INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team AxIS

User-Centered Design, Improvement and Analysis of Information Systems

Sophia Antipolis - Méditerranée - Paris - Rocquencourt

THEME COG

Table of contents

1. Team

Head of project-team
Brigitte Trousse [CR1]

Vice-head of project-team
Yves Lechevallier [DR2]

Administrative Assistant
Stéphanie Aubin [TR]
Sophie Honnorat [AI, part-time]

Research Scientists
Thierry Despeyroux [CR1, part-time 70%]
Florent Masseglia [CR1]
Fabrice Rossi [CR1, on secondment, HdR]
Bernard Senach [CR1]
Anne-Marie Vercoustre [DR2, part-time 75%]

Partners
Mireille Arnoux [Assistant Prof., Univ. Bretagne Occidentale]
Marie-Aude Aufaure [Assistant Prof., Supélec Gif-sur-Yvette, HdR]
Marc Csernel [Assistant Prof., Univ. Paris IX Dauphine]
Doru Tanasa [Assistant Prof., Univ. Monaco]

Postdoctoral Fellows
Abdouroihamane Anli [Eiffel project, from July 1]
Celine Fiot [from October 1]
Zeina Jrad [Eiffel project]
Jovan Pehcevski [until November 30]

Ph.D. Students
Abdourahamane Balde [Univ. of Paris IX Dauphine, until April 30]
Hicham Behja [France-Morocco Cooperation (STIC-GL network), Univ. Hassan II Ben M'Sik, Casablanca, Morocco]
Alzennyr da Silva [Univ. Paris IX Dauphine]
Alice Marascu [Univ. Nice Sophia Antipolis (UNSA-STIC), until October 1]

Technical Staff
Christophe Mangeat [Research engineer, MobiVIP project, until June 30]
Mohamed Sémi Gaieb [Research engineer, Epia project, until June 30]

Visiting Scientists
Malika Charrad [PhD, ENSI, Tunisie, April 7–27]
Franciso De Carvalho [Prof., Federal Univ. of Pernambuco, Brazil, August 28–September 12]
Kelly Patricia Da Silva [Student Intern, Federal Univ. of Pernambuco, September 4–26]
James Thom [Associate Prof., RMIT, Australia, invited, May 5–August 6]

Student Interns
Abdelmoujib Elkhoumri [ENSIAS Rabat, since July 23]
Reda Kabbaj [Faculté des Sciences Sidi Mohamed Ben Abdellah, Fès, Morocco, until January 25]
Cyrille Maurice [FUNDP Namur, Belgium, until January 1]
Bashar Saleh [Univ. Nice Sophia-Antipolis, until June 20]

2. Overall Objectives

2.1. Objectives

Keywords: *KDD*, *Semantic Web*, *Semantic Web mining*, *Web mining*, *World Wide Web*, *data mining*, *data stream mining*, *document mining*, *entity ranking semantics checking*, *information retrieval*, *information system*, *information system evaluation*, *information system validation*, *knowledge discovery*, *knowledge management*, *ontology extraction*, *ontology management*, *recommender system*, *tourism*, *transportation system*, *usage mining*, *user-centered design*.

AxIS is carrying out research in the area of Information Systems (ISs) with a special interest in evolving ISs such as Web based-information Systems. Our goal is to improve the overall quality of ISs, to support designers during the design process and to ensure ease of use to end users. Since ISs are in constant evolution, it is necessary to anticipate the usage and the maintenance very early in the design process. Four main applicative objectives are addressed by the team:

- supporting the design, validation/evaluation, maintenance, of evolving ISs (cf. section 3.2);
- developing methods and tools to support both usage analysis (cf. section 3.3) and information retrieval (cf. sections 3.4);
- developing methods and tools to facilitate ISs improvement or the re-design by cross-checking content &structure analysis with usage analysis;
- and finally, supporting knowledge management in designing and evaluating ISs in order to facilitate the reuse of past experiences.

To achieve such objectives in July 2003, we set up a multidisciplinary team (from Artificial Intelligence, Data Mining & Analysis, Software Engineering, Document management) and recently from Ergonomics, in the world of information systems.

The research topics related to our objectives are presented in Figure 1 according to three points of view:

- the structure and content point of view, related to the design and the evaluation of the “static” aspects of ISs (structure, documents, ontologies),
- the usage point of view, related to the dynamic aspects of ISs i.e. both the design of support tools (information retrieval, recommender systems) and the analysis of its ”dynamic use” (usage mining).
- the knowledge management point of view, i.e. the capitalization of knowledge and experience in the evaluation process of IS: expertise in combining the evaluation results according to different points of view

3. Scientific Foundations

3.1. Introduction

This section details several research questions we want to address:

- How to support the semantics specification and the design of hypertext information systems (cf. section 3.2) ?
- How to evaluate information systems by applying KDD techniques on usage data (cf. section 3.3.1) ?
- How to synthesize and exhibit information by applying KDD techniques on documents (cf. section 3.3.2) ?
- How to support users in their information retrieval task and how to design information systems supporting the evolution of user practices (cf. section 3.4) ?

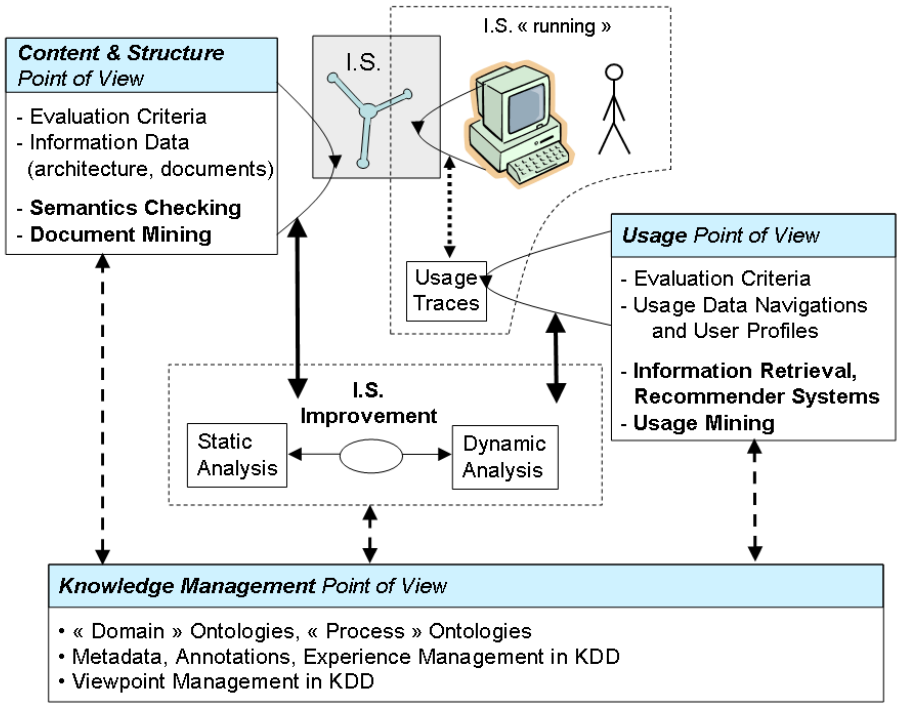

Figure 1. Global View of AxIS Research Topics

3.2. Semantics and Design of Document-Based Information Systems

Keywords: *formal semantics*, *information system design*, *semantic Web*, *semantics*, *semantics checking*.

Designing and maintaining document-based information systems, such as Web sites, is a real challenge. On the Web, it is more common to find inconsistent pieces of information than a well structured site. Our goal is to study and build tools to support the design, development and maintenance of complex but coherent sites. Our approach involves Software Engineering and Artificial Intelligence techniques. There is strong similarities between structured documents (such as Web sites) and programs; the Web is a good candidate to experiment with some of the technologies that have been developed in software engineering.

Most of the efforts deployed in the Web domain are related to languages for documents presentation (HTML, CSS, XSL) and structure (XML) and to Web sites modelling and Web services (UML). Very little has been done on Web sites formal semantics to support their quality and evolution. The initiative led by the W3C consortium on Semantic Web (XML, RDF, RDF Schema) and ontologies aims at a different objective related to resource discovery.

The term "semantics" has at least two significations: a) the meaning of words and texts, and b) the study of propositions in a deductive theory.

To address the first definition of the word semantics, we use taggers, thesaurus, ontologies, in order to add some semantics to plain texts. However we are especially interested with the latter definition, trying to give a formal semantics to Web sites.

We distinguish between the static aspects of a site that may involve a set of global constraints (not only syntactic but also semantic and context dependent) to be verified, and the dynamic aspects. Dynamic aspects formalize the Web site navigation that also needs to be specified and validated (cf. the execution of a program).

Our approach is related to the Semantic Web but yet different. The main goal of the Semantic Web is to ease computer-based information retrieval, formalizing data that is mostly textual, for further discovery. We are mostly concerned with Web sites design and production, taking into account their semantics, development and evolution. In this respect we are closer to what is called *content management* and we would like to insure that a particular Web site does follow a predefined specification. We use approaches and techniques based on logic programming and formal semantics of programming languages, in particular operational semantics.

3.3. Information Systems Data Mining

Keywords: *content mining*, *data mining*, *data warehouse*, *document mining*, *semantic data mining*, *semantic web mining*, *sematic usage mining*, *structure mining*, *usage mining*, *user behaviour*.

3.3.1. Usage Mining

There are two main motivations for usage mining in the context of ISs or search engines:

- supporting the re-design process of ISs or search engines by better understanding the user practices and by comparing the IS structure with usage analysis results;
- supporting information retrieval by reusing user groups'practices, what is called "collaborative filtering," via the design of adaptive recommender systems or ISs (cf. section 3.4).

Usage mining corresponds to data mining (or more generally KDD) applied to usage data. By usage data, we mean the traces of user behaviours in log files.
Let us consider the KDD process represented by Fig.2.

This process involves four main steps:

1. **data selection** aims at extraction, from the database or data warehouse, of the information required by the data mining step.
2. **data transformation** will then use parsers in order to create data tables usable by the data mining algorithms.
3. **data mining** techniques ranging from sequential patterns to association rules or cluster discovery.
4. finally the last step will support **re-using** previous results into an usage **analysis process**.

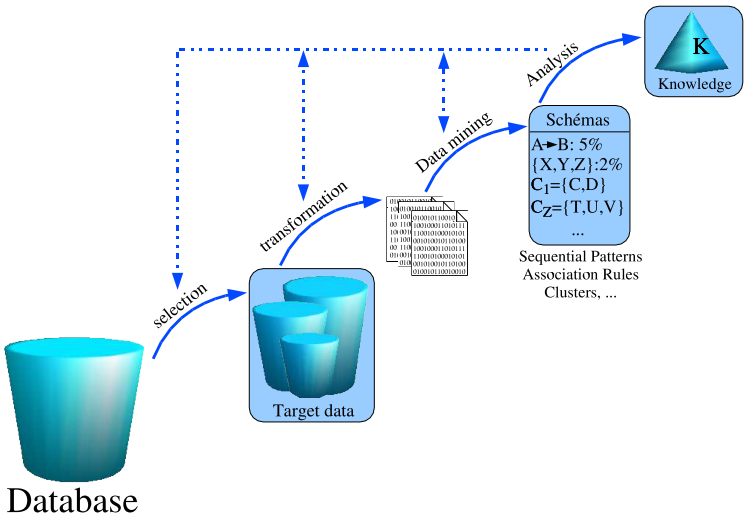

Figure 2. Steps of the KDD Process

More precisely the first three steps involve five important research directions:

3.3.1.1. Data selection and transformation

We insist on the importance of the pre-processing step in the KDD process. This step can be decomposed into selection and transformation sub-steps.

The considered KDD methods applied on usage data rely on the notion of user session, represented through a tabular model (items), an association rules model (itemsets) or a graph model. This notion of session enables us to act at the appropriate level in the knowledge extraction process from log files. Our goal is to build summaries and generate statistics on these summaries. At this level of formalization we can consider rules and graphs, define hierarchical structures on variables, extract sequences and thus build new types of data by using KDD methods.

Then, as the analysis methods come from various research fields (data analysis, statistics, data mining, AI., etc.), data transformations may be required and will be managed by appropriate parsers. Input data will come from intermediary databases, standard formatted files (XML) or a private format.

3.3.1.2. Data mining: extracting association rules

Our preprocessing tools (or generalization operators) introduced in the previous paragraph were designed to build summaries and to generate statistics on these summaries. At this level of formalization we can consider rules and graphs, define hierarchical structures on variables, extract sequences and thus build new types of data by using methods for extracting frequent itemsets or association rules.

These methods were first proposed in 1993 by R. Agrawal, T. Imielinski and A. Swami (researchers in databases at the IBM research center, Almaden). They are available in market software for data mining (IBM's intelligent miner or SAS's enterprise miner).

Our approach will rely on works from the field of generalization operators and data aggregation. These summaries can be integrated in a recommendation mechanism for helping the user. We propose to adapt frequent itemset research methods or association rules discovery methods to the Web Usage Mining problem. We may get inspired by methods coming from the genomic methods (which present common characteristics with our field). If the goal of the analysis can be written in a decisional framework then the clustering methods will identify usage groups based on the extracted rules.

3.3.1.3. Data mining: discovering sequential patterns

Knowledge about the user can be extracted based on sequential pattern discovery (which are inter transactions patterns).

Sequential patterns offer a strong correlation with Web Usage Mining purposes (and more generally with usage analysis problems). Our goal is to provide extraction methods which are as efficient as possible, and also to improve the relevance of their results. For this purpose, we plan to improve sequential pattern extraction methods by taking into account the context where those methods are involved. This can be done:

- by analyzing the causes of sequential pattern extraction failure on large access logs. It is necessary to understand and incorporate the great variety of potential behaviours on a Web site. This variety is mainly due to the large size of the trees representing the Web sites and the very large number of combination of navigations on those sites.
- by incorporating all the available information related to usage. Taking into account several information sources in a single sequential pattern extraction process is challenging and can lead to numerous opportunities.
- finally, sequential pattern mining methods will be adapted to a new and growing domain: data streams. In fact, in many practical cases, data cannot be stored for more than a specific period of time (and possibly not at all). We need to develop good solutions for adapting data mining methods to the specific constraints related to this domain (no multiple scans over the data, no blocking actions, etc.).

3.3.1.4. Data mining: clustering approach to reduce the volume of data in data warehouses

Clustering is one of the most popular techniques in knowledge acquisition and it is applied in various fields including data mining and statistical data analysis. Clustering involves organizing a set of individuals into clusters in such a way that individuals within a given cluster have a high degree of similarity, while individuals belonging to different clusters have a high degree of dissimilarity.

The definition of 'homogeneous' cluster depends on a particular algorithm: this is indeed a simple structure, which, in the absence of prior knowledge about the multidimensional shape of the data, may be a reasonable starting point towards the discovery of richer and more complex structures

Clustering methods reduce the volume of data in data warehouses, preserving the possibility to perform needed analysis. The rapid accumulation of large databases of increasing complexity poses a number of new problems that traditional algorithms are not equipped to address. One important feature of modern data collection is the ever increasing size of a typical database: it is not so unusual to work with databases containing data from a few thousands to a few million individuals and hundreds or thousands of variables. Currently, most clustering algorithms of the traditional type are severely limited regarding the number of individuals they can comfortably handle.

Cluster analysis may be divided into hierarchical and partitioning methods. Hierarchical methods yield complete hierarchy, i.e., a nested sequence of partitions of the input data. Hierarchical methods can be agglomerative or divisive. Agglomerative methods yield a sequence of nested partitions starting with the trivial clustering in which each individual is in a unique cluster and ending with the trivial clustering in which all individuals are in the same cluster. A divisive method starts with all individuals in a single cluster and performs divisions until a stopping criterion is met. Partitioning methods aim at obtaining a partition of the set of individuals into a fixed number of clusters. These methods identify the partition that optimizes (usually locally) an adequacy criterion.

3.3.1.5. Data mining: reusing usage analysis experiences

This work aims at re-using previous analysis results in current analysis: In the short term we will start with an incremental approach to the discovery of sequential motives; in the longer term, we intend to experiment with a case-based reasoning approach Very fast algorithms able to efficiently search for dependences between attributes (e.g. research algorithms with association rules), or dependences between behaviours (research algorithms with sequential motives) within large databases already exist.

Unfortunately, even though these algorithms are very efficient, but depending on the size of the database, it can take up to several days to retrieve relevant and useful information. Furthermore, the variation of parameters available to the user requires to re-start the algorithms without taking previous results into account. Similarly, when new data is added or suppressed from the base, it is often necessary to re-start the retrieval process to maintain the extracted knowledge.

Considering the size of the handled data, it is essential to propose both an interactive (parameters variation) and incremental (data variation in the base) approach in order to rapidly meet the needs of the end user.

This problem is currently regarded as an open research problem within the framework of Data Mining; Existing solutions only provide a partial solution to the problem.

3.3.2. Content and Structure Document Mining

Keywords: *classification*, *clustering*, *document mining*.

With the increasing amount of available information, sophisticated tools for supporting users in finding useful information are needed. In addition to tools for retrieving relevant documents, there is a need for tools that synthesize and exhibit information that is not explicitly contained in the document collection, using document mining techniques. Document mining objectives also include extracting structured information from rough text.
The involved techniques are mainly clustering and classification. Our goal is to explore the possibilities of those techniques for document mining.

Classification aims at associating documents to one or several predefined categories, while the objective of clustering is to identify emerging classes that are not known in advance. Traditional approaches for document classification and clustering rely on various statistical models, and representation of documents are mostly based on bags of words.

Recently much attention has been drawn towards using the structure of XML documents to improve information retrieval, classification and clustering, and more generally information mining. In the last four years, the INEX (Initiative for the Evaluation of XML retrieval) has focused on system performance in retrieving elements of documents rather than full documents and evaluated the benefits for end users. Other works are interested in clustering large collections of documents using representations of documents that involve both the structure and the content of documents, or the structure only ([103], [116], [96], [113]).

Approaches for combining structure and text range from adding a flat representation of the structure to the classical vector space model or combining different classifiers for different tags or media, to defining a more complex structured vector model [132], possibly involving attributes and links.

When using the structure only, the objective is generally to organize large and heterogeneous collections of documents into smaller collections (clusters) that can be stored and searched more effectively. Part of the objective is to identify substructures that characterize the documents in a cluster and to build a representative of the cluster [102], possibly a schema or a DTD.

Since XML documents are represented as trees, the problem of clustering XML documents is the same as clustering trees. However, it is well known that algorithms working on trees have complexity issues. Therefore some models replace the original trees by structural summaries or s-graphs that only retain the intrinsic structure of the tree: for example, reducing a list of elements to a single element, flattening recursive structures, etc.

A common drawback of the approaches above is that they reduce documents to their intrinsic patterns (sub-patterns, or summaries) and do not take into account an important characteristic of XML documents, - the notion of a list of elements and more precisely the number of elements in those lists. While it may be fine for clustering heterogeneous collection, suppressing lists of elements may result in losing document properties that could be interesting for other types of XML mining.

3.4. Supporting Information Retrieval

Keywords: *CBR*, *KDD*, *case-based reasoning*, *collaborative filtering*, *experience management*, *hypermedia*, *indexing*, *personalization*, *recommender system*, *reuse of past experiences*, *search access*, *search engine*, *social navigation*, *user behaviour*, *user profile*.

Our research on supporting information retrieval are mainly related to personalization and interface improvement:

- use and/or construction of user profiles (cf. the projects EPIA in section 7.1.1 and Eiffel (cf. section 7.1.3);
- sophisticated interfaces (cf. in the Eiffel project in section 7.1.3);
- query interface and criteria in the context of search engines (cf. section
- collaborative filtering: see our Broadway approach for designing adaptive recommender systems (cf. section 3.4.1) on which are based most of our past and current contracts. See also our software CBR*Tools and Broadway*Tools.

3.4.1. Design of Adaptive Recommender Systems

Information retrieval support tools as recommender systems are very useful in very large information systems. The objective of a recommender system is to help system users to make their choices in a field where they have little information for sorting and evaluating the possible alternatives [127], [119], [109].

A recommender system can be divided into three basic entities (cf. Figure 3): the group of recommendations producer agents, the module of recommendation computation and the group of recommendations consumers.

A major challenge in the field of recommender systems design is the following: How to produce adaptive recommendations of high quality while minimizing the effort of producers and consumers?

Two main complementary approaches are proposed in the literature: 1) approaches based on the content and the machine learning of user profiles and 2) approaches known as a collaborative filtering based on data mining techniques. The user profile is a structure of data that describes user's topics of interest in the space of the objects which can be recommended. The user profile is a structure built in the first approach or specified by the user in the second approach.

The user profile is used either to filter available objects (content-based filtering), or to recommend a user something that satisfied previous users with a similar profile (collaborative filtering) [119].

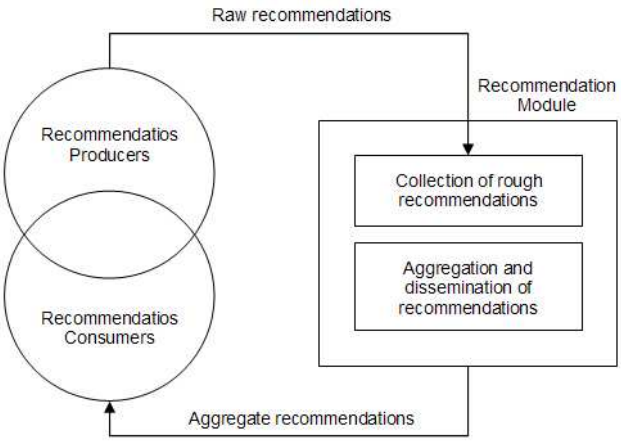

Figure 3. Architecture of a Recommender System

In the AxIS project, we continue the development of an hybrid approach for recommendations based on the analysis of visited content and on collaborative filtering; User group's past behaviours are used to compute the recommendations (collaborative filtering). This approach is able to support some usage evolutions without complete re-design. Usage analysis of the recommender system itself may be very useful to support designers in a possible re-design or improvement of their IS.

Approaches based on data mining are mainly statistical, where the sequence of events in the history is not taken into account for computing the recommendations. There are some early examples in the field of navigation assistance on the Web: the FootPrints system [130] and the system of Yan et al. [131].

The implementation challenges of our approach relate to the following aspects:

- providing techniques of identification and extraction of relevant behaviours (i.e. the learning behaviours or case behaviours) starting from raw data of past behaviours,
- defining methods and measures of similarities between behaviours,
- defining inference techniques of adaptive recommendations starting from the identified relevant past behaviours (or starting from the reminded cases).

We study the class of recommender systems, based on the re-use of a user group's past experiences, using case based reasoning techniques (CBR). **Case-Based Reasoning (CBR)** is a problem solving paradigm based on the reuse by analogy of past experiences, called "cases". In order to be found, a case is generally indexed according to certain relevant and discriminating characteristics, called "indices"; these indices determine in which situation (or context) a case can be re-used. Case-Based Reasoning [115] usually breaks up into four principal phases: retrieve, re-use, revise and retain.

Difficult problems in CBR are related to: definition and representation of a case, organization of the database containing the cases, various used indexing methods and definition of "good" similarities measurements for the case search, link between the steps research and adaptation (the best retrieved case being the most easily adaptable case), definition of an adaptation strategy starting with the found case(s), training of new indices, etc.

We focus on two types of recommender systems:

- systems where computation of recommendations is based on re-using users group's experiences in searching for information in a Web-like information system on an Internet/Intranet site. These

systems aim at providing adaptive assistance to users in their task when searching for information.

- systems where the computation of recommendations is based on the re-use of past experts'experiences, in order to assist in the design process.

We explore all the problems previously described by using case-based reasoning (CBR) techniques and more generally KDD techniques.
We pursue the evaluation of our results in CBR, in particular the indexing model by behavioral situation, the object-oriented framework CBR*Tools and toolbox Broadway*Tools in the context of our current contracts EPIA and MobiVIP (cf. section 7.1). Moreover, we pursue the study of sessions indexing techniques and plan to use some sequential pattern extraction and clustering algorithms for the on-line and off-line analysis of users'Web usage.

4. Application Domains

4.1. Panorama overview

Keywords: *Aeronautics*, *Education*, *Engineering*, *Environment*, *Health*, *Life Sciences*, *Telecommunications*, *Transportation*, *adaptive interface*, *adaptive service*, *e-CRM*, *e-business*, *e-marketing*, *information retrieval*, *personalization*, *web design*, *web usage mining*.

The project addresses any applicative field:

- on design, evaluation and improvement of huge hypermedia information systems, for which end-users are of primary concern (cf. secton 4.2).
- where knowledge management and a better understanding of use with data mining techniques could be useful (cf. Transportation and Tourism domains in sections 4.3 and 4.4).

4.2. Evolving Hypermedia Information Systems

Keywords: *Multimedia*, *Telecommunications*, *consistency verification*, *design of Information Systems*, *evaluation of Information Systems*, *ontology*, *personalization*.

We currently focus on web-based information systems (Internet, intranet), or parts of such ISs, offering one of the following characteristics:

1. presence or wanted integration of assistance in collaborative information search and personalization (ranking, filtering, addition of links, etc.);
2. a web-based IS containing information about the activities of a group of people, for instance when they are exchanging verbal or textual information.
   - a web-based IS containing information about the activities of a group of people, for example an institute (INRIA), a company, a scientific community, an European network on the Internet or intranet, etc.
   - a web-based IS indexing a wide range of productions (documents, products) from the Web or a company, according to a thematic criteria, eg. search engines (Yahoo, Voila), Internet guides for specific targets (FT Educadoc) or portals (scientific communities).
3. implicit user satisfaction (i.e. the interpretation of the user satisfaction according to the designer point of view) or explicit one, as it is the case for example for business sites, e-learning sites, and also for search engines.

The EPIA RNTL project is an example of such an evolving hypermedia information system (cf. section 7.1.1).

In summary, our fields of interest are the following:

- semantic specification and checking of an information system,
- usage analysis of an information system (Internet, intranet),
- document mining (XML documents, texts, Web pages),
- re-designing an information system based on usage analysis,
- re-designing an information system based on web mining (usage, content and structure),
- ontology construction and evolution,
- updating an ontology based on web mining (usage, content and structure),
- adaptive recommender systems for supporting information retrieval, collaborative search on the Internet,
- and in general personalization features of an Information System or a service, such as user profiling, personalized interfaces.

Ultimately, it should be noted that other fields (Life Science, Health, Transportation, etc.) could be subject to study since they provide an experimental framework for the validation of our research work in KDD, and in the reuse of experiences managing temporal data; this type of approach may be relevant in applications that are not well solved with methods in Automatics (e.g. nutrition of plants under greenhouses, control in robotics).

4.3. Transportation Systems

Several years ago we acquired experience in the design and evaluation of control rooms for transportation systems (previous work mainly with railway systems and partners such as RATP, SNCF, RTM, etc.). Presently, major evolutions in Intelligent Transportation Systems (ITS) are linked to rapid changes in communication technologies, such as ubiquitous computing, semantic web, contextual design. A strong emphasis is now put on mobility improvements. These improvements concern both the quality of traveller's information systems for trip planning and the quality of embedded services in vehicles to provide enhanced navigation aids with contextualized and personalized information.

Since 2004, The MobiVIP project (cf. section 7.1.2) has been an opportunity to collaborate with local institutions (Communauté d'Agglomération de Sophia Antipolis - CASA) and companies (VU Log) and apply AxIS' know-how in data and web mining to the field of transportation systems (cf. section 6.4.3). Cooperation about car-sharing has also been initiated this year with CASA.

4.4. Tourism

Local tourism authorities have developed Web sites in order to promote tourism and to offer services to citizens. Unfortunately the way information is organized does not necessarily meet Internet users'expectations.Mechanisms are necessary to enhance their understanding of visited sites. Tourism is a highly competitive domain. Thus if only for economical reasons, the quality and the diversity of tourism packages have to be improved, for example by highlighting cultural heritages.

AxIS is involved in the RNTL Eiffel project (cf. section 7.1.3) whose goal is to provide users with an intelligent and multilingual semantic search engine dedicated to the tourism domain. This should allow tourism operators and local territories to highlight their resources; customers could then use a specialised research tool to organize their trip on the basis of contextualised, specialised, organised and filtered information.

Other researches (cf. 6.5.2 and 6.5.3) have been carried out using log files from the city of Metz. This city was chosen because their Web site is in constant development and has been awarded several times, notably in 2003, 2004 and 2005 in the context of the Internet City label.

The objective was to extract information about tourists' behaviours from this site's log files and to identify possible benefits in designing or updating a tourism ontology [39].

AxIS is also interested in providing users whith transportation information while looking for tourism information such as cultural information, leisure etc.

5. Software

5.1. Introduction

http://www-sop.inria.fr/axis/software.html.

AxIS has developed several software or languages: 1) CLF for generating efficient parsers 2) AxISLogMIner for web usage mining (preprocessing and sequential pattern extraction), 3) a clustering toolbox used in our researches in clustering user visits, 4) CBR*Tools for Knowledge management and Broadway*Tools for designing adaptive Web-based recommendation systems. We participated also in the SODAS 2, result of a past european project and in two other softxare (Ralyx for the exploitation of INRIA activity reports and BibAdmin for the management of a collection of publications).

5.2. CLF -Computer Language Factory

Keywords: *consistency verification*, *natural semantics*, *parser*, *validation*.

Participant: Thierry Despeyroux [correspondant].

CLF is a toolbox designed to ease the development of efficient parsers in Prolog. It currently contains a couple of tools. The first one uses flex to perform lexical analysis and the second is an extension of Prolog DCGs [93], [118], [92] to perform syntactical analysis. It allows right recursion, take advantage of hash-coding of prolog clauses by modern prolog compilers and keep an automatic link to the source code to ease the development of tools as compilers with accurate error messages.

This toolbox has been used to produce a parser for XML. It has also been used to produce the specification formalism SeXML. The generated parsers have been intensively used in our team to parse and analyze XML files, mainly related to our research applied to the Inria annual activity reports.

A complete documentation is avalaible in [98].

5.3. AxISLogMiner: Preprocessing and Sequential Pattern Extraction

Keywords: *http logs*, *pre processing*, *web usage mining*.

Participants: Doru Tanasa [co-correspondant], Christophe Mangeat, Brigitte Trousse [co-correspondant].

AxISLogMiner is a software application that implements our preprocessing methodology for Web Usage Mining [128] and our work on sequential pattern extraction with low support.

We used Java to implement our application as this gives several benefits both in terms of added functionality and in terms of implementation simplicity. The application uses Perl modules for the operations carried on the log file such as: log files join, log cleaning, robot requests filtering and session/visit/episode identification. To store the preprocessed log file, in our relational model we used JDBC with Java. The result of this preprocessing is then used in data mining tool to extract, for instance, sequential patterns consisting in sequences of Web pages frequently requested by users. We endowed this software with the ability of recording the keywords employed by users in search engines to find the browsed pages.

5.4. Clustering and classification Toolbox

Participants: Marc Csernel, Alzennyr da Silva, Francesco de Carvalho, Yves Lechevallier [co-correspondant], Fabrice Rossi, Brigitte Trousse [co-correspondant].

We developed and maintained a collection of clustering and classification software, written in C++ and/or Java:

- a java library (Somlib, cf. section 6.3.2) that provides efficient implementations of several SOM variants [94], [43], [71], [70], [74], especially those that can handle dissimilarity data (available on Inria's Gforge server http://gforge.inria.fr/projects/somlib/, developed by AxIS Rocquencourt and Brieuc Conan-Guez from Université de Metz.
- a functional Multi-Layer Perceptron library, called FNET, that implements in C++ supervised classification of functional data [120], [123], [122], [121] (developed by AxIS Rocquencourt).
- two partitionning clustering methods on the dissimilarity tables issued from a collaboration between AxIS Rocquencourt team and Recife University, Brazil: CDis and CCClust [97]. Both are written in C++ and use the "Symbolic Object Language" (SOL) developed for SODAS.
- two improved and standalone versions of SODAS modules, SCluster and DIVCLUS-T [24] (AxIS Rocquencourt).
- a Java implementation of the 2-3 AHC (developed by AxIS Sophia Antipolis). The software is available as a Java applet which runs the hierarchies visualisation toolbox called the HCT for Hierarchical Clustering Toolbox (see [21]).

We developed a Web interface for the following methods: SCluster, Div, Cdis, CCClust.The interface is developed in C++ and runs on our Apache internal Web server.

5.5. CBR*Tools

Participants: Sémi Gaieb, Brigitte Trousse [correspondant].

CBR*Tools [106] is an object-oriented framework [107], [101] for Case-Based Reasoning which is specified with the UMT notation (Rational Rose) and written in Java. It offers a set of abstract classes to model the main concepts necessary to develop applications integrating case-based reasoning techniques: case, case base, index, measurements of similarity, reasoning control. It also offers a set of concrete classes which implements many traditional methods (closest neighbors indexing, Kd-tree indexing, neuronal approach based indexing, standards similarities measurements). CBR*Tools currently contains more than 240 classes divided in two main categories: the core package for basic functionality and the time package for the specific management of the behavioral situations. The programming of a new application is done by specialization of existing classes, objects aggregation or by using the parameters of the existing classes.

CBR*Tools addresses application fields where the re-use of cases indexed by behavioral situations is required. The CBR*Tools framework was evaluated via the design and the implementation of five applications (Broadway-Web, Educaid, BeCKB, Broadway-Predict, e-behaviour and Be-TRIP).

CBR*Tools is concerned by our two current contracts: EPIA (cf. section 7.1.1) and MobiVIP (cf. section 7.1.2).

CBR*Tools will be soon available for research, teaching and academic purpose under the INRIA license. The user manual can be downloaded at the URL: http://www-sop.inria.fr/axis/cbrtools/manual/.

5.6. Broadway*Tools

Participants: Sémi Gaieb, Brigitte Trousse [correspondant].

Broadway*Tools is a toolbox supporting the creation of adaptive recommendation systems on the Web or in a Internet/intranet information system. The toolbox offers different servers, including a server that computes recommendations based on the observation of the user sessions and on the re-use of user groups' former sessions. A recommender system created with Broadway*tools observes navigations of various users and gather the evaluations and annotations of those users to draw up a list of relevant recommendations (Web documents, keywords, etc).

Different recommender systems have been developed:

- for supporting Web browsing with Broadway-Web,
- for supporting browsing inside a Web-based information system with educaid (France Telecom Lannion - Inria contract), e-behaviour (Color Action, use of the mouse and eye-tracking events) and Be-TRIP (information retrieval and mobility, only specified),
- for supporting query formulation with Be-CBKB (XRCE-Inria contract), etc.

Broadway*Tools concerned our two contracts: EPIA (cf. section 7.1.1 and MobiVip (cf. section 7.1.2).

5.7. SODAS 2 Software

Participants: Yves Lechevallier [correspondant], Marc Csernel.

The SODAS 2 Software [117] is the result of the European project "ASSO" (Analysis System of Symbolic Official data), that started in January 2001 for 36 months. It supports the analysis of multidimensional complex data (numerical and non numerical) coming from databases mainly in satistical offices and administration using Symbolic Data Analysis [91].

SODAS 2 is an improved version of the SODAS software developed in the previous SODAS project, following users' requests. This new software is more operational and attractive. It proposes innovative methods and demonstrates that the underlying techniques meet the needs of statistical offices.

SODAS allows for the analysis of summarised data, called Symbolic Data. This software is now in the registration process at APP. The latest executive version (version 2.50) of the SODAS 2 software, with its user manual can be downloaded at

http://www.info.fundp.ac.be/asso/sodaslink.htm

The main contributions of AxIS to SODAS are:

1. a Symbolic Object Library (SOL [95]) that provides foundation tools, such as data loading and saving, selection, etc .
2. a divisive hierarchical clustering method on complex data tables called DIV [111]
3. a partitionning clustering method on complex data tables called SCLUST [111]
4. a supervised classification tree for symbolic data, called TREE [111]
5. a tool for extracting symbolic objects from databases [112], called DB2SO, jointly developed with EDF

Those contributions have been registered at APP.

5.8. Ralyx

Participant: Anne-Marie Vercoustre [correspondant].

In the context of her involvment with the IST department at Inria, Anne-Marie Vercoustre has been leading the Ralyx project. The goal of the Ralyx project is to publish and exploit dynamically the annual INRIA activity reports. Ralyx is based on the Xyleme system, a native XML database. Thanks to Xyleme, pages and links are no longer static but computed on the fly into different, possibly transversal views. Ralyx is operational since February 2007.

The approach works very well but may be limited by the quality of the initial data. This brings us back to one of AxIS' objectives to control and increase the quality of document-based information systems.

5.9. BibAdmin

Participant: Brigitte Trousse [correspondant].

"BibAdmin" developed by S. Chelcea. BibAdmin is a publication management tool corresponding to a collection of PHP/MySQL scripts for bibliographic (Bibtex) management over the Web. Publications are stored in a MySQL database and can be added/edited/modified via a Web interface. It is specially designed for research teams to easily manage their publications or references and to make their results more visible. Users can build different private/public bibliographies which can be then used to compile LaTeX documents. BibAdmin is made available from the end of 2005 under the GNU GPL license on INRIA's GForge server at: BibAdmin is ranked 22nd (with 1518 downloads on 1st October) among the most downloaded software on Gforge's server. For an illustration, see the publiaction server on AxIS web site.
http://gforge.inria.fr/projects/bibadmin/.
BibAdmin is used by AxIS for its Web server.

6. New Results

6.1. Introduction

This year we obtained original results in our previous four research topics: a) data transformation and knowledge management in KDD, b) data mining methods, c) Web usage and Internet mining and d) document mining and information retrieval.

First on data transformation, document validation and knowledge management in KDD (cf. section 6.2), we started new research on summarizing data streams and on a clustering approach for reducing the size of data and on a knowledge base for ontology learning. We developed a rule based language called SeXML for XML document verification based on our CLF framework (cf. section 6.2.3. We pursued also our research on feature selection (cf. section 6.2.2) and on critical edition of sanskrit texts (cf. section 6.2.6). We also studied from several years the use of metadata and ontologies (cf. the KM point of view in KDD) 1) for annotating global KDD processes in terms of viewpoints to support the management and the reuse of past KDD experiences (cf. the on-going PhD thesis of H. Behja, section 6.2.4), 2) for supporting the interpretation of extracted clusters with the definition of an ontology and an interpretation model this year (cf. the Baldé'PhD thesis defended this year [20]).

Secondly concerning data mining methods (cf. section 6.3), we published original results on a new partitioning dynamic clustering method (cf. section 6.3.1) and started research on mining solid itemsets (cf. section 6.3.7) and on extracting temporal gradual rules from sequential data. We pursued also our research on self organizing maps (cf. section 6.3.2), on functional data analysis (cf. section 6.3.3), on visualisation (cf. section 6.3.4) and on sequential pattern extraction in Data streams (cf. section 6.3.5. Let us note the defence of Chelcea'thesis on the agglomerative 2-3 Hierarchical Clustering [21].

Thirdly on information systems data mining and more precisely on usage mining, we obtained original results in the two following topics:

- construction and analysing of evolving data summaries (cf. section 6.4.1),
- mining interesting periods from Web Access Logs (cf. section 6.4.2).

Finally we pursued our research on a method based on Ergonomics and WUM for analysing a Web site (cf. section 6.4.3). We published also our work (made in 2006) on a usage mining based approach for supporting ontology evolution as a book chapter [39].

Finally on document mining and information retrieval, we started research on Entity extraction and Entity ranking, in order to validate XML-based Information Systems (IS) at a finer granularity than the one offered by the structure. We pursued our research on ontology construction from Web HTML pages (cf. section 6.5.2) and on semantic and conceptual context-aware Information Retrieval. Our work on entity extraction was published this year [56] as well as the one on scientific and technical watch [60]. For generic XML document mining, our previous work (described in our 2006 annual report) has also been published this year as a book chapter of a book on "Data Mining Patterns: New Methods and Applications" [23]. We were also very active in the context of INEX initiative involved in three tracks (cf. see section 9.1.3).

6.2. Data Transformation, Document Validation and Knowledge Management in KDD

6.2.1. Summarizing Data Streams and Clustering For Reducing the Size of Data

Keywords: *dynamic clustering algorithm*, *quantitative and qualitative data*, *unsupervised clustering*.

Participants: Alzennyr Da Silva, Yves Lechevallier.

In the data mining approach, data are not collected for a statistical analysis purpose but are available just because of the computerized nature of the management of human activities. So, goals of statistical analysis are usually not defined before the storage of data. Consequently, there is a strong need for summarizing data in order to enable future rich analysis without storing all the available data. Our work made in cooperation with Antonio Ciampi (Univ of McGill, Canada) and Georges Hébrail (ENST, Paris) differs from this work in the sense that our compression technique has a semantic basis.

Summarizing data streams

Similarly to the data mining approach, stream data are not collected for a statistical analysis purpose but are available just because of the computerized nature of the management of human activities. So, goals of statistical analyses are usually not defined before the streams to begin or before the maximum storage of data to be reached. Consequently, there is a strong need for summarizing data streams in order to enable future rich analyses without storing all the available data. Several approaches have been proposed to summarize data streams, for instance: micro-clustering techniques or sampling techniques.

Our way of getting over the problem of data stream volume is to restrict the scope of analyses by defining sliding windows on the streams. Our approach eliminates the problem of distribution drift.

We propose [30] to adapt the algorithms developed in Stephan et al. (1999) to the case of data available in the form of data streams instead of data bases.

Clustering Approach for Reducing the Size of Data

Our goal is to propose some clustering methods which reduce the volume of data in data warehouses with the possibility to perform needed analysis. Our approach is based on two key ideas [25],[33]:

A preliminary data reduction using a Kohonen Self Organizing Map (SOM) is performed. As result, the individual measurements are replaced by the means of the individual measurements over a relatively small number of micro-clusters corresponding to Kohonen neurons. The micro-clusters can now be treated as new 'cases' and the means of the original variables over micro-clusters as new variables. This 'reduced' data set is now small enough to be treated by classical clustering algorithms. A further advantage of the Kohonen reduction is that the vector of means over the micro-clusters can safely be treated as multivariate normal, owing to the central limit theorem. This is a key property, in particular because it permits the definition of an appropriate dissimilarity measure between micro-clusters.

6.2.2. Feature selection

Keywords: *Entropy*, *Feature selection*, *K Nearest Neighbor*, *Mutual information*, *Spectrometry*.

Participant: Fabrice Rossi.

Feature selection is an extremely important part in any data mining process [104]. Selecting relevant features for a predictive task (classification or regression) enables for instance specialists of the field to discover dependencies between the target variables and the input variables, that lead in turn to a better understanding of the data and of the problem. Moreover, performances of predictive models are generally higher on well chosen feature sets than on the original one, as the selection process tends to filter out irrelevant or noisy variables and reduces the effect of the curse of dimensionality.

In 2007, we have continued our work on the combination of functional data analysis and feature selection. Our previous work in this direction ([124], [125]) has been published in extended form in an international journal [38]. We have also started another approach [64] that consists in clustering the spectral variables via a simple correlation measure. The clustering is constrained to produce interval clusters, i.e., to select sub-interval of the spectral range under analysis (this can be considered as a "functional aware" variable clustering method). Each cluster of variables is replaced by a mean variable. Then, the resulting variables are processed by our previously proposed mutual information based feature selection method [126].

We have also studied approaches to automatize feature selection [29]. Our main idea is to use resampling methods to investigate the variability of an estimator of a dependency measure of two variables, e.g., the mutual information. This allows to choose automatically the parameters of the estimator by minimizing its variability. The same strategy can be used to stop a forward selection measure by estimating in a robust way the increase in the dependency measure induced by adding a new variable.

6.2.3. XML document validation

Keywords: *XML*, *consistency verification*, *natural semantics*, *validation*.

Participant: Thierry Despeyroux.

Following previous experiments [99] we develop a new methodology for XML documents verification [7] offering a rule based specific specification language (SeXML). The design of this language and all the parsers used by the system are based on CLF [98] (Computer Language Factory), a framework developed to ease parsers generation.

SeXML is based on Structural Operational Semantics (SOS) and Natural Semantics, but environments are hidden to users and are managed in an automatic manner. Basic objects are XML patterns with logical variables. As SeXML is compiled to Prolog, there is a complete access to Prolog predicates and external tools. For example Treetagger calls from SeXML have been used in other part of the project.

SeXML has also been used for structure based extraction in XML files [56].

6.2.4. Viewpoint Managment for Annotating a KDD Process

Keywords: *annotation*, *complex data mining*, *metadata*, *viewpoint*.

Participants: Hicham Behja, Brigitte Trousse.

This work was performed in the context of H. Behja's Ph.D (France-Morocco Cooperation - Software Engineering Network).

Our goal is to make explicit the notion of "viewpoint" from analysts during their activity and to propose a new approach in integrating this notion in a multi-views Knowledge Discovery from Databases (KDD) analysis.

Past years, we designed and started the implementation of an object platform (design patterns and UML using Rational Rose) for KDD integrating the definitions of viewpoints. This platform used the Weka library and contains our conceptual model integrating the "viewpoint" concept and an ontology for the KDD process. Such an ontology is composed of original components we propose for the pre-processing step and others components based on the DAMON ontology for the data mining step. For the ontology, we have used the Protégé-2000 system.
We proposed a new metadata format to annotate the KDD process. The proposed model is based on use cases to annotate the KDD process in terms of viewpoints, and on the systematic use of design patterns to comment and justify design decision. Our approach proposed object-oriented models for the KDD process, characterized by its complexity, and allows the capitalization of corporate objects for KDD. This year we pursued the implementation of the platform under Protegé-2000 and the redaction of the thesis document.

6.2.5. Knowledge Base For Ontology Learning

Keywords: *knowledge base*, *ontology acquisition*, *ontology learning*.

Participant: Marie-Aude Aufaure.

Many approaches dedicated to ontology extraction were proposed these last years. They are based on linguistic techniques (using lexico-syntactic patterns), on clustering techniques or on hybrid techniques. However no consensus has emerged for this rather difficult task. This is likely due to the fact that ontology construction relies on many dimensions such as the usage of the ontology, the expected ontology type and the actors to which this ontology is dedicated.

Knowledge extraction from web pages is a complex process starting from data cleaning until the evaluation of the extracted knowledge. The main process is web mining.

Our objective here is to propose a semi-automatic construction of ontologies from web pages [22]. To achieve such an objective, we build a knowledge base to represent web knowledge which is specified using a metaontology containing the knowledge related to the task of domain knowledge extraction. Our architecture is based on ontological components, defined by the metaontology, and related to the content, the structure and the services of a determined domain. In this architecture, we specify three interrelated ontologies: the domain ontology, the structure ontology and the services ontology [42]. Our metaontology is able to store the knowledge related to different techniques and methods for ontology construction. We have defined an on-line information retrieval system using this web knowledge architecture [41] . The on-line information retrieval system enriches the user query with domain concepts and classifies the web documents according to the concepts and the services; it also gives the user the opportunity to detect a set of services related to a given concept. A prototype has been developed and experiments have been realized in the tourism domain (cf. section 4.4).

6.2.6. Comparison of Sanskrit Texts for Critical Edition

Keywords: *Sanskrit*, *critical edition*, *distance*, *text comparison*.

Participants: Marc Csernel, Yves Lechevallier.

These results have been obtained in the context of the EuropeAid AAT project and the CNRS ACI action (cf. section 8.2.2). Our objective is to compare around 50 versions of the same text copied by hand along the centuries. During that period numerous changes in the text were introduced by the different scribes, most of the time, without meaning it.

Our aim is to obtain a critical edition of this text, i.e. an edition where all the differences between the different manuscripts are highlighted. One text is arbitrarily chosen as a reference version, and all the manuscripts are compared one by one with this reference text.

The main difficulties in doing this comparison, from an algorithmic point of view, are:

- The lack of space between the words.
- The morpho-syntactic transformation that arises, in Sanskrit, between two consecutive words without separation between. These transformations, perfectly defined by the Sanskrit grammar, are called *sandhi*.
- A number of altered manuscripts, partially destroyed by insects, mildew, rodents etc.

To address these difficulties we use a complete lemmatized reference version called *pādapāthā* (according to a special kind of recitation of Sanskrit texts) where each Sanskrit word is distinctively separated from the others by a blank or another separator. Each manuscript text (called *mātrikāpathā*) will be compared with this reference version. In the text of the *mātrikāpathā*, where few blanks occur, words facing each other are transformed according to the *sandhi*.

The expected results are expressed as an *edit distance*, but in terms of words, instead of characters as in the usual string *diff*: which are the words that have been added, deleted, replaced from the *pādapāthā* to obtain the text of the manuscript (the *mātrikāpathā*).

We first developed an HTML interface for critical edition of Sanskrit texts.

Then we focused on the comparison of Sanskrit texts, which present some difficulties because of the *sandhi*. *Pādapāthā* where *sandhi* do not aplly and *mātrikāpathā* where *sandhi* apply are not homogenous.

The comparison is made letter by letter, using the algorithm of the Longest Common Subsequence (L.C.S), as basic support, in order to determine which are the words in the *mātrikāpathā*.

This year we completed the whole process, and some new problems arise which were impossible to consider at the begining of the project, and on the other hand getting a complete chain of treatment suggested us some obvious ameliorations.

Some results have been summarized in a paper written in collaboration with François Patte (Université Paris Descartes) [45], and presented in the First International Sankrit Computational Linguistics Symposium (October 2007).

The software produces results which have been evaluated as satisfactory by some Sankrit philologists, even considering the imperfections discovered at the end of the project.

Another paper written in collaboration with Patrice Bertrand (ENST Bretagne) [45], has been presented at the "Société Francophone de Classification" congress of september 2007, and an other [26]included in a book edited by Springer.

6.3. Data Mining Methods

Keywords: *Self Organizing Map*, *complex data*, *hierarchical clustering*, *hierarchies*, *neural networks*, *symbolic data analysis*, *unsupervised clustering*.

6.3.1. Adaptive Distances in Clustering Methods

Keywords: *distances table*, *dynamic clustering algorithm*, *unsupervised clustering*.

Participants: Marc Csernel, F.A.T. de Carvalho, Yves Lechevallier.

The adaptive dynamic clustering algorithm [100] optimizes a criterion based on a fitting measure between clusters and their prototypes, but the distances used to compare clusters and their prototypes change at each iteration. These distances are not determined absolutely and can be different from one cluster to another. The advantage of these adaptive distances is that the clustering algorithm is able to recognize clusters of different shapes and sizes. The main difference between these algorithms lies in the representation step, which has two stages in the adaptive case. The first stage, where the partition and the distances are fixed and the prototypes are updated, is followed by a second one, where the partition and their corresponding prototypes are fixed and the distances are updated.

The idea of dynamical clustering with adaptive distances is to associate a distance to each cluster, which is defined according to its intra-class structure. We proposed an approach [53], which generalizes easily the dynamic cluster method for the case of the adaptive and non-adaptive non-euclidean distances.

6.3.2. Self Organizing Maps on Dissimilarity Matrices

Keywords: *clustering*, *dissimilarity*, *neural networks*, *self organizing maps*, *visualization*.

Participant: Fabrice Rossi.

In 2007, we have continued our previous work on the adaptation of the Self Organizing Map (SOM) to dissimilarity data. For the SOM based on a generalized median, we have improved our software implementation available on INRIA's GForge (cf. section 5.4) with two new methods. First, we have studied the effect of prototype collisions, i.e., when two neurons of the SOM share the same prototype [71]. Those collisions generally lead to maps of bad quality (in term of topology preservation) and should be avoided. We have proposed several strategies to prevent them. Second, we have used the branch and bound principle [110] to speed up the median SOM [43]. This solution divides by up to 2.5 the time needed to obtain the results compared to our previous work [94]. As in our previous works, results of the new algorithm are strictly identical to the ones obtained by the standard naive implementation.

We have also started to investigate a quite different approach to SOM analysis of dissimilarity data. This approach is called the "relational approach" and is based on pioneering works by Hathaway, Davenport and Bezdek [105]. The main idea is to extend a dissimilarity in order to compute dissimilarities between virtual linear combinations of the original observations and those observations. We have proposed to apply this approach to topographic processing (i.e., to SOM and to Neural Gas) in [57]. While this approach doesn't suffer from the prototype collision problems mentioned above for the median based SOM, the obtained algorithms are quite slow. We have therefore started to optimize them, especially by constraining the virtual linear combinations to have only a limited number of non zero terms [70].

Finally, we have started to work on a kernel version of the SOM, which happen to be very close to the relational version described in the previous paragraph. We have in particular obtained a best paper award at the WSOM conference 2007 with our paper on the comparison of median dissimilarity SOM and a batch version of the kernel SOM for graph analysis [74].

6.3.3. Functional Data Analysis

Keywords: *curves classification*, *functional data*, *machine learning*, *neural networks*, *support vector machines*.

Participant: Fabrice Rossi.

Functional Data Analysis is an extension of traditional data analysis to functional data. In this framework, each individual is described by one or several functions, rather than by a vector of R^n. This approach allows to take into account the regularity of the observed functions.

In 2007, we have continued our work on joining functional methods with feature selection methods. Details can be found in section 6.2.2. In summary, our main idea is to reduce the number of features submitted to a feature selection method by leveraging the functional nature of the data, either via some spline representation [38] or with a "functional aware" variable clustering method [64].

6.3.4. Visualization

Keywords: *data visualization*, *graph visualization*, *machine learning*, *metric studies*, *non linear projection*.

Participant: Fabrice Rossi.

Our work on Self Organizing Map for dissimilarity data (see Section 6.3.2) is now mature enough to enable visualization of such data. We have in particular studied hyperbolic SOM visualization of macroarray data and of Proteins, based on non Euclidean metrics [57].

6.3.5. Sequential Pattern Extraction in Data Streams: Incremental Approach

Keywords: *data streams*, *sequential patterns*.

Participants: Alice Marascu, Florent Masseglia, Yves Lechevallier.

This work was conducted in the context of A. Marascu's Ph.D study.

In recent years, emerging applications introduced new constraints for data mining methods. These constraints are mostly related to new kinds of data that can be considered as complex data. One typical such data are known as *data streams*. In data stream processing, memory usage is restricted, new elements are generated continuously and have to be considered as fast as possible, no blocking operator can be performed and the data can be examined only once. In 2006 ([10]) we have proposed a method called SMDS (Sequence Mining in Data Streams) for extracting sequential patterns from data streams. This year, our main goal was to improve the quality of the results. Actually, most data stream mining methods (including SMDS) are not able to manage the history of the knowledge in terms of content (evolution of the content of patterns). To this end, we have proposed the ICDS (Incremental Clustering in Data Streams) method [66]. ICDS is based on the algorithmic schema of SCDS and improves it by managing the evolution of the content of the patterns. To summarize this method, we cut the data stream in batches of a same size and we process the batches one by one. For the first batch, at the very beginning, we add the first sequence s_1 in a cluster c_1 and decide that the centroid of c_1 (i.e. ς_{c_1}) is equal to s_1. Then, for each other sequence s_i of the cluster, we perform the following steps:

1. Compare s_i with all clusters' centroids;
2. Find the nearest cluster c_j;
3. Add s_i to c_j;
4. Update $\varsubsetneq_{c_j}$ the centroid of cluster c_j.

The main difference with SMDS is that we don't start from scratch from one batch to another. Actually, at the end of this processing of the first batch, we keep the centroid of each cluster and use them in the processing of the next batch. The steps above are then iterated again, but the clusters from the previous batch are considered as a guide for the next batch processing.

ICDS has been tested over both real and synthetic datasets. Experiments could show the efficiency of our approach and the relevance of the extracted patterns on the Web site of Inria Sophia Antipolis. This work is the first step towards a better management of the kowledge extracted from a data stream. It allows managing the history of the content of the clusters and their evolution in time (which was not the case of SMDS). Our goal is now to propose a history management dedicated to the frequency of the extracted patterns with optimal sensitivity to the strengths of variations and awareness to the available resources.

6.3.6. Extracting Temporal Gradual Rules from Sequential Data

Keywords: *gradual rules*, *outlier detection*, *temporal data*.

Participants: Céline Fiot, Florent Masseglia.

This work aims at characterizing atypical behaviours by means of gradual data mining techniques. Our main objective is to take into account the temporal information contained within sequential data nature of the data while mining for knowledge.

First we studied existing works on atypical behaviours discovery as well as anomaly or intrusion detection using data mining approaches or gradual data mining methods. We tried to make an exhaustive and comprehensive survey. From this state of the art, we note that most approaches intended to detect atypical behaviours are based on outlier discovery by a prior clustering of the data. This clustering results are used to assess the atypicity of data, compared to the whole dataset.

Secondly this survey shows that (1) only few methods uses the temporal information that may exist within the data (Rensselaer Polytechnic Institute, New York, USA ; University of Pennsylvania, USA), (2) when atypical behaviours are observed, they are only partially explained (University of Alberta, Canada; University of Tübingen, Germany ; Mississippi State University) and (3) often not really intelligibly.

Therefore we are working on outlier detection based on sequential data clustering and on comprehensive description of atypicity using temporal gradual rules.

With this work our goal is now to provide a definition of temporal gradual rules, i.e. a new kind of gradual rules that include the temporal aspect of sequential data, for characterizing the content of the clusters (at this time, there are no propositions in the litterature for handling time in gradual rules). Such a temporal gradual rule may be for instance "The higher the number of messages having the same subject at time t, the higher the risk of mail server crash at time $t + x$".

6.3.7. Mining Solid Itemsets

Keywords: *itemsets*, *optimal window size*, *temporal itemsets*.

Participants: Bashar Saleh, Florent Masseglia.

Association rule mining algorithms aim to obtain, among a very large set of records, the frequent correlations between the items of the database. However, for many real world applications, this definition of frequent itemsets is not well adapted. Possible interesting itemsets might remain undiscovered despite their very specific characteristics. In fact, interesting itemsets are often related to the moment during which they can be observed. We may consider, for instance, the behaviors of the users on the web site of an on-line store after a special discount on recordable DVDs and CDs, advertised on TV.

In [90], we propose to find itemsets that are frequent over a contiguous subset of the database. For instance, navigations on the web page of recordable CDs and DVDs occur randomly all year, but the correlation between both items is not frequent if we consider the whole year. However, the frequency of this behavior will certainly be higher within the few hours (or days) that follow the TV spot. Therefore, the challenge is to find the time window that will optimize the support of this behavior. In other words, we want to find B, a contiguous subset of D where the support of the behavior on B is above the minimum support and the size of B is optimal. we introduced the definition of solid itemsets, which represent a coherent and compact behavior over a specific period, and we propose SIM, an algorithm for their extraction.

SIM introduces a new paradigm for the counting step of the generated candidates and extends the Generating-Pruning principle of apriori in order to generate candidate solid itemsets and count their support. The generating principle is provided with a filter on the possible intersection of the candidates (*i.e.* if two solid itemset of size k have a common prefix but do not share a common period, then they are not considered for generating a new candidate).

However, the counting step (or "pruning" in apriori) is not straightforward in our case and our goal is to build "kernels" of the candidate temporal itemsets over their period of possible frequency. Then, the kernels will be merged in order to find the corresponding solid itemsets.

Our experiments showed that SIM is able to extract the solid itemsets from very large datasets and provide useful and readable results such as behaviors corresponding to annual events, navigations on conferences Web sites or downloading of a software after the announce of a release. This work has been accepted in EGC 2008.

6.4. Web Usage Mining

6.4.1. Construction and analysis of evolving data summaries

Keywords: *dynamic clustering algorithm*, *symbolic data analysis*, *unsupervised clustering*, *web usage mining*.

Participants: Alzennyr Da Silva, Yves Lechevallier, F.A.T. de Carvalho.

The Web access patterns tend to be very dynamic in nature due not only to the dynamics of Web site content and structure, but also to changes in the user's interests. Consequently, the models associated with these patterns must be continuously updated in order to reflect the actual patterns of user access. One solution to this problem was proposed and described in [48] [51] [46] [27] [50]. The goal of our approach is to update the models using summaries obtained by means of an evolutionary approach based on clustering strategies. The approach proposed in these works consists in dividing the time period analyzed into more significant sub-periods (in our case, the months of the year) with the aim of discovering the evolution of old patterns or the emergence of new ones. After that, a clustering method is carried out on data of each sub-period, as well as over the complete period. The results provided for each clustering are then compared. We proposed four types of clustering strategies: Global clustering (performed on all existing data), Local independent clustering (performed on each time sub-period separately), Local previous clustering (performed by means of the affectation of the data in each time sub-period to the prototypes from the previous clustering) and Local dependent clustering (performed with an initialization of the clustering algorithm with the prototypes of the clusters from the previous time sub-period, the algorithm is run until it reaches the convergence). The statistical fundamentals of this approach were presented in [49] [52]. Moreover, a survey of techniques taking into account the temporal dimension in such analyses was proposed in [28].

6.4.2. Mining Interesting Periods from Web Access Logs

Keywords: *WUM*, *Web logs*, *periods*, *sequential pattern*.

Participants: Alice Marascu, Florent Masseglia.

In this work done in collaboration with M. Teisseire (LIRMM) and P. Poncelet (Ecole des Mines d'Alès), we have focused on a particular problem that has to be considered by Web Usage Mining techniques: the arbitrary division of the data which is done today. This problem was introduced in [114]. This division comes either from an arbitrary decision in order to provide one log per *x* days (*e.g.* one log per month), or from a wish to find particular behaviours (*e.g.* the behaviour of the Web site users from November 15 to December 23, during Christmas purchases).

The outline of our method [34] is the following: enumerating the sets of periods in the log that will be analyzed and then identifying which ones contain frequent sequential patterns. Our method will process the log file by considering millions of periods (each period corresponds to a sub-log). The principle of our method will be to extract frequent sequential patterns from each period. Our proposal is a heuristic-based miner, our goal is to provide a result having the following characteristics:

For each period p in the history of the log, let $realResult$ be the set of frequent behavioural patterns embedded in the navigation sequences of the users belonging to p. $realResult$ is the result to obtain (*i.e.* the result that would be exhibited by a sequential pattern mining algorithm which would explore the whole set of solutions by working on the clients of C_p).We want to find most of the sequences occurring in $realResult$ while preventing the proposed result becoming larger than it should (otherwise the set of all client navigations would be considered as a good solution, which is obviously wrong).

In the new version of this work [34], we have improved the genetic operators that are involved in our method. These operator range from the mere extension of a sequence with a frequent item to the more complex crossing of sequences. We now propose a comparison of the efficiency between those operators in a new set of experiments. We have also provided a better comparison with existing methods for mining sequential patterns. This comparison is based on the support of the extracted patterns, as well as the ability of existing method to extract some of those patterns.

In our experiments, we have extracted interesting behaviours. Those behaviours show that an analysis based on multiple division of the log (as described in this paper) allows obtaining behavioural patterns embedded in short or long periods.

6.4.3. Web site analysis based on an Ergonomic and Web usage Mining Approach

Participants: Bernard Senach, Brigitte Trousse.

In 2006 AxIS has began to set up a new method for web site evaluation, articulating usage mining approach and human factors expertise (cf. our 2006 annual report). The first study during the MobiVIP Project [84] showed that combining Ergonomic and Web usage Mining Approaches was very fruitful and we want to go further in this direction. A rapid analysis of the state of the art as shown that the two INRIA research teams which have their focus on user interface design and evaluation (In-situ and Merlin) as well as other french academic laboratories (LIC/IIHM, LIG/Multicom, IRIT/I3C) and specialised laboratories in usage analysis (Laboratoires des Usages: Marsouin, LUCE, Lutin, Lucsi, LDU of Sophia Antipolis), do not presently use data mining technologies when evaluating web sites. Due to the place at which usability evaluation methods have to be run after design changes, the international effort is mainly oriented toward automatization of the evaluation processes. Previous attemps have tried to compute web metrics (e.g. Rating Game, WebTango, etc.), to connect log files analysis and task interaction models (for instance, QUIP, KALDI) or to implement human factors expertise in knowledge bases (for instance: Sherlock, Ergoval, Synop, Ergo-conceptor). As full automatization is still often decepting, we believe much more in a cognitive coupling in which web site evaluation relies both on human ability and powerful technologies. The effort will be pursued with the FOCUS platform (see 8.1.2).

6.5. Document Mining and Information Retrieval

Keywords: *Classification*, *Clustering*, *Context-Aware Information Retrieval*, *Data Mining*, *Entity Ranking*, *INEX*, *Mining Complex data*, *XML Document*, *XML mining*, *ontology construction*.

6.5.1. Entity Ranking

Keywords: *Entity ranking*, *categories*, *linkrank*, *named entities*.

Participants: Anne-Marie Vercoustre, Jovan Pehcevski, James Thom.

The goal of *entity ranking* is to retrieve entities as answers to a query. The objective is no longer to tag the names of the entities in documents (in batch mode) but rather to return a list of the relevant entity names, and possibly a page or some description associated with each entity. We have developped a system for Entity Ranking in Wikipedia that addresses two specific tasks: a task where the category of the expected entity answers is provided; and a task where a few (two or three) examples of the expected entity answers are provided.

In our approach, candidate pages are ranked by combining three different scores: a linkrank score, a category score, and the initial search engine similarity score. The architecture of our system provides a general framework for evaluating entity ranking which allows for replacing some modules by more advanced modules and evaluate alternatives or different combinations of the score funtions [88], [129]. We also experimented with different category similarity between the category of the entity examples and the potential entity answers [72]. An evaluation module assists in tuning the parameters of the system and to globally evaluate the entity ranking approach [88], [73].

The current system has been developped in the context of the INEX (Initiative for the Evaluation of XML Retrieval) track on Entity Ranking.

6.5.2. Web HTML Pages Clustering For Ontology Construction

Keywords: *Web pages*, *clustering*, *ontology construction*.

Participant: Marie-Aude Aufaure.

We proposed an approach for ontology construction from Web pages that is based on a *contextual* and incremental clustering of terms. Our approach defines and evaluates a context-based clustering algorithm for ontology learning included in a global architecture for knowledge discovery for the semantic Web [108]. This algorithm is based on an incremental use of the partitioning K-means algorithm and is guided by a structural context. This context is based on the HTML structure and the location of words in the documents. This contextual representation guides the clustering algorithm to delimit the context of each word by improving the word weighting, the word pair's similarity and the semantically closer cooccurent selection for each word. Our algorithm refines the context of each word cluster and improves the conceptual quality of the resulting clusters and consequently of the extracted concepts [62]. This year, we have defined a set of criteria for evaluating the ontological concepts [31]. We also experiment the contextual clustering algorithm on a HTML document corpus related to the tourism domain (in French) and we evaluate the extracted ontological concepts with our contextual algorithm. The results show that the appropriate context definition and the successive refinements of clusters improve the relevance of the extracted concepts in comparison with a simple K-means algorithm. Our evaluation of ontological concepts can be applied to any domain and provides qualitative and quantitative criteria.

6.5.3. Semantic and Conceptual Context-Aware Information Retrieval

Keywords: *contextual information retrieval*, *formal concept analysis*, *semantics*.

Participant: Marie-Aude Aufaure.

In this work, we define an information retrieval methodology that uses Formal Concept Analysis in conjunction with semantics to provide contextual answers to Web queries [40]. The conceptual context defined can be global - i.e. stable- or instantaneous- i.e. bounded by the global context. Our methodology consists first in a pretreatment providing the global conceptual context and then in an online contextual processing of users requests, associated with an instantaneous context. The pretreatment consists in computing offline a conceptual lattice from data sources in order to build an overall conceptual context. Then, the information retrieval is performed in real-time: users formulate their query with terms from the thesaurus/ontology. Users may then navigate within the lattice by generalizing or on the contrary by refining their query [69]. This year, we define a similarity measure to find the closer concepts starting from an entry point of the lattice, in order to help the user to navigate (master thesis of Saoussen Sakji, University Paris-Dauphine). Our information retrieval process was illustrated through experimentations in the tourism domain. One interest of our approach is to perform a more relevant and refined information retrieval, closer to the users expectation. We add a semantic layer to the conceptual and data ones. The similarity measure helps the user to navigate through big lattices by ranking the neighbour concepts. This method is generic and can be applied to any heterogeneous data sources (Web data, personal data, etc.).

This method has several advantages:

- Results are provided according to both the context of the query and the context of available data. For example, only query refinements corresponding to existing tourism pages are proposed;
- The added semantics can be chosen depending on the target user(s);

More powerful semantics can be used, in particular ontologies. This allows enhanced query formulation and provides more relevant results.

Our information retrieval process is illustrated through experimentation results in the tourism domain. One interest of our approach is to perform a more relevant and refined information retrieval, closer to the users expectation. In this work, we strongly collaborate with Bénédicte Le Grand and Michel Soto, from LIP 6 Laboratory.

7. Contracts and Grants with Industry

7.1. Grants with Industry

7.1.1. EPIA: a RNTL Project (2003-2007)

Participants: Semi Gaieb, Yves Lechevallier, Bernard Senach, Brigitte Trousse [resp].

Inria Contract Reference: S04 AO485 00 SOPML00 1

The EPIA project “Evolution of an Adaptive Information Portal” got labeled by RNTL 2002, was started on September 2003 and completed in june 2007. Partners are Dalkia, Ever(Mediapps) and Inria.

The objectives of this pre-competitive project is to use clustering methods for usage analysis [79] and to use collaborative filtering to help intranet users to access information inside a large portal. After understanding the user needs for Net.Portal (construction tool for intranet portals), we finalized this year the specification of the NetPortal trace engine (cf. the deliverable D3 [77], [78]). The solution for the user interface lay-out and dynamics of a recommender system called Net.CanalRecommender was defined from the study of several alternatives (pull-down, menu bar, tabs) and the trade-off was chosen according to ergonomics criterias. In order set up an experiment, few usage scenarios were set up [82] and a simple one was used for a demonstration of the recommender which suggests documents selected by people in previous sessions. The system [83] combines in an original way users profiles and navigations similarities to improve documents. An evaluation plan with evaluation criterias and a query about user satisfaction have been defined but no real users were available to conduct an evaluation.

7.1.2. MobiVIP: a PREDIT Project (2004-2007)

Participants: Christophe Mangeat, Guilaume Pilot, Alex Thibau, Bernard Senach [co-resp], Brigitte Trousse [co-resp.].

Inria Contract Reference: 2 03 A2005 00 00MP5 01 1
MobiVIP, Individual Public Vehicles for Mobility in town centers, is a research project of Predit 3 (Integration of the Communication and Information systems Group). It involved five research laboratories and seven small business companies (SME), in order to experiment, show and evaluate the impact of the NTIC on a new service for mobility in town centers. This service is made up of small urban vehicles completing existing public transportation. The MobiVIP project has developped key technological bricks for the integrated deployment of mobility services in urban environment. The strengths of the project are:

1. the integration between assisted and automatic control, telecommunications, transportation modeling, evaluation of service;
2. the demonstrations on five complementary experimental sites;
3. the evaluation of possible technology transfer.

The MobiVIP Project has ended in June 2006 for most of the partners but has been extended up to June 2007 for AxIS and some others partners [87]. In this project, our team has been in charge of the evaluation module and this continuation has allowed us to process real usage data of electrical vehicules in Antibes. Our data mining and classification tools have been used to set up a structured methodology of mobility logs analysis linking usage data in a complete processing chain from vehicule reservation up to trip destination.

This year we had two main tasks: the work related to the task 5.4 and the VU Log experimentation. In task 5.4 we proposed methodological tools to analyse profiles in a mobility context and to classify trip dataset coming from surveys and from simulated vehicule usage [85]. Our classifications tools were used in relation with cartographic visualization to provide a better understanding of urban mobility. In this task we also realized a review of spatial cognition studies in order to apply their results in the design and evaluation of travellers information systems [84] In task 5.5. [86] the experimentation was conducted in collaboration with the company VU Log who provides services for urban mobility in Antibes: subscribers can use electrical vehicules for short trips around the town center. We first provided a technical assistance to the company in several pretests with users (potential customers) in order to understand their perception of the services and the way they would use them. Two analysis of real usage dataset were then conducted : the first one concerned the usage of a vocal server and showed very good technical performances of the system and a variety of usage patterns; the second focused on vehicules usage. The detailed analysis described in D5.1 (decomposition trip structure from vehicule reservation to arrival at destination) and D5.2. (evaluation criterias) were used to specified the dataset to record. Then, vizualisation tools were used in relation with classification tools to provide insight of pattern usages. The results revealed special usage patterns corresponding to specific situations knowned by VU Log team. The company has a strong demand for further analysis as it is convinced that usage mining is the key for quality in future mobility services. A complex process including several free software (QGis, Grass) in different environments (Windows, Linux) has been used to map trip data and visualization tools and a user'manual has been written [81].

7.1.3. Eiffel "E-tourism and Semantic Web": a RNTL Project (2006-2009)

Participants: Abdouroihamane Anli, Marie-Aude Aufaure, Zeina Jrad, Yves Lechevallier [resp.], Guillaume Pilot, Bernard Senach, Doru Tanasa, Brigitte Trousse.

Inria Contract Reference: 105D1499 00 21173 01 0

The EIFFEL project related to semantic web and e-Tourism was labelled in 2006 by the RNTL program and started this year. Industrial partners are Mondeca and Antidot (leadership) and academic partners are LIRMM and University of Paris X (Nanterre).

The main goal of the Eiffel project is to provide users with an intelligent and multilingual semantic search engine dedicated to the tourism domain. This solution should allow tourism operators and local territories to highlight their resources; the end users will then use a specialised research tool allowing them to organize their trip on the basis of contextualised, specialised, organised and filtered information. Queries and results will be guided by user profiles extracted from usage analysis. These profiles will facilitate the access to distributed and highly heterogeneous data. In this project, AxIS is in charge of the sub-package SP8 and will define new paradigms dedicated to knowledge searching and visualizing, and will extract and exploit users' models and profiles from web logs. A first deliverable (june 2007) concerns the user interface specification for knowledge searching [76], [59]. We are working on the user model definition to represent and store the most relevant characteristics of the user [58]. We investigate the use of ontologies in modelling not only the user's preferences but also his global context (time, place, material, history, etc). Our objective is to automatically adapt the content and structure, based on information contained in the user profile. This information is obtained either explicitly or implicitly. AxIs strongly collaborated with Bénédicte Le Grand and Michel Soto from LIP6.

7.1.4. Others actions with Industry

Two new grants with industry were accepted and will start in 2008:

- the "MIDAS" project in response to the ANR's call for proposals on "Masses de Données / Connaissances Ambiantes" has been accepted. MIDAS (MINING DATA STREAMS) will gather together 5 academic partners: AxIS, ENST (Paris), LIRMM (Montpellier), LGI2P (Nimes), GRIMAAG (Martinique); and two industrial partners: EDF R&D and France Télécom R&D.
- The "Intermed" project in resposne to the ANR's TechLOg call for proposals has been accepted. The Intermed kick-off meeting previously planned in december 2007 has been delayed to 2008. Academic partners are Cemagref (G-EAU et TETIS) LIRMM, CEPEL and industrial partners are SCRIPTAL, SIRENA, Normind, PIKKO. The aim of the InterMed project is to design and implement a substructure of tools fitting the requirements of users in charge of territory planning. The goal is to use appropriate technologies to establish a functionnal link between citizen and local authorities. The technologies we are looking for will be progressively adapted to deal with human factors and constraints of the "field". The proposed experimental approach will rely on several iterations and active participation of people involved in the discussions.

We had different industrial contacts during this year:

- VU Log, startup offering software and services dedicated to urban mobility (located in Antibes). Support for an experimentation with a small electric urban vehicule (speed limited to 45 km/h).
- Agglomeration Community of Antibes Sophia Antipolis (CASA). A previous successfull collaboration during the MobiVIP project (See Delivrable 5.3. [84] gave opportunity for further work. We had several metting with the CASA transportation team to plan a cooperation and a usage analysis of a Web site dedicated to car-sharing will probably be engaged.
- Zetoo: AxIS participated to a meeting organised by valorization board of Inria to discuss an Alcatel' project about e-commerce and social networks.
- Morocco Telecom: The exchanges with Morocco Telecom, Casablanca University and ENSAM (Meknes) about the WRUM proposal (redesign methology based on site usage mining) were extended this year and have are not yet resulted in effective cooperation.
- SAP, Sophia Antipolis related to data mining and data streams (security and environnement problems) in relation with Marascu's PhD thesis. Contact: B. Trousse and F. Masseglia
- DENSO Research: Mr Sazaki asked for a presentation of AxIS'expertise while visiting INRIA.

8. Other Grants and Activities

8.1. Regional Initiatives

Due to the bi-localization of the team, we are involved with two regions: PACA and Ile-de-France.

8.1.1. "Pôles de compétitivité"

- "Pôle de compétitivité SCS - Solutions Communicantes Sécurisées":

  AxIS (B. Trousse and B. Senach) were involved in the preparation of the project "Clic&Go". The goal of this project is to provide tools for proximity e-commerce. New technologies are used not to sell products on-line but to support new relationships between customers and shopkeepers. A better information will be given to the clients (availability of products with well-defined attributes - size, colors, etc.) and merchants will be able to develop loyalty through technologies. Usage mining will be used to improve user profiling and push suggestions according to similarity of buyers'profils. In Clic&Go project academic partners are LSIS (UMR CNRS 6168), INRIA (AxIS) and industrial partners are STID, COMLINKS and AGEVIA. The project has been labelled by "Pôle de competitivité SCS" in June and will be proposed in December to the region for a grant.
- "Pôle de compétitivité SCS - Solutions Communicantes Sécurisées": AxIS (B. Trousse and B. Senach) are involved in the preparation of a new project (AccèsCité) aiming at providing guidance for thematic urban travelling. The device will be used, in particular, by disabled people and usage mining tools will allow to improve information content delivered to users.

8.1.2. Other initiatives

- Axis (B. Trousse and B. Senach) participated with several academic laboratories (LIRMM, Cirad and Cemagref) to the founding of the METISSE working group on "Territory intelligence and sustainable development"
- CPER Télius, FOCUS plateform : last year in the framework of a grant between government and regional administration, AxIS proposed the design of a usage mining platform. The FOCUS platform, that is now granted by Region will provide to its "clients" tools and services to get a precise understanding of how people use new technologies. This year we had several meeting with academic partners (Eurecom and UNSA) in order to choose classes of application which will be studied within the plateform.
- TICE-Med : Bernard Senach went to TICE'Med colloqium in Marseille to envision with Serge Agostinelli futur collaborations (LSIS, SIC project "Sciences de l'Information et de la Communication")
- Supelec: Marie-Aude Aufaure from Supelec collaborates with others Supelec members [69].

8.2. National Initiatives

AxIS is involved in one ARC from INRIA and several national working groups.

8.2.1. ARC "SéSur": Sécurité et Surveillance dans les flots de données

SéSur (2007-2008, 1 post-doc 1 year + 49Ke) is an ARC which involves Dream (IRISA), LGI2P (Ecole des Mines d'Alès) and LIRMM (Montpellier). The goal of SéSur is to propose solutions for the security, monitoring and diagnosis of data streams. Data streams have two major characteristics: 1) they are the vital signals of the considered system and their analysis is of great interest and 2) their production rate is so high that actual technology is not able to process them in a satisfying way. We are mainly interested in monitoring the systems that produce data streams. The expected result is a monitoring system able to detect, in the stream, signals that are typical of the good or bad condition of this stream. The general framework of data streams implies extracting simultaneously and on-the-fly any pattern that indicates a dysfunction.

8.2.2. CNRS Action Concertée Incitative: "Histoire des savoirs"

This initiative (ACI RNR TTT Grammaire et mathématique dans le monde indien 17/01/03 - 17/01/06) associates several French research teams from various research fields, such as computer science, data analysis, and Sanskrit literature. The main goal of this action is to provide help for the construction of critical edition

of Indian manuscripts in Sanskrit, and to provide pertinent information about the manuscripts classification (construction of cladistic trees). The expected tools will not be restricted to Sanskrit language in every aspects. This action is complemented by the European AAT project support which allows us to collect more Sankrit manuscripts and to care about some interactive aspect that we where not able to take into account with the ACI dotation.

The end of action has been delayed until december 2007, a meeting in ENS Ulm (from 29th November until 1st December).where every team presented its results.

8.2.3. EGC Association: National Group on Mining Complex Data

URL: http://eric.univ-lyon2.fr/~gt-fdc/ AxIS members participated actively this year to the Working Group "Fouille de données complexes" created by D.A Zighed in June 2003 in the context of the EGC association:

- F. Masseglia with O. Boussaïd (ERIC, Lyon) co-organised and co-chaired the fourth workshop "Fouille de données complexes dans un processus d'extraction de connaissances" (23 January, 2007) [16]. M-A. Aufaure, B. Trousse and Y. Lechevallier were members of the program committee.
- F. Masseglia with O. Boussaïd co-animate one of the three topics: "Organisation and Structuration of Complex Data".
- B. Trousse with S. Després (CRIP5 - Université de Paris V) co-animate the topic " Knowledge in Complex Data Mining".

8.2.4. SFDS association: InfoStat Group

SFDS is the French Society of Statistics : URL: http://www.sfds.asso.fr/.
AxIS members participated actively this year in the workshops "Les après-midis d'InfoStat" of the InfoStat Group which is leaded by Y. Lechevallier (president):

- March 29, Paris: "les Fouilles de Données Textuelles".
- November 15, Paris: "NIVS : Nouvelles Idées Via les Software".

8.2.5. GDR-I3

AxIS is concerned by three working groups of the GDR-PRC~I3 National Research Group "Information - Interaction - Intelligence" of CNRS: working Group 3.4 (GT) on Data Mining, GRACQ, Working Group 3.7 on "Sécurité des Systèmes d'Information".

8.2.6. Other Collaborations

- LIP6: We work with Bénédicte Le Grand and Michel Soto in the Eiffel RNTL project on visualisation and navigation for enhancing semantic web retrieval in the tourism domain. Marie-Aude Aufaure works also with them on semantic and conceptual context-aware information retrieval [69].
- ENST Paris: Y. Lechevallier collaborated with Georges Hébrail (ENST) [30].
- Paris Descartes: Marc Csernel collaborate regularly with François Patte on every aspects concerning the Sankrit.
- ENST Bretagne: In the framework of the ACI "Histoire des savoirs" we have a regular collaboration with some ENST B researchers, namely P. Bertrand, M Le Pouliquen, J-P Barthélémy on classification and comparison of the Sanskrit Manuscripts.
- CNAM and Loria (Cortex Team): contacts have been established with research teams in human and social sciences.
- University of Bordeaux 1 and 2 (MAP laboratory): Y. Lechevallier collaborated with M. Chavent [24].
- GRIMM-SMASH team (Université Toulouse Le Mirail): F. Rossi works with N. Villa Self Organizing Map for dissimilarity matrices (cf Section 6.3.2 and [74]).

- LITA EA3097 (Université de Metz): F. Rossi works with Brieuc Conan-Guez on the Self Organizing Map for dissimilarity matrices (see Section 6.3.2 and [43]) .
- IRENav (Ecole Navale): MA Aufaure collaborates with C. Claramunt on spatial web personnalisation [75].

8.3. European Initiatives

8.3.1. Other Collaborations

- Germany: AxIS participated to the project "Core Technology Cluster" of the AII program "QUAERO" in the multimedia domain (ontology construction, personalization)
- Germany, Clausthal University of Technology, Department of Informatics (Prof. Barbara Hammer & Alexander Hasenfuss) and IPK Gatersleben, Pattern Recognition Group (Dr. Marc Strickert): F. Rossi [57], [70]
- Italy, University of Napoli II (Prof. R. Verde) [54],
- Italy, University La Sapienza (Roma) Prof Rafaele Torella and Dr Vincenzo Vergiani collaboration in the framework of th IT Asia Project on comparison of Sanskrit manuscripts.
- Belgium, Université Catholique de Louvain, DICE Laboratory (Prof. Michel Verleysen, Prof. Vincent Wertz, Dr. Damien François & Catherine Krier): F. Rossi [38], [64], [29]
- Belgium: a belgium delegation from Namur visited AxIS Sophia Antipolis on december 8th.

8.4. International Initiatives

8.4.1. Australia

We welcomed James Thom (associate professor at RMIT) as invited professor for 3 months at Inria Paris - Rocquencourt and also Jovan Pehcevski (RMIT) as postdoctoral fellow [88], [73], [68], [72], [67], [35], [61], [36]. James Thom made a seminar in July at AxIS Sophia Antipolis - Méditerranée.

8.4.2. Brazil

We continue our collaboration on clustering and web usage mining with F.A.T. De Carvalho from Federal University of Pernambuco (Recife) and his team.

- A scientific project submitted by Francisco De Carvalho and Yves Lechevallier has been accepted by FACEPE and INRIA. The project started from 04/2006 and ends on 03/2008. Researches and students are concerned by this project from AxIS and CIn-UFPE side. It aims at developing methods of clustering analysis and web usage mining tools.
- Francisco de Carvalho and Renata Souza visited AxIS project. During their stays, in collaboration with Yves Lechevallier, they participated to the design of dynamic clustering models based on adaptative distances and they finalized the conception of dynamic clustering models based on adaptative distances suitable to symbolic interval data. A complete paper has been submitted to the "Pattern Recognition" journal.
- In collaboration with Yves Lechevallier, Alzennyr Da Silva and Fabrice Rossi, Francisco de Carvalho has participated to the conception of an approach concerning the contruction of summaries via clustering methods of data which evolve overtime. An application of this approach has been done on data from web usage which evolves on the time.
- Marc Csernel visited CIn-UFPE during the two first weeks of february and during the two last june weeks. In collaborattion with Francisco de Carvalho they submitted a paper to "Pattern Recognition Letter". He worked also with Kelly Silva on her Master Thesis on the adaptation of Fuzzy clustering on symbolic data, specially in presence of rules as background knowledge.
- Kelly Silva spend three weeks at the begining of september to work with Marc Csernel and Yves Lechevallier.

8.4.3. Canada

Y. Lechevallier pursued his collaboration with A. Ciampi (Univ of McGill, Montréal).

8.4.4. China

Marie-Aude Aufaure collaborates with Yanwu Yang, Institute of Automation, Chinese Academy of Science, Beijing, on user modelling for the semantic web [75] and Yves Lechevallier collaborates with Hueiwen Wang, BUAA, Beijing on clustering methods.

8.4.5. Morocco

B. Trousse pursued her co-supervision with Abdelaziz Marzark (University of Casablanca) of a Ph.D. student: H. Behja (ENSAM, Meknès, Morocco). AxIS welcomed Mustapha We welcomed two students: Reda Kabbaj (University of Fès) [60] and Abdelmoujib Elkhoumri (ENSIAS Rabat, University of Settat). A co-supervision of their thesis was studied in collaboration with their professors. B. Trousse visited Reda Kabbaj during the VSTT conference at Marrakech.

8.4.6. Romania

We maintained our contacts with the Computer Science department of the West University of Timisoara (Prof Viorel Negru), in particular via the SYNASC conference every year.

8.4.7. Tunisia

Marie-Aude Aufaure and Yves Lechevallier are involved in co-supervision of masters and/or thesis (Riadi Lab, ENSI Tunis). These masters and thesis subjects are about web mining (usage, content and structure, using different methods) and ontology construction from heterogeneous sources [41] [22] [42].

9. Dissemination

9.1. Promotion of the Scientific Community

9.1.1. Journals and Books

AxIS is involved in the management and the edition of 3 journals and 2 books:

- Special issue on “Multimedia Data Mining” of the Multimedia Tools and Applications (MTAP Journal, vol 35:1, Springer, 2007): F. Masseglia (co-editor)
- the second special issue of the RNTi journal on “Fouille de données complexes”: B. Trousse (co-editor)
- member of the RSTI scientific committee related to the “ISI, L’OBJET, RIA, TSI” journals (Hermès publisher): B. Trousse
- La revue MODULAD (electronic journal, http://www.modulad.fr/): Y. Lechevallier is one of the four editors. F. Rossi is a member of the editorial board and S. Aubin is the webmaster of the web site.
- Two books on data mining, published by IGI Global in 2007: “Data Mining Patterns: New Methods and Applications” (ISBN 978-1599041629) and “Successes and New Directions in Data Mining” (ISBN 978-1599046457): F. masseglia (co-editor)

AxIS members belongs to editorial boards of four international journals, six national journals (or some of their special issues):

- Neurocomputing: F. Rossi (http://www.elsevier.com/locate/issn/09252312)
- the Co-Design Journal (Editor: S. Scrivener, Coventry University, UK - Publisher: Swets & Zeitlinger): B.Trousse
- the Journal of Symbolic Data Analysis (JSDA) (Editor: E. Diday, electronic journal http://www.jsda.unina2.it): Y. Lechevallier, F. Rossi and B. Trousse.
- European Journal of GIS and Spatial Analysis (“Revue Internationale de Géomatique”) http://geo.e-revues.com/: M-A. Aufaure
- the RSTI RIA journal (“Revue d’Intelligence Artificielle”) (Hermès publisher; editor-in-chief: M. Pomerol): B. Trousse.
- the I3 (Information, Interaction, Intelligence) electronic journal of the GDR-I3 (editor-in-chief: C. Garbay et H. Prade) http://www.Revue-I3.org/: B. Trousse.
- Special Issue "Points de vue" de la revue RSTI “L’OBJET’ (Publisher, Hermès Editor-in-chief: B. Coulette): B. Trousse
- RNTI Special Issue "Modélisation des connaissances" (Publisher, Cépaduès Editions, Editor-in-chief: H. Briand and J. Blanchard): B. Trousse
- RNTI Special Issue "Fouille du Web" (Publisher, Cépaduès Editions, editors-in-chief: C. raynaud and G. Venturini): M-A. Aufaure
- the I3 electronic journal of the GDR-I3 (editor-in-chief: C. Garbay et H. Prade) - Special issue “Visualisation et extraction des connaissances”: M-A. Aufaure

AxIS members were reviewers for 13 international and national journals and for 4 international books:

- INS International Journal on Information Sciences: F. Masseglia (http://ees.elsevier.com/ins/)
- JAIR Journal of Artificial Intelligent Research: F. Masseglia (http://www.jair.org)
- IJBET International Journal of Biomedical Engineering and Technology (special issue on Warehousing and Mining Complex Data: Applications to Biology, Medicine, Behavior, Health and Environment): F. Masseglia (http://www.inderscience.com/ijbet/)
- TKDE Transactions on Knowledge and Data Engineering: F. Masseglia, B. Trousse (http://www.computer.org/tkde/)
- IJHCS International Journal of Human-Computer Studies: M-A. Aufaure
- DKE International Journal on Data Knowledge and Engineering: M-A. Aufaure, F. Masseglia
- AAI (Applied Artificial Intelligence): M-A. Aufaure
- Econometric Reviews: F. Rossi
  (http://www.informaworld.com/smpp/title~content=t713597248)
- IEEE Transactions on Neural Networks: F. Rossi (http://ieee-cis.org/pubs/tnn/)
- IEEE Transactions on Pattern Analysis and Machine Intelligence: F. Rossi (http://www.computer.org/tpami/)
- Computational Statistics and Data Analysis: F. Rossi (http://www.elsevier.com/locate/csda)
- Neural Processing Letters: F. Rossi
  (http://www.springer.com/journal/11063/about)
- Neurocomputing: F. Rossi
  (http://www.elsevier.com/locate/issn/09252312)

9.1.2. Program Committees

Several AxIS members were involved in national or international conferences/worhshops as members of Program Committee.

9.1.2.1. National Conferences/Workshops

- EGC 2007: Namur, Belgique (Jan.23-26): Y. Lechevallier, B. Trousse
- INFORSID 2007: Perros-Guirec, France (May 22-25): B. Trousse
- EDA 2007: Poitiers, France (June 7-8): M-A. Aufaure
- Plate-forme AFIA 2007 (Atelier RàPC): Grenoble, France (July 2, 2007): B. Trousse
- ASD 2007 (Atelier sur les systèmes décisionnels): Sousse, Tunisie (October 19-20, 2007) - M-A. Aufaure (http://eric.univ-lyon2.fr/~asd/asd2007/)
- H2PTM 2007: Hammamet, Tunisie (Oct. 29-31): B. Trousse
- Ateliers EGC 2007: Namur, Belgium, January 2007
  - "Fouille de Données Complexes": M-A. Aufaure, B. Trousse, F. Masseglia
  - "Modélisation de Connaissances": B. Trousse

9.1.2.2. International Conferences/Workshops

- ESTSP: Otaniemi, Espoo, Finland (Feb. 7-9): F. Rossi
- EuroIMSA 2007: Chamonix, France (March 14-16): A-M. Vercoustre
- WDSA 2007: Caserta, Italy (March 15-16): Y. Lechevallier
- ESANN 2007: Bruges, Belgium (Apr. 25-27): F. Rossi
- IF&GIS 2007 (Third International workshop on Information Fusion and Geographic Information Systems): St-Petersbourg, Russia (May 25-27, 2007) - M-A. Aufaure (http://www.springer.com/east/home?SGWID=5-102-22-173674025-0)
- CSCWD 2007: Nanjing, China (April 26-28) - B. Trousse (http://2006.cscwid.org/
- IWANN 2007: San Sebastián, Spain (June 20–22): F. Rossi
- ACM SIGIR 2007: Amsterdam, the Netherlands (July 23-27): A-M. Vercoustre
- ISPA 2007: Niagara Falls, ON, Canada (August 29-31): F. Masseglia
- DocEng 2007: Winnipeg, Canada (August 28-31): A-M. Vercoustre
- ICANN 2007: Porto, Portugal (Sept. 9–13): F. Rossi
- ICTAI 2007: Patras, Greece (October 28-31): F. Masseglia, B. Trousse (reviewer) (http://ictai07.ceid.upatras.gr/)
- ICDM 2007: Omaha, NE, USA (Oct. 28-31): F. Masseglia
- PIKM 2007: Lisbon, Portugal (Nov. 5-10): F. Masseglia
- OTM COOPIS 2007: Vilamoura, Portugal (Nov. 28-30, 2007): B. Trousse (http://www.cs.rmit.edu.au/fedconf/2007/index.html?page=coopis2007cfp)
- SYNASC 2007: Timisoara, Romania (Sept. 26-29, 2006): B. Trousse (http://synasc07.info.uvt.ro/)
- ADCS 2007: Melbourne, Australia (Dec. 10): A-M.Vercoustre

9.1.3. Organization of Conferences or Workshops

We are involved in other organization tasks:

- Co-organisation of the Fourth workshop FDC at EGC07: F. Masseglia (http://www-sop.inria.fr/axis/fdc-egc07/)

- INEX: co-organisation of the track on XML mining [55]: A-M. Vercoustre.
- INEX: co-organisation of the new track on Entity Ranking (2007): A-M. Vercoustre. (http://inex.is.informatik.uni-duisburg.de/2007/xmlSearch.html)
- INEX: participation in the definition of the official measures of retrieval effectiveness planned to be employed for the ad hoc track of INEX 2007 [67]: Jovan Pehcevski.
- Organisation of our annual AxIS workshop at Inria Rocquencourt (July 2-4): S. Aubin, S. Honnorat, Y. Lechevallier and B. Trousse. Furthermore, monthly team meetings were organised by videoconference between AxIS Sophia Antipolis and AxIS Rocquencourt.
- Organisation of a regular internal seminar by J. Pehcevski at AxIS Rocquencourt.

9.1.4. AxIS Web Server

AxIS maintains an external and an internal Web site allowing the access to lots of information, including software developed in the team, our publications, relevant events (conferences, workshops) and information related to the conferences and seminar we organise. URL:http://www-sop.inria.fr/axis/.
AxIS uses its own publication management tool called "BibAdmin" developed by S. Chelcea (cf. section 5.9) available on Inria's Gforge server http://gforge.inria.fr/projects/bibadmin/.

9.1.5. Activities of General Interest

- Y. Lechevallier is the president of the "InfosStat, Logiciels et Data Mining" Group of the SFDS society http://www.sfds.asso.fr/groupes/logiciel.htm. Three InfoStat seminars were organized in January, March and October.
- T. Despeyroux is involved (30 %) as president of AGOS (Inria Works Council), a permanent member of the "Commission technique paritaire (CTP)" and a member of the Inria Board of Directors (Conseil d'Administration) as a scientific staff representative.
- B. Senach is involved (10 %) in the support committee (Inria Sophia Antipolis) of the world-wide competitivity pole "Solutions Communicantes Sécurisées". During this year he has been involved in four Inria workings groups for Sophia Antipolis: 1) the "Comorale" workgroup (Inria Sophia Antipolis) which is in charge of an internal survey to identify users' needs in information exchange cooperation and to drive the choice of future communications tools, 2) the "Circulation de l'information" workgroup (Inria Sophia Antipolis) which has to analyse the information flow inside Inria Sophia Antipolis Unit and to improve information dissemination, 3) an internal "think tank" working group (CUMIR) which has to envision the futur user's need within INRIA to plan the requiered evolution of technicological ressources and 4) the "Refonte du site Web Inria Sophia" workgroup which has to give specifications for the redesign of the Research Center Web Site.
- A-M. Vercoustre is involved (25%) in the Department for Scientific Information and Communication (DISC), working on Inria policy and tools for scientific publications, in particular the development of the Open Archive HAL, in cooperation with CNRS. She is a member of the COST (Comité scientifique et technique du Comité stratégique) for the extension of HAL to become the French National Open Archive. As part of her DISC involvement, A-M. Vercoustre is also leading the Ralyx project for exploiting the INRIA Activity Report (cf. section 5.8).

9.2. Formation

9.2.1. University Teaching

AxIS is "associated team" for the STIC Doctoral school at the University of Nice Sophia Antipolis (UNSA) and AxIS team members are teaching in various universities:

- "Master PMLT" (resp. Mr Kounalis) at UNSA Sophia Antipolis: Tutorial (12h) on *Data Mining and Web Mining*: F. Masseglia, D. Tanasa, B. Trousse (resp.).

- "Licence professionnelle franco-italienne: Statistiques et Traitement Informatique de Données (STID)" (resp. J. Lemaire) at UNSA, Menton: Supervision of a student project (60h by students, 8 students, 30h supervised) on *Mining HTTP Logs From Inria's Web Sites*: S. Gaieb (co-resp.), D. Tanasa, B. Trousse (co-resp.).
- Master 2 Recherche "Systèmes intelligents" (resp: S. Pinson) of the University Paris IX-Dauphine: Tutorial (8h) on *"Analyse des connaissances numériques et Symboliques"*: Y. Lechevallier.
- Master 2 Pro "Mathématiques appliquées et sciences économiques"(resp: P. Cazes) of the University Paris IX-Dauphine: Tutorial (18h) on *"Méthodes neuronales en classification"*: Y. Lechevallier.
- Master 2 Pro "Ingénierie de la Statistique" (resp: G. Saporta) of CNAM (12h) on *"Méthodes neuronales"*: Y. Lechevallier.
- ENSAE ("Ecole Nationale de la Statistique et de l'Administration Economique"): Tutorial (12h) on *"Data Mining"*: Y. Lechevallier.
- Master 2 recherche Informatique, Paris XI: Tutorial (3h) on *Ontology construction*: M-A. Aufaure.
- Master 1 Pro "Ingénierie Mathématique pour les Sciences du Vivant" (resp: B. Le Roux et M. Kratz) of Université Paris V, introduction to artificial neural networks (15h): F. Rossi.

9.2.2. H.D.R and Ph.D. Thesis

Ph.D. in progress:

1. **H. Behja**, (start: end of 2002), "Gestion de points de vues multiples dans l'analyse d'un observatoire sur le Web", University of Casablanca, (directors: A. Marzark and B. Trousse). This thesis is done in the context of the STIC Software engineering network of France-Morocco cooperation (2002-2005).
2. **A. Marascu**, (start: October 2005), "Extraction de Motifs Séquentiels dans les Data Streams", Université de Nice-Sophia Antipolis (director: Yves Lechevallier, with the participation of F. Masseglia).
3. **A. Da Silva**, (start: October 2005), "Modélisation de données agrégées ou complexes par l'approche symbolique, application au Web Usage Mining", University of Paris IX Dauphine (directors: Edwin Diday and Yves Lechevallier).

F. Rossi is a member of the thesis committees of **C. Krier** (start: October 2005) on "Analyse de données de grande dimension en particulier en spectométrie", Université Catholique de Louvain, Belgium (director: Michel Verleysen).
M-A Aufaure is co-superviser with Mohammed Ben Ahmed for R. Djedidi's thesis (start: end of 2005): "Towards a generic approach for ontology construction from heterogeneous sources", University Paris XI and University La Manouba (Tunisia).

Y. Lechevallier is co-superviser with G. Saporta for M. Charrad's thesis (start: end of 2005), CNAM. and University La Manouba (Tunisia).

AxIS researchers were members of H.D.R or Ph.D. committees in 2007:

- **L. Berti**, H.D.R. "Quality Awareness for Managing and Mining Data", University of Rennes 1 (Computer Science) : B. Trousse
- **A. Baldé**, PH.D, "Utilisation de métadonnées et ontologie pour l'aide à l'interprétation des résultats de classification", May: Y. Lechevallier, M-A. Aufaure and B. Trousse;
- **S. Chelcea**, Ph.D, "Agglomerative 2-3 Hierarchical Clustering: theoretical and applicative study", Université de Nice-Sophia Antipolis (directors: J. Lemaire and B. Trousse with the support of P. Bertrand on 2-3 AHC), : Y. Lechevallier, B. Trousse
- **N. Delannay**, Ph.D, "Aspects of probabilistic modelling for data analysis", October, Université Catholique de Louvain: F. Rossi

9.2.3. Internships

We welcomed five students in AXIS this year:

1. **R. Kabbaj** (supervisors B. Senach and B. Trousse), Master 2, Faculté des Sciences Sidi Mohamed Ben Abdellah de Fès, Morocco, “Mise en correspondance des appels d’offre et des équipes de recherche d’un organisme: application à l’Inria” [89].
2. **C. Maurice** (supervisor Y. Lechevallier), Master 2, (Université de Namur, Belgique),“ Navigation et visualisations graphiques sur PDA” [80].
3. **G. Pilot** (supervisors B. Senach and B. Trousse), Licence Pro STID, IUT Menton, “Projet MobiVIP: Véhicules individuels publics pour la mobilité en centre ville: analyse de données et manuel utilisateur du logiciel d’analyse” [81].
4. **B. Saleh** (supervisor F. Masseglia), Master 2, (Université de Nice Sophia-Antipolis), “Découverte de connaissance avec un découpage optimal des données” [90].
5. **A. Thibau**, Johns Hopkins University, Baltimore, MD, USA.

9.3. Participation to Workshops, Conferences, Seminars, Invitations

Furthermore we attended the following conferences or workshops:

- Colloque STIC, November 5-7 2007, Paris: M-A. Aufaure
- INEX’07 Workshop (Initiative for the Evaluation of XML Retrieval), Schloss Dagstuhl, Germany, December 17-19: A-M. Vercoustre, J. Pehcevski
- Seminars Jean-Pierre Fénelon on "Analyse des données": Y. Lechevallier

F. Rossi was invited professor for one month at the Université Catholique de Louvain (Belgium) for 2006-2007: he visited the university three weeks in 2007.

10. Bibliography

Major publications by the team in recent years

[1] E. GUICHARD (editor). *Mesures de l’internet*, ouvrage collectif suite au Colloque Mesures de l’internet, Nice, France, 12-14 Mai, 2003, Les Canadiens en Europe, 2004.

[2] P. BERTRAND, M. F. JANOWITZ. *The k-weak hierarchical representations: an extension of the indexed closed weak hierarchies*, in "Discrete Applied Mathemathics", vol. 127, n^{o} 2, April 2003, p. 199–220.

[3] M. CHAVENT, F. DE CARVALHO, Y. LECHEVALLIER, R. VERDE. *New clustering methods for interval data*, in "Computational Statistics", vol. 21, n^{o} 23, 2006, p. 211-230, http://www-sop.inria.fr/axis/Publications/uploads/pdf/ChaventDeCarvalhoLechevallierVerdeFinalVersion.pdf.

[4] S. CHELCEA, P. BERTRAND, B. TROUSSE. *Un Nouvel Algorithme de Classification Ascendante 2-3 Hiérarchique*, in "Actes de 14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle (RFIA 2004), Centre de Congrès Pierre BAUDIS, Toulouse, France", vol. 3, 28-30 Janvier 2004, p. 1471-1480, http://www.laas.fr/rfia2004/actes/ARTICLES/388.pdf.

[5] B. CONAN-GUEZ, F. ROSSI, A. EL GOLLI. *Fast Algorithm and Implementation of Dissimilarity Self-Organizing Maps*, in "Neural Networks", vol. 19, n^{o} 6-7, August 2006, p. 855–863, http://dx.doi.org/10.1016/j.neunet.2006.05.002.

[6] A. DA SILVA, Y. LECHEVALLIER, F. DE CARVALHO, B. TROUSSE. *Mining Web Usage Data for Discovering Navigation Clusters*, in "11th IEEE Symposium on Computers and Communications (ISCC'06), Pula-Cagliari, Italy", IEEE Computer Society, 26-29 June 2006, p. 910-915, http://doi.ieeecomputersociety.org/10.1109/ISCC.2006.102.

[7] T. DESPEYROUX. *Practical Semantic Analysis of Web Sites and Documents*, in "The 13th World Wide Web Conference, WWW2004, New York City, USA", 17-22 May 2004, http://www-sop.inria.fr/axis/papers/04www/despeyroux-www2004.pdf.

[8] A. EL GOLLI, B. CONAN-GUEZ, F. ROSSI, D. TANASA, B. TROUSSE, Y. LECHEVALLIER. *Une application des cartes topologiques auto-organisatrices à l'analyse des fichiers Logs*, in "Actes des onzièmes journées de la Société Francophone de Classification, Bordeaux, France", Septembre 2004, p. 181–184.

[9] G. HÉBRAIL, Y. LECHEVALLIER. *Data mining et analyse des données*, in "Analyse des données", Hermes, June 2003, p. 340-360.

[10] A. MARASCU, F. MASSEGLIA. *Mining Sequential Patterns from Data Streams: a Centroid Approach*, in "Journal of Intelligent Information Systems (JIIS).", vol. 27, n^{o} 3, November 2006, p. 291-307.

[11] F. MASSEGLIA, P. PONCELET, M. TEISSEIRE, A. MARASCU. *Web Usage Mining: Extracting Unexpected Periods from Web Logs*, in "Data Mining and Knowledge Discovery (DMKD) Journal.", DOI 10.1007/s10618-007-0080-z, 2007, http://www-sop.inria.fr/axis/Publications/uploads/pdf/period2.pdf.

[12] F. MASSEGLIA, D. TANASA, B. TROUSSE. *Web Usage Mining: Sequential Pattern Extraction with a Very Low Support*, in "Advanced Web Technologies and Applications: 6th Asia-Pacific Web Conference, APWeb 2004, Hangzhou, China. Proceedings", LNCS, vol. 3007, Springer-Verlag, 14-17 April 2004, p. 513–522.

[13] F. ROSSI, B. CONAN-GUEZ. *Functional Multi-Layer Perceptron: a Nonlinear Tool for Functional Data Analysis*, in "Neural Networks", vol. 18, n^{o} 1, January 2005, p. 45-60, http://hal.inria.fr/inria-00000599.

[14] D. TANASA, B. TROUSSE. *Advanced Data Preprocessing for Intersites Web Usage Mining*, in "IEEE Intelligent Systems", vol. 19, n^{o} 2, March-April 2004, p. 59–65.

Year Publications

Books and Monographs

[15] F. BOUALI, F. MASSEGLIA, L. KHAN (editors). *Special Issue on Multimedia Data Mining (MDM/KDD'06)*, Multimedia Tools and Applications (MTAP), vol. 35:1, Springer, 2007, http://www.springerlink.com/content/qm783022j1l0/?p=9e9a7978dc414397a96e4cb243334524&pi=1.

[16] O. BOUSSAID, F. MASSEGLIA (editors). *Actes de FDC'07, le quatrième atelier sur la " Fouille de données complexes dans un processus d'extraction de connaissances "*, 2007.

[17] F. MASSEGLIA, P. PONCELET, M. TEISSEIRE (editors). *Successes and New Directions in Data Mining*, Information Science Reference, ISBN 978-1599046457, Idea Group, 2007.

[18] P. PONCELET, F. MASSEGLIA, M. TEISSEIRE (editors). *Data Mining Patterns: New Methods and Applications*, Premier Reference Source, ISBN 978-1599041629, Idea Group, 2007.

[19] Z. (. ZHANG, F. MASSEGLIA, R. JAIN, A. D. BIMBO (editors). *Proceedings of MDM'06, the seventh international workshop on " Multimedia Data Mining "*, (held in conjunction with KDD'06), 2007, http://www.fortune.binghamton.edu/MDM2006/.

Doctoral dissertations and Habilitation theses

[20] A. BALDÉ. *Utilisation de métadonnées et d'ontologie pour l'aide à l'interprétation des résultats de classification*, Thèse de doctorat, University of Paris Dauphine UFR EDDIMO, May 2007.

[21] S. CHELCEA. *Agglomerative 2-3 Hierarchical Classification: Theoretical and Applicative Study*, Ph. D. Thesis, University of Nice Sophia Antipolis, March 2007.

Articles in refereed journals and book chapters

[22] H. BAAZAOUI-ZGHAL, M.-A. AUFAURE, N. BEN MUSTAPHA. *Extraction of Ontologies from web pages: conceptual modelling and tourism application*, in "Journal of Internet Technology", 2007.

[23] L. CANDILLIER, L. DENOYER, P. GALLINARI, M.-C. ROUSSET, A. TERMIER, A.-M. VERCOUSTRE. *Mining XML documents*, in "Data Mining Patterns: New Methods and Applications", P. PONCELET, M. TEISSEIRE, F. MASSEGLIA (editors), chap. 9, Idea Group Reference, September 2007, p. 198-219, http://hal.inria.fr/inria-00188899/en/.

[24] M. CHAVENT, Y. LECHEVALLIER, O. BRIANT. *DIVCLUS-T: A monothetic divisive hierarchical clustering method*, in "Computational Statistics and Data Analysis", vol. 52, 2007, p. 687-701.

[25] A. CIAMPI, Y. LECHEVALLIER. *Statistical Models and Artificial Neural Networks: Supervised Classification and Prediction Via Soft Trees*, in "Advances in Statistical Methods for the Health Sciences", J.-L. AUGET, N. BALAKRISHNAN, M. MESBAH, G. MOLENBERGHS (editors), Statistics for Industry and Technology, chap. 16, Birkhäuser, 2007, p. 239–262.

[26] M. CSERNEL, P. BERTRAND. *Sanskrit Manuscript Comparison For Critical Edition an Classficication*, in "Selected Contributions in Data Analysis and Classification", P. BRITO, P. BERTRAND, G. CUCUMEL, F. DE CARVALHO (editors), Springer, 2007, p. 557–566.

[27] A. DA SILVA. *Analyzing the Evolution of Web Usage Data.*, in "Special issue on Data Stream Analysis of MODULAD (Monde des Utilisateurs de L'Analyse de Données)", ISSN: 17697387, vol. 1, n^{o} 36, May 2007, p. 75-84.

[28] A. DA SILVA. *Diverses approches permettant l'introduction du temps dans la fouille de données d'usage du Web*, in "Revue des Nouvelles Technologies de l'Information", Edited by Ch. Reynaud and G. Venturini, vol. RNTI W-1 Fouille du Web, 2007, p. 35-55.

[29] D. FRANÇOIS, F. ROSSI, V. WERTZ, M. VERLEYSEN. *Resampling methods for parameter-free and robust feature selection with mutual information*, in "Neurocomputing", vol. 70, n^{o} 7–9, March 2007, p. 1276–1288, http://hal.inria.fr/inria-00174298.

[30] G. HÉBRAIL, Y. LECHEVALLIER. *Building Symbolic Objects from Data Streams*, in "Selected Contributions in Data Analysis and Classification", P. BRITO, P. BERTRAND, G. CUCUMEL, F. DE CARVALHO (editors), Studies in Classification, Data Analysis and Knowledge Organization, Springer, 2007, p. 82-94, http://www-sop.inria.fr/axis/Publications/uploads/pdf/ED2007Hebrail.pdf.

[31] L. KAROUI, M.-A. AUFAURE. *Revealing Criteria for the Ontology Evaluation Task*, in "Journal of Internet Technology", 2007.

[32] Y. LECHEVALLIER, A. CIAMPI. *Multilevel Clustering for large Databases*, in "Advances in Statistical Methods for the Health Sciences", J.-L. AUGET, N. BALAKRISHNAN, M. MESBAH, G. MOLENBERGHS (editors), Statistics for Industry and Technology, chap. 10, Birkhäuser, January 2007, p. 263–274, http://www-sop.inria.fr/axis/Publications/uploads/pdf/Lechevallier-Ciampi-Nantes-2004.pdf.

[33] C. L. MANUEL, A. CIAMPI, Y. LECHEVALLIER. *Hierarchical Clustering of Subpopulations with a dissimilarity based on the likelihood ratio statistic: Application to clustering massive data sets.*, in "Pattern Analysis & Applications", 2007.

[34] F. MASSEGLIA, P. PONCELET, M. TEISSEIRE, A. MARASCU. *Web Usage Mining: Extracting Unexpected Periods from Web Logs*, in "Data Mining and Knowledge Discovery (DMKD) Journal.", DOI 10.1007/s10618-007-0080-z, 2007, http://www-sop.inria.fr/axis/Publications/uploads/pdf/period2.pdf.

[35] J. PEHCEVSKI, B. LARSEN. *Relevance*, in "Encyclopedia of Database Systems", L. LIU, M. T. ÖZSU (editors), Springer-Verlag, 2007, http://hal.inria.fr/inria-00174160/en/.

[36] J. PEHCEVSKI, B. PIWOWARSKI. *Evaluation metrics*, in "Encyclopedia of Database Systems", L. LIU, M. T. ÖZSU (editors), Springer-Verlag, 2007, http://hal.inria.fr/inria-00174152/en/.

[37] J. PEHCEVSKI, B. PIWOWARSKI. *Specificity*, in "Encyclopedia of Database Systems", L. LIU, M. T. ÖZSU (editors), Springer-Verlag, 2007, http://hal.inria.fr/inria-00174155/en/.

[38] F. ROSSI, D. FRANÇOIS, V. WERTZ, M. VERLEYSEN. *Fast Selection of Spectral Variables with B-Spline Compression*, in "Chemometrics and Intelligent Laboratory Systems", vol. 86, n^{o} 2, April 2007, p. 208–218, http://hal.inria.fr/inria-00174299.

[39] B. TROUSSE, M.-A. AUFAURE, B. LE GRAND, Y. LECHEVALLIER, F. MASSEGLIA. *Web Usage Mining for Ontology Management*, in "Data Mining with Ontologies: Implementations, Findings and Frameworks.", N. HÈCTOR OSCAR, GONZALEZ CISARO. SANDRA ELISABETH G, X. DANIEL HUGO (editors), chap. 3, Information Science Reference, 2007, p. 37-64.

Publications in Conferences and Workshops

[40] M.-A. AUFAURE, B. LE GRAND, M. SOTO. *Sémantique et contextes conceptuels pour la recherche d'information*, in "7èmes journées francophones " Extraction et Gestion des Connaissances ", EGC, Revue des Nouvelles Technologies de l'Information RNTI-E-9", 2007.

[41] M.-A. AUFAURE, R. SOUSSI, H. BAAZAOUI-ZGHAL. *SIRO: On-Line Semantic Information Retrieval using Ontologies*, in "The Second International Conference on Digital Information Management (ICDIM'2007)", 28-31 October 2007.

[42] N. BEN MUSTAPHA, H. BAAZAOUI-ZGHAL, M.-A. AUFAURE. *A prototype for Knowledge Extraction from Semantic Web based on Ontological Components Construction*, in "3rd International Conference on Web Information Systems and Technologies (WEBIST)", 3-6 March 2007, p. 451-454.

[43] B. CONAN-GUEZ, F. ROSSI. *Speeding Up the Dissimilarity Self-Organizing Maps by Branch and Bound*, in "Proceedings of 9th International Work-Conference on Artificial Neural Networks (IWANN 2007), San Sebastian (Spain)", S. F., A. PRIETO, J. CABESTANY, M. GRANA (editors), Lecture Notes in Computer Science, n^{O} 4507, June 2007, p. 203–210.

[44] M. CSERNEL, P. BERTRAND. *Différences, Distances entre Texte Sanskrit, élaboration d'édition critique.*, in "XIV èmes rencontres de la Société Francophone de Classification (SFC), Enst Paris , France", September 2007, p. 74–77.

[45] M. CSERNEL, F. PATTE. *Sanskrit Manuscript Comparison for Critical Edition*, in "First International Sanskrit Computational Linguistics Symposium, Inria Rocquecourt", October 2007, p. 103–121.

[46] A. DA SILVA. *Analyzing the Evolution of Web Usage Data*, in "European Workshop on Data Stream Analysis (WSDA 2007), Caserta, Italy", article published by the MODULAD magazine, March 2007, p. 75–84.

[47] A. DA SILVA, F. DE CARVALHO, Y. LECHEVALLIER. *Analysing Distance Measures for Symbolic Data Based on Fuzzy Clustering*, in "Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007), Rio de Janeiro, Brazil", L. MOURELLE, N. NEDJAH, J. KACPRZYK (editors), ISBN: 0-7695-2976-3. doi: 10.1109/ISDA.2007.52, IEEE Computer Society, 22-24 October 2007, p. 109-114, http://doi.ieeecomputersociety.org/10.1109/ISDA.2007.40.

[48] A. DA SILVA, Y. LECHEVALLIER, F. ROSSI, F. DE CARVALHO. *Classifications non supervisées de données évolutives : application au Web Usage Mining*, in "Atelier N r c4 : Flux de données, 7ème journées Extraction et Gestion des Connaissances (EGC 2007), Namur, Belgique", 23 January 2007, p. 31-40.

[49] A. DA SILVA, Y. LECHEVALLIER, F. ROSSI, F. DE CARVALHO. *Clustering Strategies for Detecting Changes on Web Usage Data*, in "56th Session of the International Statistical Institute (ISI 2007), Lisbon, Portugal", 22-29 August 2007.

[50] A. DA SILVA, Y. LECHEVALLIER, F. ROSSI, F. DE CARVALHO. *Construction and Analysis of Evolving Data Summaries: an Application on Web Usage Data*, in "Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007), Rio de Janeiro, Brazil", L. MOURELLE, N. NEDJAH, J. KACPRZYK (editors), ISBN: 0-7695-2976-3. doi: 10.1109/ISDA.2007.51, IEEE Computer Society, 22-24 October 2007, p. 377-380, http://doi.ieeecomputersociety.org/10.1109/ISDA.2007.59.

[51] A. DA SILVA, Y. LECHEVALLIER, F. ROSSI, F. DE CARVALHO. *Construction et analyse de résumés de données évolutives : application aux données d'usage du Web*, in "Actes des 7ème journées Extraction et Gestion des Connaissances (EGC 2007), Revue des Nouvelles Technologies de l'Information (RNTI), Namur, Belgique", vol. II, Cépaduès-éditions, 23-26 January 2007, p. 539-544, http://www.info.fundp.ac.be/egc2007/programme-courts.php.

[52] A. DA SILVA, Y. LECHEVALLIER, F. ROSSI, F. DE CARVALHO. *Groupement de données évolutives dans la fouille d'usage du Web*, in "39èmes Journées de Statistique de la SFdS, Angers, France", 11-15 June 2007.

[53] F. DE CARVALHO, Y. LECHEVALLIER. *Une méthode de partionnement sur un ensemble de tableaux de distances*, in "Actes des XIVes Rencontres de la Société Francophone de Classification, Paris", ENST, Société Francophone de Classification, September 2007, p. 79–82.

[54] F. DE CARVALHO, Y. LECHEVALLIER, R. VERDE. *Clustering approach on interval data*, in "Classification and Data Analysis (Sixth Meeting of the CLAssification and Data Analysis Group (CLADAG) of the Italian Statistical Society), Macerata (Italy)", Economia – Statistica, Eum (edizioni università di macerata), CLADAG, 12-14 September 2007, p. 139-142.

[55] L. DENOYER, P. GALLINARI, A.-M. VERCOUSTRE. *Report on the XML Mining Track at INEX 2005 and INEX 2006, Categorization and Clustering of XML Documents*, in "5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006, Dagstuhl, Germany", N. FUHR, M. LALMAS, S. MALIK, G. KAZAI (editors), Lecture Notes in Computer Science, vol. 4518, Springer, 2007, p. pp. 432-443, http://hal.inria.fr/inria-00173420/en/.

[56] T. DESPEYROUX, E. FRASCHINI, A.-M. VERCOUSTRE. *Extraction d'entités dans des collections évolutives*, in "7ièmes Journées francophones Extraction et Gestion des Connaissances EGC 2007, Namur, Belgium", G. VENTURINI, M. NOIRHOMME-FRAITURE (editors), Revue des Nouvelles Technologies de l'Information (RNTI-E-9), vol. 76300, Cépaduès, 2007, p. pp. 533-538, http://hal.inria.fr/inria-00116910/en/.

[57] B. HAMMER, A. HASENFUSS, F. ROSSI, M. STRICKERT. *Topographic Processing of Relational Data*, in "Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 07), Bielefeld (Germany)", ISBN: 978-3-00-022473-7, September 2007, http://dx.doi.org/10.2390/biecoll-wsom2007-121.

[58] Z. JRAD, M.-A. AUFAURE, M. HADJOUNI. *A Contextual user model for Web personalization*, in "Personalized Acces to Web Information (PAWI'2007), Nancy, france", 3-7 December 2007, 12.

[59] Z. JRAD, M.-A. AUFAURE. *Personalized interfaces for a semantic web portal: tourism information search*, in "11th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES2007), Vietri sul Mare, Italy", 12-14 September 2007.

[60] R. KABBAJ, B. TROUSSE, B. SENACH. *Aide à la soumission aux appels d'offres : mapping bidirectionnel par classification de textes*, in "Cinquième Colloque VSST Veille Stratégique Scientifique & Technologique", 21-25 October 2007.

[61] J. KAMPS, M. LALMAS, J. PEHCEVSKI. *Evaluating relevant in context: document retrieval with a twist*, in "Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands", ISBN 978-1-59593-597-7, 2007, p. 749-750, http://hal.inria.fr/inria-00174140/en/.

[62] L. KAROUI, M.-A. AUFAURE, N. BENNACER. *Contextual Concept Discovery Algorithm*, in "FLAIRS Conference", 2007, p. 460-465.

[63] L. KAROUI, M.-A. AUFAURE, N. BENNACER. *Contextualization Vs Decontextualization : impact on the expert's evaluation*, in "Fourth IEEE International Multiconference on Systems, Signals and Devices SSD07", 2007.

[64] C. KRIER, F. ROSSI, D. FRANÇOIS, M. VERLEYSEN. *Feature clustering and mutual information for the selection of variables in spectral data*, in "Proceedings of XVth European Symposium on Artificial Neural Networks (ESANN 2007), Bruges (Belgium)", April 2007, p. 157–162.

[65] M. LALMAS, G. KAZAI, J. KAMPS, J. PEHCEVSKI, B. PIWOWARSKI, S. ROBERTSON. *INEX 2006 evaluation measures*, in "Comparative Evaluation of XML Information Retrieval Systems: 5th International

Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006, Dagstuhl Castle, Germany, December 17-20, 2006, Revised and Selected Papers", Lecture Notes in Computer Science, ISBN 978-3-540-73887-9, vol. 4518, Springer Berlin / Heidelberg, 2007, p. 20-34, http://hal.inria.fr/inria-00174121/en/.

[66] A. MARASCU, F. MASSEGLIA. *Limites d'une approche incrémentale pour la segmentation de séquences dans les flux*, in "Fouille de données complexes dans un processus d'extraction des connaissances (FDC), Namur, Belgique", 23 January 2007, p. 49-60.

[67] J. PEHCEVSKI, J. KAMPS, G. KAZAI, M. LALMAS, P. OGILVIE, B. PIWOWARSKI, S. ROBERTSON. *INEX 2007 evaluation measures*, in "Pre-Proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007, Dagstuhl Castle, Germany, December 17-19, 2007", 2007, http://hal.inria.fr/inria-00174184/en/.

[68] J. PEHCEVSKI, J. A. THOM. *Evaluating focused retrieval tasks*, in "Proceedings of the SIGIR 2007 Workshop on Focused Retrieval, Amsterdam, The Netherlands", ISBN 978-0-473-12333-8, 2007, p. 33-40, http://hal.inria.fr/inria-00166790/en/.

[69] G. POLAILLON, M.-A. AUFAURE, B. LE GRAND, M. SOTO. *FCA for contextual semantic navigation and information retrieval in heterogeneous information systems*, in "workshop on Advances in Conceptual Knowledge Engineering, in conjunction with DEXA 2007", 3-7 September 2007.

[70] F. ROSSI, A. HASENFUSS, B. HAMMER. *Accelerating Relational Clustering Algorithms With Sparse Prototype Representation*, in "Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 07), Bielefeld (Germany)", ISBN: 978-3-00-022473-7, September 2007, http://dx.doi.org/10.2390/biecoll-wsom2007-144.

[71] F. ROSSI. *Model collisions in the dissimilarity SOM*, in "Proceedings of XVth European Symposium on Artificial Neural Networks (ESANN 2007), Bruges (Belgium)", April 2007, p. 25–30, http://apiacoa.org/publications/2007/dsom-collision-esann.pdf.

[72] J. THOM, J. PEHCEVSKI, A.-M. VERCOUSTRE. *Use of Wikipedia Categories in Entity Ranking*, in "The 12th Australasian Document Computing Symposium (ADCS'07), Melbourne, Australie", December 2007, http://hal.inria.fr/inria-00188790/en/.

[73] A.-M. VERCOUSTRE, J. PEHCEVSKI, J. THOM. *Using Wikipedia Categories and Links in Entity Ranking*, in "Pre-Proceedings of the Sixth International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007), Schloss Dagstuhl, Germany", Postprint, 2007, http://hal.inria.fr/inria-00192489/en/.

[74] N. VILLA, F. ROSSI. *A comparison between dissimilarity SOM and kernel SOM for clustering the vertices of a graph*, in "Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 07), Bielefeld (Germany)", ISBN: 978-3-00-022473-7, September 2007, http://dx.doi.org/10.2390/biecoll-wsom2007-139.

[75] Y. YANG, M.-A. AUFAURE, C. CLARAMUNT. *Towards a DL-Based semantic user model for web personalization*, in "The First International Workshop on Knowledge-based User Interface, KUI 2007, Athens, Greece", IEEE Computer Society Press, 19 June 2007.

Internal Reports

[76] Z. JRAD, L. NOËL, B. SENACH. *Dossier de conception : Analyse de l'existant, scénarios d'usage et scénarios d'interaction*, 109 pages, Délivrable Projet RNTL EIFFEL, n^{o} LSP 8.1, Inria, July 2007.

[77] M. JURCA, S. GAÏEB, B. TROUSSE. *Contexte expérimental et moteur de traces dans Net.Portal*, 44 pages. Non public, Délivrable Projet RNTL EPIA, n^{o} D3, Inria, June 2007.

[78] M. JURCA, Y. LECHEVALLIER, A. EL GOLLI, F. MASSEGLIA, B. SENACH, B. TROUSSE. *Contexte expérimental et moteur de traces dans Net.Portal - Analyse des besoins utilisateurs*, Non public, Délivrable Projet RNTL EPIA, n^{o} D3-Annexe, Inria, June 2007.

[79] Y. LECHEVALLIER, L. BEAUDOIS, A. EL GOLLI, B. TROUSSE. *Canal.Net pour Mediapps.Net Component.Net et Admin.Usage pour Net.Portal : Outils de classification*, 34 pages. Non public, Délivrable Projet RNTL EPIA, n^{o} D2 D4 D6, Inria, June 2007.

[80] C. MAURICE. *Extraction de données touristiques et visualisation sur PDA*, Technical report, Facultés unviversitaires Notre Dame de la Paix, Namur, 2007, http://www-sop.inria.fr/axis/Publications/uploads/pdf/Rapport.CyrillMaurice.pdf.

[81] G. PILOT, B. SENACH. *Manuel utilisateur de l'environnement d'analyse des données VIP*, 36 pages, User manual, Inria, July 2007.

[82] B. SENACH, B. TROUSSE. *Evaluation de la maquette Net.CanalRecommender - Conception de l'expérimentation*, 23 pages. Non public, Délivrable Projet RNTL EPIA, n^{o} D7, Inria, June 2007.

[83] B. TROUSSE, S. GAÏEB, B. SENACH. *Spécification et réalisation du système Net.CanalRecommender, un assistant d'aide à la navigation dans un portail d'entreprise*, 53 pages. Non public, Délivrable Projet RNTL EPIA, n^{o} D5-part1, Inria, June 2007.

[84] B. TROUSSE, B. SENACH, C. MANGEAT, G. CLOUET. *Analyse de l'usage d'un site d'information voyageurs (rapport final)*, 89 pages, Délivrable Projet PREDIT MobiVIP, n^{o} D5.3, Inria, March 2007.

[85] B. TROUSSE, B. SENACH, C. MANGEAT, G. CLOUET. *Classification des profils et comportements utilisateurs pour l'évaluation du service de mobilité*, 81 pages. Non public, Délivrable Projet PREDIT MobiVIP, n^{o} D5.4, Inria, May 2007.

[86] B. TROUSSE, B. SENACH, C. MANGEAT, G. PILOT. *Définition du plan d'évaluation du service VIP sur Antibes*, 66 pages. Non public, Délivrable Projet PREDIT MobiVIP, n^{o} D5.5, Inria, May 2007.

[87] B. TROUSSE, B. SENACH. *Retour d'expérience et valorisation*, 17 pages. Non public, Délivrable Projet PREDIT MobiVIP, n^{o} D5.6, Inria, June 2007.

[88] A.-M. VERCOUSTRE, J. THOM, J. PEHCEVSKI. *Entity Ranking in Wikipedia*, Research Report, n^{o} RR-6294, INRIA, 2007, http://hal.inria.fr/inria-00172511/en/.

Miscellaneous

[89] R. KABBAJ. *Mise en correspondance entre appels d'offres et équipes de recherche: mapping bidirectionnel*, Technical report, Faculté des Sciences Sidi Mohamed Ben Abdellah de Fès, Maroc, 2007.

[90] B. SALEH. *Optimizing the Division of Data For Knowledge Discovery*, Technical report, Université de Nice Sophia-Antipolis, 2007.

References in notes

[91] H.-H. BOCK, E. DIDAY (editors). *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*, Springer Verlag, 2000.

[92] P. BLACKBURN, J. BOS, K. STRIEGNITZ. *Learn Prolog Now!*, Texts in Computing, vol. 7, College Publications, 2006.

[93] W. F. CLOCKSIN, C. S. MELLISH. *Programming in Prolog*, 5th edition, Springer Verlag, 2003.

[94] B. CONAN-GUEZ, F. ROSSI, A. EL GOLLI. *Fast Algorithm and Implementation of Dissimilarity Self-Organizing Maps*, in "Neural Networks", vol. 19, n^{o} 6–7, August 2006, p. 855–863, http://hal.inria.fr/inria-00174196.

[95] M. CSERNEL. *Software Requirements Specification for the S.O.M. (Symbolic Object Manipulation)*, November 1997, Deliverable of the WP1 of the Sodas Project.

[96] T. DALAMAGAS, T. CHENG, K.-J. WINKEL, T. SELLIS. *Clustering XML Documents using Structural Summarie*, 2004, In Proc. of ClustWeb - International Workshop on Clustering Information over the Web in conjunction with EDBT 04, Crete, Greece.

[97] F. DE CARVALHO, R. M. C. R. DE SOUZA, M. CHAVENT, Y. LECHEVALLIER. *Adaptive Hausdorff distances and dynamic clustering of symbolic interval data*, in "Pattern Recognition Letters", vol. 27, n^{o} 3, February 2006, p. 167–179.

[98] T. DESPEYROUX. *Developing efficient parsers in Prolog: the CLF manual (v1.0)*, Technical Report, n^{o} RT-0328, INRIA, 2006, http://hal.inria.fr/inria-00120518/en/.

[99] T. DESPEYROUX, B. TROUSSE. *De la sémantique des langages de programmation à la vérification sémantique des sites Web*, in "Journées scientifiques de l'action spécifique Web Sémantiques, CNRS, Paris", October 2002, http://www.lalic.paris4.sorbonne.fr/stic/octobre/programme0209.html.

[100] E. DIDAY, G. GOVAERT. *Classification automatique avec distance adaptatives*, in "R.I.A.R.O Informatique Computer Science", vol. 11, n^{o} 4, 1977, p. 329-349.

[101] M. E. FAYAD, D. C. SCHMIDT. *Object-Oriented Application Frameworks*, in "Communication of the ACM", vol. 40, n^{o} 10, 1997, p. 32-38.

[102] S. FLESCA, G. MANCO, E. MASCIARI, L. PONTIERI, A. PUGLIESE. *Detecting Structural Similarities between XML Documents*, in "WebDB", 2002, p. 55-60.

[103] M. GAROFALAKIS, A. GIONIS, R. RASTOGI, S. SESHADRI, K. SHIM. *XTRACT: a system for extracting document type descriptors from XML documents*, 2000, p. 165–176, http://citeseer.ist.psu.edu/garofalakis00xtract.html.

[104] I. GUYON, S. GUNN, M. NIKRAVESH, L. A. ZADEH. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*, Springer, 2006.

[105] R. J. HATHAWAY, J. W. DAVENPORT, J. C. BEZDEK. *Relational duals of the c-means clustering algorithms*, in "Pattern Recognition", vol. 22, n^{o} 2, 1989, p. 205–212.

[106] M. JACZYNSKI, B. TROUSSE. *Patrons de conception dans la modélisation d'une plate-forme po ur le raisonnement à partir de cas*, in "Revue l'Objet", Numéro Spécial sur les patterns orientés objets, D. Rieu et J-P. Giraudon (guest editors), vol. 5, n^{o} 2, 1999.

[107] R. E. JOHNSON, B. FOOTE. *Designing Reusable Classes*, in "Journal of Object-oriented programming", vol. 1, n^{o} 2, 1988, p. 22–35.

[108] L. KAROUI, M.-A. AUFAURE, N. BENNACER. *Context-based Hierachical Clustering for the Ontology Learning*, in "IEEE/WIC/ACM International Conference on Web Intelligence, Hong-Kong, China", 18-22 December 2006, p. 420-427.

[109] J. A. KONSTAN, B. N. MILLER, D. MALTZ, J. L. HERLOCKER, L. R. GORDON, J. RIEDL. *GroupLens: Applying collaborative filtering to usenet news*, in "Communications of the ACM", vol. 40, n^{o} 3, 1997, p. 77-87.

[110] A. H. LAND, A. G. DOIG. *An Automatic Method for Solving Discrete Programming Problems*, in "Econometrica", vol. 28, 1960, p. 497–520.

[111] Y. LECHEVALLIER. *Scientific technical report for WP6*, juin 2002, IST-2000-25161 WP6/D1.1.

[112] Y. LECHEVALLIER. *Symbolic Objects Methodology*, novembre 2002, IST-2000-25161 WP2.2/D1.1.

[113] W. LIAN, D. W.-L. CHEUNG, N. MAMOULIS, S.-M. YIU. *An Efficient and Scalable Algorithm for Clustering XML Documents by Structure*, in "IEEE Trans. Knowl. Data Eng", vol. 16, n^{o} 1, January 2004.

[114] F. MASSEGLIA, P. PONCELET, M. TEISSEIRE, A. MARASCU. *Web Usage Mining: Extracting Unexpected Periods from Web Logs*, in "Proceedings of the 2nd Workshop on Temporal Data Mining (TDM 2005), held in conjunction with the Fifth IEEE International Conference on Data Mining (ICDM'05), Houston, USA", 27 November 2005, http://www-sop.inria.fr/axis/Publications/uploads/pdf/tdm_icdm_period2.pdf.

[115] A. NAPOLI, ET AL. *Aspects du raisonnement à partir de cas*, in "Actes des 6 èmes journées nationales PRC-GDR Intelligence Artificielle", S. PESTY, P. SIEGEL (editors), Hermes, Paris, mars 1997, p. 261-288.

[116] A. NIERMAN, H. V. JAGADISH. *Evaluating Structural Similarity in XML Documents*, in "Proceedings of the Fifth International Workshop on the Web and Databases (WebDB 2002), Madison, Wisconsin, USA", June 2002, http://citeseer.ist.psu.edu/nierman02evaluating.html.

[117] M. NOIRHOMME-FRAITURE, ET AL.. *User manual for SODAS 2 Software*, version 1.0, FUNDP, Belgique, april 2004.

[118] R. A. O'KEEFE. *The craft of Prolog*, MIT Press, Cambridge, MA, USA, 1990.

[119] P. RESNICK, H. R. VARIAN. *Recommender systems*, in "Communications of the ACM", vol. 40, n° 3, 1997, p. 56-58.

[120] F. ROSSI, B. CONAN-GUEZ. *Functional Multi-Layer Perceptron: a Nonlinear Tool for Functional Data Analysis*, in "Neural Networks", vol. 18, n° 1, January 2005, p. 45–60, http://hal.inria.fr/inria-00000599.

[121] F. ROSSI, B. CONAN-GUEZ. *Un modèle neuronal pour la régression et la discrimination sur données fonctionnelles*, in "Revue de Statistique Appliquée", vol. LIII, n° 4, 2005, p. 5–30, http://hal.inria.fr/inria-00001190.

[122] F. ROSSI, B. CONAN-GUEZ. *Theoretical Properties of Projection Based Multilayer Perceptrons with Functional Inputs*, in "Neural Processing Letters", vol. 23, n° 1, February 2006, p. 55–70, http://hal.inria.fr/inria-00001191.

[123] F. ROSSI, N. DELANNAY, B. CONAN-GUEZ, M. VERLEYSEN. *Representation of Functional Data in Neural Networks*, in "Neurocomputing", vol. 64, March 2005, p. 183–210, http://hal.inria.fr/inria-00000666.

[124] F. ROSSI, D. FRANÇOIS, V. WERTZ, M. VERLEYSEN. *Sélection de groupes de variables spectrales par information mutuelle grâce à une représentation spline*, in "Actes de la conférence Chimiométrie 2005, Villeneuve d'Ascq (France)", November–December 2005.

[125] F. ROSSI, D. FRANÇOIS, V. WERTZ, M. VERLEYSEN. *A functional approach to variable selection in spectrometric problems*, in "Artificial Neural Networks (Proceedings of the 16th International Conference on Artificial Neural Networks, ICANN 2006), Athens, Greece", S. KOLLIAS, A. STAFYLOPATIS, W. DUCH, E. OJA (editors), Lecture Notes in Computer Science, vol. 4131, Springer, September 2006, p. 11–20.

[126] F. ROSSI, A. LENDASSE, D. FRANÇOIS, V. WERTZ, M. VERLEYSEN. *Mutual information for the selection of relevant variables in spectrometric nonlinear modelling*, in "Chemometrics and Intelligent Laboratory Systems", vol. 80, n° 2, February 2006, p. 215–226, http://hal.inria.fr/inria-00174077.

[127] U. SHARDANAND, P. MAES. *Social Information Filtering: Algorithms for Automating Word of mouth*, in "CHI'95: Mosaic of creativity, Denver, Colorado", ACM, May 1995, p. 210-217.

[128] D. TANASA, B. TROUSSE. *Data Preprocessing for WUM*, in "IEEE Potentials", vol. 23, n° 3, August 2004, p. 22–25.

[129] A.-M. VERCOUSTRE, J. THOM, J. PEHCEVSKI. *Entity Ranking in Wikipedia*, in "the 23rd Annual ACM Symposium on Applied Computing, Fortaleza, Brazil", to appear, March 2008, http://hal.inria.fr/inria-00189149/en/.

[130] A. WEXELBLAT, P. MAES. *Using History to Assist Information Browsing*, in "Proceedings of the RIAO'97 Symposioum: Computer-Assisted Information Retrieval on the Internet, Montreal, Canada", June 1997.

[131] T. W. YAN, M. JACOBSEN, H. GARCIA-MOLINA, U. DAYAL. *From user access patterns to dynamic hypertext linking*, in "Computer Network and ISDN systems", (proceedings of the 5th international WWW conference), vol. 28, mai 1996, p. 1007-1014.

[132] J. YI, N. SUNDARESAN. *A classifier for semi-structured documents*, in "KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA", ACM Press, 2000, p. 340–344.