of Data Mining and Knowledge engineering in order to analyze the fermentation process of Production of the vaccine "acellular pertussis"; to build a model of process control to explain the observed shifts on durations of industrial culture and to try to control this important factor in the production of vaccines.
<i>Duration and financing</i>	Duration : 2007 Financed by SANOPHI-PASTEUR : 6000 €

V. INTERNATIONAL COLLABORATIONS

University of Laval at Quebec, Canada

<i>Identifying Partners</i>	Prof. Nadir Belkhiter and Prof. Guy Mineau
<i>Collaboration in teaching</i>	We have, for many years, regular collaboration with the University of Laval in Quebec City. Professor Nadir Belkhiter. Is a regular guest at the University Lyon 2 to give master's courses in the field of search data interfaces and human-machine communication.
<i>Collaboration in research</i>	s to the expertise of Professor Belkhiter in the field of communication interfaces man-machine interface, we are developing a methodological research on the interfaces and data mining. Indeed, users of these data mining techniques are potentially very many, but these tools will only be used if they really are easy to grasp. This research aims to study the modes of interaction and visualization techniques.

University of Laval at Quebec, Canada

<i>Identifying Partners</i>	CRG Laboratory (Center for Research in Geomatics) Professor Yvan Badard
<i>Collaboration in research</i>	Warehouses spatial data assets

University of Quebec at Outaouais, Canada

<i>Identifying Partners</i>	Laboratory LARIM Professor Rokia Missaoui
<i>Collaboration in research</i>	Our collaboration on coupling OLAP - Data Mining.

University of Geneva, Switzerland

<i>Identifying Partners</i>	Professor Gilbert Ritschard
<i>Collaboration in teaching</i>	Professor G. Ritschard intervenes, since 1999, as a visiting professor in a master's course in ECD.
<i>Collaboration in research</i>	We are working with Professor G. Ritschard for many years and we have many common publications on the decision trees, the discretization of attributes, the application of text mining techniques etc.

University of Prague, Czech Republic

<i>Identifying Partners</i>	Professor Jan Rauch and professor Petr Berka
<i>Collaboration in research</i>	Development of softwares for association rules.

National School of informatics of Tunis, Tunisia

<i>Identifying Partners</i>	Ms. Hajer Bazaoui
<i>Collaboration in research</i>	Modeling and analysis of Data Marts. After building a generic data marts spatio-temporal, we are currently working on an exploratory including descriptive analyses, the OLAP and extraction of knowledge.

University of Oklahoma, Norman, USA

<i>Identifying Partners</i>	Professeur Le Gruenwald
<i>Collaboration in research</i>	A research project on the use of techniques of data mining for the self-administration of data warehouses has resulted in several joint publications (on self-indexing, mainly). We regularly send students from Master internship in the United States since 2001. We are strengthening cooperation on the project of self-administration. From a scientific point of view, this is a part of extending our approach to self-indexing for other ways of optimising performance (materialization of views, in particular) and, secondly, test different data mining techniques in this context (frequent item sets, sequential patterns, classification...) to find the most suitable for each case. Applications with complex data are planned.

National school of informatics, University of Fianarantsoa, Madagascar

<i>Identifying Partners</i>	Victor Manantsoa
<i>Collaboration in research</i>	Performances of complex data warehouse.
<i>Perspectives</i>	Develop collaboration between ENI and ERIC. Both laboratories have research subjects very close and have the desire to develop joint projects. The relationship between our two research institutions is currently provided, in large part, by Mr. Ralaivao, whose thesis work are materialized via this collaboration and is getting stronger each of his visits to the laboratory ERIC.

University of Zagreb, Croatia

<i>Identifying Partners</i>	Professor Bojana DALBELO BASIC. Supported by the Foreign Office (programme EGIDE since 2004)
<i>Collaboration in research</i>	Data mining methods applied on medical data and joint organization of the "International Workshop on Intelligent Data Analysis and Data Mining Application in Medicine" for several years. Meetings in both countries to compare our methods of data mining applied to epidemiological data.

University of Ljubljana, Slovenia

<i>Identifying Partners</i>	Professor Blaz ZUPAN Supported by the Foreign Office (programme EGIDE since 2004)
<i>Collaboration in research</i>	Data mining methods applied on medical data and joint organization of the "International Workshop on Intelligent Data Analysis and Data Mining Application in Medicine" for several years. meetings in both countries to compare our methods of data mining applied to epidemiological data.

National University of Economics, Kharkov, Ukraine

<i>Identifying Partners</i>	Professors Olexandr PUSKAR et Irina ZOLOTORIEVA
<i>Collaboration in teaching</i>	Establishing of a Franco-Ukrainian master's degree in decision-making. Financed by the Ministry of Foreign Affairs from 2005-2006.

University of Alessandria, Italy

<i>Identifying Partners</i>	Professor Lorenza Saitta
<i>Collaboration in teaching</i>	Professor Saitta Lorenza intervenes for several years as a teacher in the master EDC. Partner in the draft European master Erasmus Mundus
<i>Collaboration in research</i>	Co-supervision of the thesis of Julien Charbel

University Polytechnica of Barcelone, Espagne

<i>Identifying Partners</i>	Professor Tomas Aluja
<i>Collaboration in teaching</i>	Professor Tomas Aluja intervenes in 2007-2008 as a visiting professor in a master's course EDC.. Partner in the draft European master Erasmus Mundus

University polytechnica of Bucharest, Romania

<i>Identifying Partners</i>	Professors Eugenia Kalisz and Stefan Trausan
<i>Collaboration in teaching</i>	Professor S. Trausan intervenes in 2007-2008 as a visiting professor in a master's course ECD.. Partner in the draft European master Erasmus Mundus

