INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team AxIS

User-Centered Design, Improvement and Analysis of Information Systems

Sophia Antipolis - Rocquencourt

THEME COG

Table of contents

1. Team **1**
2. Overall Objectives **2**
2.1. Objectives 2
3. Scientific Foundations **4**
3.1. Introduction 4
3.2. Semantics and Design of Document-Based Information Systems 4
3.3. Information Systems Data Mining 4
3.3.1. Usage Mining 5
3.3.1.1. Data selection and transformation 5
3.3.1.2. Data mining: extracting association rules 6
3.3.1.3. Data mining: discovering sequential patterns 6
3.3.1.4. Data mining: clustering approach to reduce the volume of data in data warehouses 6
3.3.1.5. Data mining: reusing usage analysis experiences 7
3.3.2. Content and Structure Document Mining 7
3.4. Supporting Information Retrieval 8
3.4.1. Design of Adaptive Recommender Systems 8
4. Application Domains **10**
4.1. Panorama overview 10
4.2. Evolving Hypermedia Information Systems 10
4.3. Transportation Systems 11
4.4. Tourism 11
5. Software **12**
5.1. Introduction 12
5.2. CLF -Computer Language Factory 12
5.3. AxISLogMiner 12
5.4. SODAS 2 Software 13
5.5. Clustering Toolbox 13
5.6. CBR*Tools 14
5.7. Broadway*Tools 14
5.8. Ralyx 14
5.9. BibAdmin 15
6. New Results **15**
6.1. Introduction 15
6.2. Data Transformation and Knowledge Management in KDD 16
6.2.1. Feature selection 16
6.2.2. Viewpoint Managment for Annotating a KDD Process 16
6.2.3. Cluster Interpretation Process Metamodel based on a Clustering Ontology 17
6.2.4. Knowledge Base For Ontology Learning 19
6.2.5. Comparison of Sanskrit Texts for Critical Edition 19
6.3. Data Mining Methods 20
6.3.1. Partitioning Methods on Interval Data 20
6.3.2. Self Organizing Maps on Dissimilarity Matrices 20
6.3.3. Functional Data Analysis 21
6.3.4. Visualization 21
6.3.5. Sequential Pattern Extraction in Data Streams 21
6.3.6. Agglomerative 2-3 Hierarchical Classification (2-3 AHC) 22
6.4. Web Usage and Internet Usage Mining Methods 24
6.4.1. Dynamic Clustering of Web usage Data For Charactering Visitors groups 24
6.4.2. Crossed Clustering in Web Usage Mining 24

6.4.3. Discovering Generalized Usage Patterns: the GWUM method 25
6.4.4. Mining Interesting Periods from Web Access Logs 25
6.4.5. P2P Usage Mining 27
6.4.6. Web Usage Mining for Ontology Evolution 27
6.4.7. Web Site Analysis based on an Ergonomic and Web Usage Mining Approach 28
6.5. Document Mining and Information Retrieval 29
6.5.1. XML Document Mining 29
6.5.2. Entity Extraction From XML Documents 30
6.5.3. Document Mining for Scientific and Technical Watch 30
6.5.4. Web HTML Pages Clustering For Ontology Construction 31
6.5.5. Formal Concept Analysis and Semantics for Contextual Information Retrieval 31
6.5.6. Web Pages Mining for Improving Search Engines 32
7. Contracts and Grants with Industry **32**
7.1. Industrial Contracts 32
7.1.1. EPIA: a RNTL Project (2003-2007) 32
7.1.2. MobiVIP: a PREDIT Project (2004-2007) 33
7.1.3. Eiffel "E-tourism and Semantic Web": a RNTL Project (2006-2009) 35
7.1.4. Industrial Contacts 35
8. Other Grants and Activities **36**
8.1. Regional Initiatives 36
8.1.1. "Pôles de compétitivité" 36
8.1.2. Other initiatives 36
8.2. National Initiatives 37
8.2.1. CNRS Action Concertée Incitative: "Histoire des savoirs" 37
8.2.2. EGC Association: National Group on Mining Complex Data 37
8.2.3. SFDS association: InfoStat Group 38
8.2.4. GDR-I3 38
8.2.5. Other Collaborations 38
8.3. European Initiatives 39
8.3.1. EuropeAID Project: For Archaeology of Ancient Asian Texts (AAT) 39
8.3.1.1. The objective of the AAT 39
8.3.1.2. Contributions to program 39
8.3.2. Other Collaborations 40
8.4. International Initiatives 40
8.4.1. Australia 40
8.4.2. Brazil 41
8.4.3. Canada 41
8.4.4. China 41
8.4.5. India 41
8.4.6. Morocco 41
8.4.7. Romania 42
8.4.8. Tunisia 42
8.4.9. Other Collaborations 42
9. Dissemination **42**
9.1. Promotion of the Scientific Community 42
9.1.1. Journals 42
9.1.2. Program Committees 43
9.1.2.1. National Conferences/Workshops 44
9.1.2.2. International Conferences/Workshops 44
9.1.3. Organization of Conferences or Workshops 44
9.1.4. AxIS Web Server 45

9.1.5. Activities of General Interest 45
9.2. Formation 45
9.2.1. University Teaching 45
9.2.2. H.D.R and Ph.D. Thesis 46
9.2.3. Internships 47
9.3. Participation to Workshops, Conferences, Seminars, Invitations 48
10. Bibliography **48**

1. Team

Team Leader

Brigitte Trousse [Research Scientist (CR1), Inria Sophia Antipolis]

Team Vice-Leader

Yves Lechevallier [Research Scientist (DR2), Inria Rocquencourt]

Administrative Assistant

Stéphanie Aubin [TR INRIA, Inria Rocquencourt]
Sophie Honnorat [AI INRIA, part-time, Inria Sophia Antipolis]

Research Scientists

Thierry Despeyroux [Research Scientist (CR1), Inria Rocquencourt]
Florent Masseglia [Research Scientist (CR1), Inria Sophia Antipolis]
Fabrice Rossi [Research Scientist (CR1), on secondment, Inria Rocquencourt, HdR]
Bernard Senach [Research Scientist (CR1), Inria Sophia Antipolis]
Anne-Marie Vercoustre [Research Scientist (DR2), Inria Rocquencourt]

Partners

Mireille Arnoux [Assistant Prof., Univ. Bretagne Occidentale, Inria Sophia Antipolis]
Marie-Aude Aufaure [Assistant Prof., Supélec Gif-sur-Yvette, Inria Rocquencourt, HdR]
Marc Csernel [Assistant Prof., Univ. Paris IX Dauphine, Inria Rocquencourt]

Postdoctoral Fellows

Zeina Jrad [Univ. Evry Val d'Essone, Eiffel project, since Nov.1st, Inria Rocquencourt]
Jovan Pehceski [RMIT, Melbourne, Australia, since Dec. 1st, Inria Rocquencourt]

Ph.D. Students

Abdourahamane Balde [Univ. of Paris IX Dauphine, Inria Rocquencourt]
Hicham Behja [France-Morocco Cooperation (STIC-GL network), Univ. Hassan II Ben M'Sik, Casablanca, Morocco, Inria Sophia Antipolis]
Sergiu Chelcea [Univ. Nice Sophia Antipolis (UNSA-STIC), Inria Sophia Antipolis]
Alzennyr da Silva [Univ. Paris IX Dauphine, from October 1st, Inria Rocquencourt]
Alice Marascu [Univ. Nice Sophia Antipolis (UNSA-STIC), Inria Sophia Antipolis]

Technical Staff

Ghuilaine Clouet [Research engineer, MobiVip project, since July 1st, Inria Sophia Antipolis]
Mohamed Sémi Gaieb [Research engineer, Epia project, since July 1st, Inria Sophia Antipolis]
Christophe Mangeat [Research engineer, MobiVip project, since Oct. 1st, Inria Sophia Antipolis]
Doru Tanasa [Research engineer, EPIA project, until Jul. 31, Inria Sophia Antipolis]

Visiting Scientists

Guy Cucumel [Prof., Univ. of Québec,Montréal, July 4-august 5, Inria Rocquencourt]
Renata Cardoso De Souza [Prof., Federal Univ. of Pernambuco, Brazil, June 21-30, Inria Rocquencourt]
Franciso De Carvalho [Prof., Federal Univ. of Pernambuco, Brazil, September 15-October 4, Inria Rocquencourt]
Ghislain Lévesque [Prof., UQAM Montréal, Canada, June 14, Inria Sophia Antipolis]
James Thom [Associate Prof., RMIT, Australia, November 27 - December 1, Inria Rocquencourt]
Kenneth Reed [Associate Prof., University of Melbourne, Australia, June 24- July 28, Inria Rocquencourt]

Student Interns

Mustapha Eddahibi [Univ. Cadi Ayyad Marrakech, Morocco, SARIMA program, October, Inria Sophia Antipolis]
Eduardo Fraschini [Univ. Of the Republic, Uruguay, May-Sept., Inria Rocquencourt]
Mounir Fegas [Univ. Paris Sud XI LRI, April, Inria Rocquencourt]
Saba Gul [MIT, until January, Inria Rocquencourt]

Reda Kabbaj [Faculté des Sciences Sidi Mohamed Ben Abdellah, Fès, Morocco, since July 17, Inria Sophia Antiplis]
Nicomedes Lopes Cavalcanti Junior [Federal University of Pernambuco, Brazil, until March, Inria Rocquencourt]
Cyrille Maurice [FUNDP Namur, Belgique, Sept-Jan., Inria Rocquencourt]
Jean-Nicolas Turlier [IIE Evry, Feb.-July, Inria Rocquencourt]

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*, *information retrieval*, *information system*, *information system evaluation*, *information system validation*, *knowledge discovery*, *knowledge management*, *ontology extraction*, *ontology management*, *recommender system*, *semantics checking*, *tourism*, *transportation*, *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 ultimate 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. We are convinced that to reach this goal, according to the constant evolution of web based ISs, it is necessary to anticipate the usage and the maintenance very early in the design process. Four main applicative objectives are then addressed by the team:

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

To achieve such objectives, we have set up in July 2003 a multidisciplinary team that involves people from different computer sciences domains (Artificial Intelligence, Data Mining & Analysis, Software Engineering, Document management) and recently from Ergonomics, all of them being involved 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 dynamic aspects of ISs i.e. both the design of support tools (information retrieval support tools, recommender systems), the IS use and then the dynamic” analysis (usage mining).
- the knowledge management point of view related to the capitalization of knowledge and experience in the evaluation process of IS: expertise in combining the evaluation results according to different points of view and more general KDD[1] expertise applied on information systems data.

[1] KDD: Knowledge Discovery from Databases

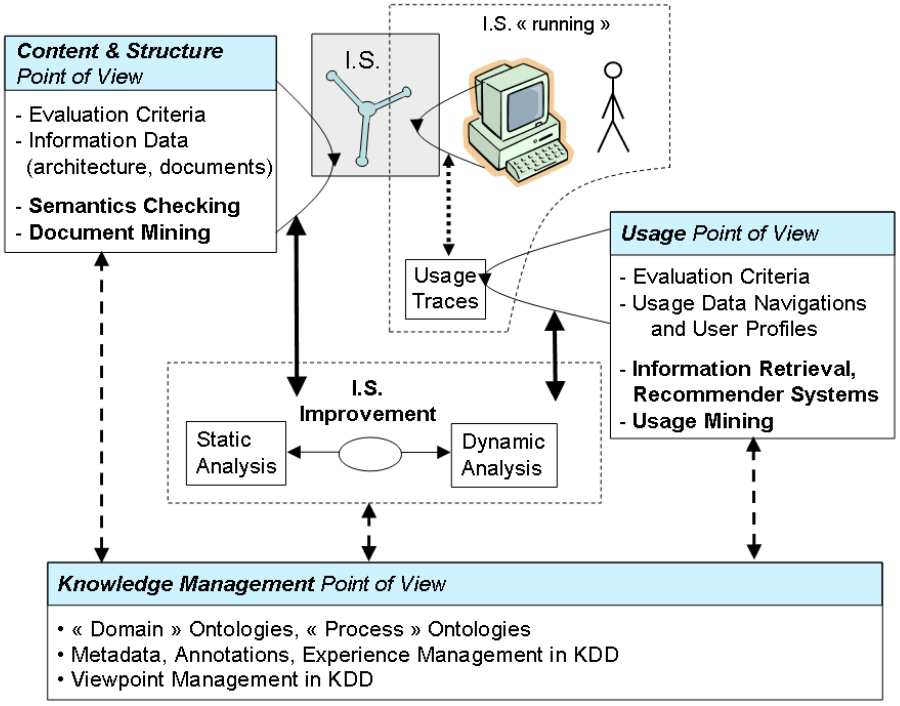

Figure 1. Global View of AxIS Research Topics

3. Scientific Foundations

3.1. Introduction

This section details several 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 technics on usage data (cf. section 3.3.1) ?
- How to synthesize and exhibit information by applying KDD technics 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) ?

The second and third questions concern “information systems data mining”.

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 is multi-disciplinary, involving Software Engineering and Artificial Intelligence techniques. There is a strong relation between structured documents (such as Web sites) and a program; 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), to Web sites modelling and Web services (UML), but not to the formal semantics of Web sites 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 navigation in a Web site which 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 concerned first by 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

The main motivations of usage mining in the context of ISs or search engines are double:

- 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 to 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, 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**.

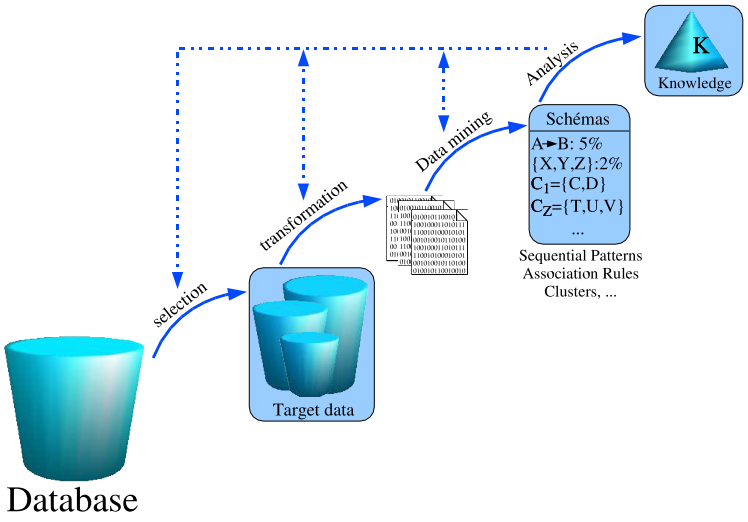

Figure 2. Steps of the KDD Process

More precisely the first third 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 in 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 or from 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 individual 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 a priori 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 from a few thousands to a few millions of individuals and hundreds or thousands of variables. Now, 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 splitting 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 an approach based on case-based reasoning. Very fast algorithms already exist 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.

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 ([90], [102], [83], [98]).

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 models [115], 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 [89], 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 those 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 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 researches for supporting information retrieval are mainly related to personalization and the following topics:

- use and/or construction of user profiles (cf. the projects EPIA (cf. section 7.1.1) and Eiffel (cf. section 7.1.3);
- sophisticated interfaces as in the Eiffel project (cf. section 7.1.3);
- query interface and ranking criteria in the context of search engines (cf. section 6.5.6);
- 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 [109], [105], [94].

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 minimizing the effort of producers and the consumers?

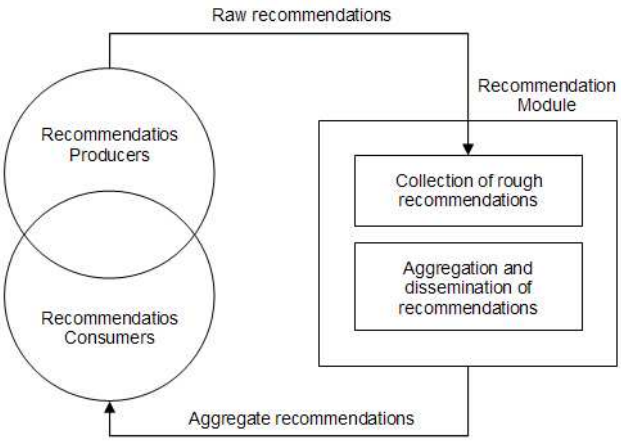

Figure 3. Architecture of a Recommender System

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) [105].

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 [113] and the system of Yan et al. [114].

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 [101] 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* *e-CRM*, *Transportation*, *adaptive interface*, *adaptive service personalization*, *e-business*, *e-marketing*, *information retrieval*, *web design*, *web usage mining*.

The project addresses any applicative field

- on design, evaluation and improvement of a 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 technics 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 services in the collaborative search of information and personalization (ranking, filtering, addition of links, etc.);
2. frequent evolution of the content (information, ontology), generating many maintenance problems, for example:
   - 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) resulting from the Web or a company, according to a thematic criteria, eg. search engines (Yahoo, Voila), Internet guides for specific targets (FT Educado) or portals (scientific communities).

3. interpretation of the user satisfaction (according to the designer point of view) or explicit user satisfaction, 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),
- ontology construction and evolution,
- re-designing an information system based on usage analysis,
- re-designing an information system based on web mining (usage, content and structure),
- updating an ontology based on web mining (usage, content and structure),
- adaptive recommender systems for supporting information retrieval, Collaborative search of Information 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, Transports, 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 in story management: 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. An 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 partner 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 6.4.7). Cooperation about car-sharing has also been initiated this year with CASA.

Let us note also past work published this year on a hybrid clustering approach to approximate fastest paths on urban networks [18].

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. If only for economical reasons, the quality and the diversity of tourism packages have to be improved, for example by highlighting the 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. sections 6.4.6, 6.5.4) and 6.5.5 have been carried out using log files from the city of Metz. This city was chosen because their Web site keeps developing and has been rewarded 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 log files and to evaluate what could be the benefice in designing or updating a tourism ontology (cf. section 6.4.6).

AxIS is also interested in providing users whith transport 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: CLF for the development of efficient parsers, AxISLogMIner for web usage mining, SODAS 2 and Clustering Toolbox for Clustering, CBR*Tools for knowledge management and Broadway*Tools for designing adaptive Web-based recommendation systems, 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].

The initial goal of the Computer Language Factory (CLF) was to enhance the trilogy of formalisms Metal/PPML/Typol used in Centaur [76] to support prototyping the syntax and the semantics of computer languages. The Centaur system was a syntactic editor, written in Lisp, and was able to call external modules as parsers or semantic tools that was specified using specific formalisms. However, with the premature end of the development of the Centaur sytem, this goal was not completely achieved.

The current version of CLF has been rebuilt to permit a fast and easy developpement of efficient parsers in Prolog, including XML parsers. 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 [80], [104], [75] to perfom syntactical analysis.

This toolbox has been used to produce a parser for XML. This parser has been extended to produce a specification formalism. The generated parsers have been intensively used in our team to parse and analyse XML files, mainly related to our research applied to the Inria annual activity reports (cf. section 6.5.1).

A complete documentation is valaible in [63].

5.3. AxISLogMiner

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

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

AxISLogMiner Preprocessing is a software application that implements our preprocessing methodology for Web Usage Mining [111]. 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. Recently, we extensed this software with the ability of recording the keywords employed by users in search engines to find the browsed pages.

The keywords extracted from the http referrer field can therefore be associated with the Web pages and used to build a dissimilarity matrix for those Web pages. Furthermore, this allows extracting clusters of similar pages in terms of content filtered through users' views (keywords). The results of such clustering were used in the experiments conducted in the GWUM work [59] (cf. section 6.4.3).

5.4. SODAS 2 Software

Participants: Yves Lechevallier [correspondant], Marc Csernel.

The SODAS 2 Software [103] is the result of the European project called "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 [73].

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. It uses the SOM (Self Organizing Map) library [81].

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 (PDF format), can be downloaded at

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

5.5. Clustering Toolbox

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

For clustering, we maintained a clustering toolbox, written in C++ and Java, that includes several clustering methods developed by the team over time, and uses the SOM library developed by M. Csernel. This library offers a common data interface to every algorithm. This toolbox supports developers in integrating various classification methods, and testing and comparing with other methods. Currently it integrates several methods:

- from AxIS Rocquencourt: 1) a partitionning clustering method on complex data tables called SCluster [103], 2) Div (in C++) [78], 2) a java library that provides efficient implementations of several SOM variants, especially those that can handle dissimilarity data called DSOM [86],[20] which is available on Inria's Gforge server http://gforge.inria.fr/projects/somlib/ (cf. section 6.3.2) and 3) a functional Multi-Layer Perceptron Method called FNET for the classification of functional data [106];
- two partitionning clustering methods on the dissimilarity tables issued from a collaboration between AxIS Rocquencourt team and Recife University, Brazil: 1) CDis (in C++) [103] and 2) CCClust (in C++);
- 2-3 AHC (in Java)[64] from AxIS Sophia Antipolis which is available as a Java applet which runs the "hierarchies visualisation" toolbox.

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

5.6. CBR*Tools

Participants: Sergiu Chelcea, Sémi Gaieb, Brigitte Trousse [correspondant].

CBR*Tools is an object-oriented framework [93], [88] 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). We showed that, for each application, the thorough expertise necessary to use CBR*Tools is based on only 20% to 40% of the hot spots thus validating the assistance brought by our platform on design as well as on the implementation, thanks to the re-use of its abstract architecture and its components (index, similarity).

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/.

5.7. 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 current contracts: EPIA (cf. section 7.1.1 and MobiVip (cf. section 7.1.2).

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. Xyleme stores the XML version of the activity reports and supports queries to the reports involving both their structure and their content. The first objective was to offer a tool for browsing the activity report with the same interface as the legacy HTML version. However, thanks to Xyleme, pages and links are no longer statics, but computed on the fly, which can offer more flexibility in the future (different styles of display, editing and updating along the year, etc). The next step was to design and implement some views that could meet different users' expectations (Research scientists, teams, INRIA departments, INRIA managers, etc). Each view is delivered in HTML but also in XML, and possibly in text for import into Excel; Ralyx will be put in production for the Raweb 2006 (February 2007).

The idea behind Ralyx is to exploit XML documents the same way one can exploit data in a database, i.e. by querying the documents and assembling the answers (parts of the initial documents) into a new document. The approach works 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.

Morever to access fine granularity information embedded into the text we need to use or develop more advanced techniques based on natural language processing.

5.9. BibAdmin

Participant: Sergiu Chelcea [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: http://gforge.inria.fr/projects/bibadmin/.
BibAdmin is used by AxIS for its Web server.

6. New Results

6.1. Introduction

Keywords: *KDD*, *annotation*, *data transformation*, *dissimilarities*, *distances*, *knowledge management*, *metadata*, *ontology*, *preprocessing*, *reusability*, *viewpoint*.

This year we obtained original results as previous years in our four research topics: data transformation and knowledge management, data mining, Web usage and Internet mining and document mining and Information retrieval.

Let us note new researches this year for supporting ontology construction and evolution (cf. sections 6.2.4, 6.5.4 et 6.4.6) and on information visualization in data mining (cf. section 6.3.4). Let us also note that some previous works described in our 2005 annual report have been published this year on "Dissimilarities for Web usage Mining" ([55], [54]) and on XML Document mining (cf. section 6.5.1). More an hybrid clustering approach to approximate fastest paths on urban networks has also been published this year [18].

First on data transformation and knowledge representation (cf. section 6.2), we pursued our researches on feature selection (cf. section 6.2.1) and on critical edition of sanskrit texts (cf. section 6.2.5). We studied also the use of metadata (cf. the KM point of view), in particular in two ongoing PhD thesis related to semantic web and KDD, conducted by H. Behja and A. Baldé. Ontologies and metadata have been used 1) for annotating global KDD processes in terms of viewpoints to support the management and the reuse of past KDD experiences (cf. section 6.2.2), 2) for supporting the interpretation of extracted clusters with the definition of an ontology and an interpretation model this year (cf. section 6.2.3).

Secondly on data mining methods (cf. section 6.3), we published new results on a new partitioning dynamic clustering method (cf. section 6.3.1), on self organizing maps (cf. section 6.3.2), on functional data analysis (cf. section 6.3.3) and on an agglomerative 2-3 Hierarchical Clustering in the context of Chelcea'PhD thesis (cf. section 6.3.6). This year we pursued actively the research topic started in 2005 related to mining data streams in the context of Marascu's PhD thesis (cf. section 6.3.5) and started a new research topic on Visualization (cf. section 6.3.4).

Thirdly on information systems data mining and more precisely on usage mining, we pursued our researches on mining Web user visits via applying in an original way dynamic clustering (cf. section 6.4.1) and crossed clustering (cf. section 6.4.2). We proposed also five original methods this year:

- the GWUM method for extracting generalized usage patterns (cf. section 6.4.3),
- a method for mining interesting periods from Web Access Logs (cf. section 6.4.4),
- an approach based on a genetic-inspired algorithm for improving resource searching in a dynamic and distributed database such as a P2P system (cf. section 6.4.5),
- a method based on usage mining for supporting the evolution of a Web site ontology (cf. section 6.4.6,
- and finally 5) a method based on Ergonomics and WUM for analysing a Web site (cf. section 6.4.7).

Finally we pursued our researches on XML or HTML document mining and its applications such as the exploitation of a large collection of XML documents (cf. section 6.5.1), ontology construction (cf. section 6.5.4), scientific and technical watch (cf. section 6.5.3) or the improvement of information retrieval based on contextual aspects or ranking criteria (cf. section 6.5.6). Our researches aimed more precisely clustering or classifying XML documents based on their structure and content (cf. section 6.5.1), entity extraction from XML documents (cf. section 6.5.2), document mining for scientific and technical watch (cf. section 6.5.3), clustering HTML pages (cf. section 6.5.4) and contextual information retrieval (cf. section 6.5.5).

6.2. Data Transformation and Knowledge Management in KDD

6.2.1. *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 [92]. 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.

We have been working since 2004 [97] on the application of feature selection methods to spectrometric data. This type of data introduces specific challenges as they correspond to a small number of spectra (a few hundred) described by a very high number of correlated spectral variables (up to several thousands). We have shown how a recently proposed high dimensional estimator of the mutual information [95] could be used, together with a forward backward search procedure, to select relevant spectral variables in non linear regression problems [25].

We have also started to combine our work on functional data analysis with our feature selection research (see section 6.3.3 for details).

6.2.2. *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 integrating this notion in a multi-views Knowledge Discovery from Databases (KDD) analysis. We define a viewpoint in KDD as the analyst perception of a KDD process which is referred to its own knowledge. The KDD process implies various kinds of knowledge, which makes it complex. Our purpose is to facilitate both reusability and adaptability of a KDD process, and to support this complexity via storing past analysis viewpoints. The KDD process will be considered as a view generation and transformation process annotated by metadata related to the semantics of a KDD process.

In 2004 and 2005, we started with an analysis of the state of the art and identified three directions: 1) the use of the viewpoint notion in the Knowledge Engineering Community including object languages for knowledge representation, 2) modelling a KDD process adopting a semantic web based approach and 3) annotating a KDD process. Then we designed and implemented 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.
This year we propose a new metadata format to annotate the KDD process in order to reuse the analysis of experts based on their preferences and formalized by the "viewpoints" analysts. Secondly, in order to facilitate the management and the use of our scheme in a complete KDD analysis, we propose an object-oriented framework that integrates specializable "viewpoints" and reusable components. 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.

6.2.3. Cluster Interpretation Process Metamodel based on a Clustering Ontology

Keywords: *Dublin Core*, *PMML*, *RDF*, *XQuery*, *clustering's interpreting*, *metadata*.

Participants: Abdourahamane Baldé, Yves Lechevallier, Brigitte Trousse, Marie-Aude Aufaure.

This work was conducted in the context of A. Baldé's Ph.D.
The main goal of this thesis is to help end-users to interpret, automatically, the results of their clustering methods. Thus, this work addresses the last step of KDD process (post processing) and its anticipation from the data mining step.

In 2005, we designed this process by using metadata model as one solution to this problem. hen, we implemented this model with the Dynamic clustering algorithm (*SClust*) developed in the AxIS project.

In 2006, we began by using our approach in the Weka Software [2] [28] and we have shown that using our approach is very helpful in this context.

Based on our previous study [28], this year we propose a new metamodel, based on our clustering ontology (cf. Figure 4) for cluster interpreting process. A meta-model is an explicit model of the constructs and rules needed to build specific models within a domain of interest. The main interest of this metamodel is to explicitly explain the main concepts used in the interpretation domain, and the relationships between them.

This metamodel, mainly inspired from and based on the Common Warehouse Metamodel proposed by (*CWM*), allows us to elaborate the automatic interpretation process. Then, we defined some interpretation scenarios in our tool. Our minimal ontology of clustering domain helps us within the definition of these scenarios. This ontology is constructed using the *Protégé2000* software.

[2] *Weka* is a collection of machine learning algorithms for data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Figure 4. Ontology of the Clustering Domain

We experimentally validated this new approach on Weka and on Sclust Algorithm.

Our main contributions can be summarized as follows:

- construction of a clustering ontology, supporting the automation of the interpretation process. This ontology is used to define various interpretation scenarios,
- creation and implementation of a metamodel based on this ontology,
- extension of our metadata architecture based on the *Saxon* processor.

6.2.4. 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.

We proposed two approaches for ontology construction from web pages. The first one (cf. section 6.5.4) is based on a contextual and incremental clustering of terms, while the second one designs a knowledge base for learning ontologies from the web.

Indeed the objective of our second approach is to build a knowledge base for ontology learning from web pages [30]. This knowledge base is specified using a metaontology. This metaontology contains 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 ontologies: the domain ontology, the structure ontology and the services ontology [29]. These components are interrelated. For example, the relation between the domain ontology and the service ontology is useful to determine the set of concepts and relations identifying each service. Our ontology learning approach is based on the synthesis of the research work in this field. A prototype has been developed and experiments have been realized in the tourism domain (cf. section 4.4).

6.2.5. Comparison of Sanskrit Texts for Critical Edition

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

Participants: Marc Csernel, Jean-Nicolas Turlier, Yves Lechevallier.

These results have been obtained in the context of the EuropeAid AAT project (cf. section 8.3.1) and the CNRS ACI action (cf. section 8.2.1). Our objective was to compare around 50 versions of the same text copied by hand along the centuries. During that duration 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 arbitrary 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 given below:

- 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 are transformed according to the *sandhi*.

The expected results are expressed as an *edit distance*, in terms of words, instead of the usual string diff: the sequence of words that are added, deleted, replaced from the *pādapāthā* to obtain the text of the manuscript.

In 2005, after addressing the graphic problems related to sanskrit, we developed an HTML interface for a critical edition of Sanskrit texts, and we made the first step of the processing of Sanskrit manuscripts.
This year we focusedmostly on the comparison of Sanskrit texts.

The comparison is done according to the following steps:

- A parser makes the two versions homogeneous
- The comparison is made letter by letter, using the algorithm of the Longest Common Subsequence (L.C.S), to determine which are the words in the *mātrikāpathā*. The separation between the words of the *pādapāthā*, are used as a pattern for this determination.
- Once the L.C.S completed, we can not examine all the possible results provided, because their number is enormous , 10^{10} is quite common and can be frequently oversized.
- The strategy developed is a navigation through the L.C.S matrix associated with some rules based on the common sense.
- The rules based on common sense are quite simple, such as *"two words are not considered as replacing each others if they don't have at least 50% of letters in common"*.

The results [65], which have been obtained without specific Sankrit knowledge, are quite good, according to some Sankrit philologists. They are due to some common sense rules with some specific algorihmic methods.

6.3. Data Mining Methods

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

6.3.1. Partitioning Methods on Interval Data

Keywords: *Quantitative Data*, *dynamic clustering algorithm*, *unsupervised clustering*.

Participants: Marc Csernel, F.A.T. de Carvalho, Yves Lechevallier, Rosanna Verde, Renata Souza.

In the publication of a special issue on interval data edited by F. Palumbo [19] we propose a survey of the partionning methods of interval data. They use different homogeneity criteria as well as different kinds of clusters representation (prototypes). For the first two methods we introduce some tools to interpret the final partitions. Finally the methods are compared and corroborated on a real data set.

This year we progressed our research on adaptive distances. An article was published [21]. The main contribution is the proposal of a new partitional dynamic clustering method for interval data based on the use of an adaptive or a global Hausdorff distance at each iteration. 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. The advantage of this approach is that the clustering algorithm recognizes different shapes and sizes of clusters. Here the adaptive distance is a weighted sum of Hausdorff distances. Explicit formulas for the optimum class prototype, as well as for the the weights of the adaptive distances, are found. When used for dynamic clustering of interval data, these prototypes and weights ensure that the clustering criterion decreases at each iteration.

6.3.2. Self Organizing Maps on Dissimilarity Matrices

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

Participants: Fabrice Rossi, Nicolas Lopes, Yves Lechevallier.

In 2006, we have continued our previous work on the adaptation of the Self Organizing Map (SOM) to dissimilarity data (the DSOM). We have in particular improved the quality of our software implementation available on INRIA's GForge (cf. section 5.5).

Our previous work on the DSOM and of its applications in collaboration with ex-members of AxIS i.e. A. El Golli and B. Conan-Guez has been published in various journals and conference [20], [22], [34].

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 2006, we have continued our work on Support Vector Machines (SVMs) for functional data analysis. We have shown in particular how some specific spline based kernels can be use to define SVMs on the derivatives of the input functions, without calculating explicitly those derivatives. We showed that the SVMs defined this way are consistent (i.e., they can reach the Bayes error rate asymptotically) [62], [27].

We have also started to combine our work on feature selection (see section 6.2.1) with our work on functional data analysis in collaboration with the DICE laboratory (Belgium, Louvain). One of the limitations of the method proposed in [25] is its computational cost, related to the high number of original spectral variables. We have investigated in [107] and [56] how a B spline representation of the spectra can be used to reduce the number of features prior to the application of the feature selection method studied in [25]. We use the locality of the B spline representation to preserve interpretation possibilities. The new features, i.e. the coordinates of the spectra on a B splines basis, are obtained from limited ranges of the original spectral band: a spectral interval can be associated to each selected feature and used for interpretation purpose.

We have in addition started to work on the application of FDA to time series prediction in [40]. We applied a general idea from [74] in which a time series is splited into sub-series. Each sub-series is considered as a function, which leads to a function value time series. The resulting series is predicted via an autoregressive model. In our approach, Radial Basis Function networks are used to represent the functions and a functional Least Square Support Vector Machine is used to implement the autoregressive model.

In 2006, our earlier works on functional neural methods made in collaboratiuon with N. Villa from GRIMM-SMASH team (Université Toulouse Le Mirail) have been published in international journals [24], [26].

6.3.4. Visualization

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

Participant: Fabrice Rossi.

In 2006, we conducted two surveys on information visualization [57], [58]. The first one [57] outlines the important relationships between machine learning and information visualization, while the second survey [58] is dedicated to the usage of visualization methods for metric studies (such as bibliometrics, for instance). Metric studies provide challenging non vector data, generally large graphs with different types of nodes and links. While we have not applied our work on self organizing map for dissimilarity data [20] to metric studies, this is a promising research topic.

6.3.5. Sequential Pattern Extraction in Data Streams

Keywords: *data stream*, *sequential pattern*.

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 2005 ([99],[51]) 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 execution time and meanwhile the quality of the results. To this end, we have proposed the SCDS (Sequence Clustering in Data Streams) method [49], [50], [23]. To summarize this method, we cut the data stream in batches of a same size and we process the batches one by one. For each batch, at the very beginning, we place 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 ς_{c_j} the centroid of cluster c_j.

The general idea of this method is illustrated in Figure 6.
We needed to define: 1) A computation method of the centroid of a cluster; 2) A similitude measure between a sequence and a centroid; 3) An update method, performed after adding a new sequence to a cluster.

The centroid of a cluster is found thanks to an alignment method. Let's consider the following cluster:

Centroid
(a:3, b:2, c:1, d:1):4 (e:4):4 (h:2, i:2, j:1):3 (m:4, n:2):4
Sequences
<(a,c) (e) (m,n)>
<(a,d) (e) (h) (m,n)>
<(a,b) (e) (i,j) (m)>
<(b) (e) (h,i) (m)>

Figure 5. Cluster

As illustrated in the figure 5, the centroid of a cluster is the result of an alignment method applied to the sequences contained in that cluster. Because this alignment process is often applied we have optimized the method with an incremental alignment based on a sort of the sequences applied in real time. In fact, the quality of the alignment depends on the order of the sequences. This order thus has to be maintained in real time.

All those steps have to be performed as fast as possible in order to meet the constraints of a data stream environment. Approximation has been recognized as a key feature for this kind of applications, explaining our choice for an alignment method for extracting the summaries of clusters. The dynamic feature of data streams imposes an execution time constraint, but, in meantime, we must assure a good quality of results. To this aim, we have performed some quality tests.

SCDS 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.

6.3.6. Agglomerative 2-3 Hierarchical Classification (2-3 AHC)

Keywords: *2-3 HAC*, *AHC*, *evaluation*, *stress*.

Participants: Sergiu Chelcea, Brigitte Trousse, Yves Lechevallier.

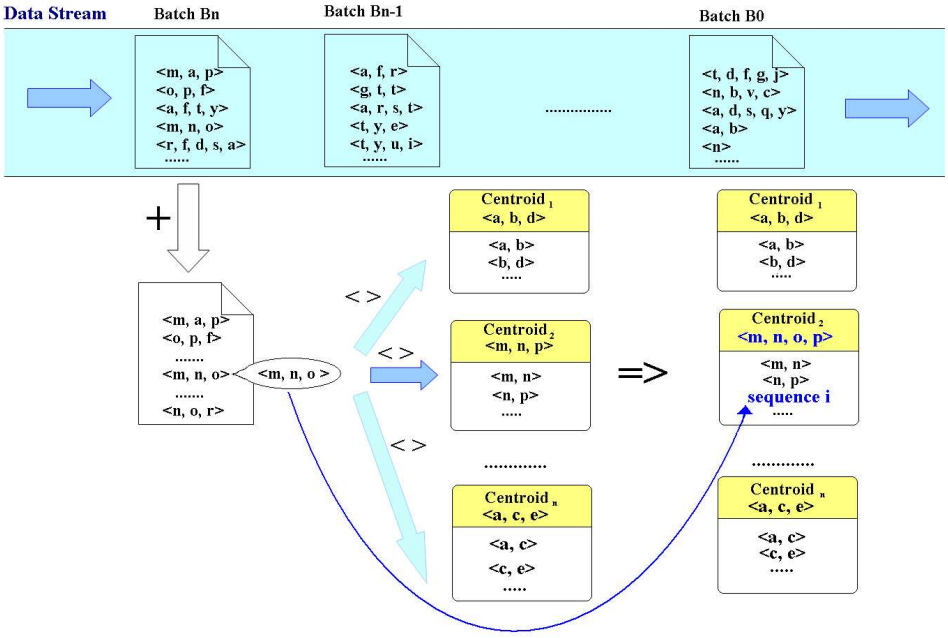

Figure 6. The SCDS method

This work has been done in the context of S. Chelcea's thesis. The past years, we proposed a new Agglomerative 2-3 Hierarchical Classification general algorithm in collaboration with P. Bertrand (ENST B) and four 2-3 AHC algorithm variants which can create different 2-3 hierarchies on the same dataset according to the use or not of the blind merging. A previous theoretical complexity analysis of our 2-3 AHC algorithm proved that the complexity was reduced from O(n^3) in the initial 2-3 AHC algorithm to O($n^2 log(n)$) for our algorithm.

This year, in collaboration with J. Lemaire (IUT Menton University of Sophia Antipolis), we pursued our tests on the obtained 2-3 hierarchies and the classical hierarchy on different datasets (Ruspini, urban itineraries, simulated data, Abalone) for complexity execution times and structure quality. The obtained execution times verified our theoretical complexity of O($n^2 log(n)$). To determine the created structures quality, we have chosen the Stress coefficient for comparing the initial data and the induced dissimilarity matrices. Using the complete link, we obtained an average gain of 23% (for the Stress) while the maximum gain was around 84% on the Abalone dataset.

Moreover, we finalised our study of the applicability of our 2-3 AHC method in two fields: Web Mining and XML Document Mining.. For Web Mining field, we found that the 2-3 AHC produced interesting results, richer than the classical AHC and better than the AxIS ones obtained with another method [87]. For XML document mining, we applied our 2-3 AHC algorithm on the INRIA activity reports. One objective was to compare different 2-3 AHC algorithms using as reference the classical AHC one: we found that the best results are obtained with the 2-3 AHC algorithm avoiding the blind merging (V3), which was the only one to always have a positive Stress gain compared to the classical AHC [64]. For applicative results, see the section 6.5.1.

6.4. Web Usage and Internet Usage Mining Methods

6.4.1. Dynamic Clustering of Web usage Data For Characterizing Visitors groups

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

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

The analysis of a web site based on its usage data is an important task as it provides insight into the organization of the site and its adequacy regarding user needs. Such knowledge is especially interesting for business applications. In this context, analyzing such data can help organizations, among other things, to plan cross marketing strategies and effectiveness of promotional campaigns. We thus defined an approach for discovering the profiles of visitor groups. To this purpose, we map user interests into symbolic objects which represent a user's successful interaction with the site. Symbolic Objects constitute the bases of the Symbolic Data Analysis (SDA). The general aim of this analysis is to extend the processing of classical data types to support more complex data. In conventional datasets, the objects are individualized, whereas in symbolic datasets they are unified by means of relationships. In our proposition, we identify groups of users with similar behaviour by means of a dynamic clustering algorithm which applies a context dependent dissimilarity measure defined by Francisco De Carvalho. The benchmark data set consists in a one-year log file coming from the web site of the CIn (Informatics Centre of UFPE, Brazil). Our approach was capable to identify the profiles of distinct typologies of users based on their navigational preferences. Although the method was carried out to identify visitor groups of an educational web site this approach is generic enough to be applied on any other domain. The results of our experiments were published this year in two international conferences [35], [36].

6.4.2. Crossed Clustering in Web Usage Mining

Keywords: *contingence table*, *crossed clustering algorithm*, *unsupervised clustering*, *web usage mining*.

Participants: Alzennyr Da Silva, Yves Lechevallier, Sergiu Chelcea, Doru Tanasa, Brigitte Trousse.

The emergence of new information technologies such as the Word Wide Web had for consequence the explosion of the amount of data. The necessity for summarizing these data has thus become obvious. In this context, we proposed an approach to automatically build homogeneous classes from these data and to define new statistical units to describe them. By reducing the initial amount of data, the summarization results contain a maximum of information. This kind of problem is addressed in the Web Usage Mining framework. Our approach is based on the crossed clustering method whose objective is to obtain simultaneously a row partition and a column partition from a contingency table. This method represents an effective solution for both search of a typology of individuals (represented by the lines of the table) and the construction of a taxonomy on variables values (represented by the columns of the table). As a result, we identify dominant groups of users as well the sets of pages visited by each group. One of the goals of this analysis is to better understand users' behaviour and for consequence to propose changes in the web site organization in order to better serve the users. We applied our proposition on the Web log data provided by the IT centre of UFPE (Recife, Brazil) [66] and also on the Web log data registering access on seven different e-commerce Web sites from the Czech Republic [33].

6.4.3. Discovering Generalized Usage Patterns: the GWUM method

Keywords: *Generalization*, *WUM*, *sequential pattern*.

Participants: Doru Tanasa, Florent Masseglia, Brigitte Trousse, Yves Lechevallier.

This work [46], [59], [110] proposes an original method for Web usage analysis based on a user-driven generalization of Web pages. The information extracted for these pages, for the clustering purpose, regards the users' access to the pages. The information is obtained from the referrer field of the Web access logs when the user employed a search engine to find the page. The main idea is to characterize a Web page by the keywords that have been given to a search engine in order to find this page. For instance, if most of the accesses to the Web page about job opportunities in the AxIS team ("/axis/jobs-sop.htm") come from a search engine with the keywords "Position" and "Internship", then this page may be generalized by (characterized with) the keywords "Position,Internship". This principle of generalization in illustrated in figure 7.

Then the traditional data mining step will not be applied to the Web pages but on their generalization. The experiment that we carried out illustrates our methodology and shows some of the benefits obtained with such an approach in the discovery of frequent sequential patterns. These benefits consist in obtaining generalized patterns with a higher support and easier to interpret.

6.4.4. 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 [100]. 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). In order to better understand our goal, let us consider student behaviours when they are connected for a working session. Let us assume that these students belong to two different groups having twenty students. The first group was connected on 31/01/05 while the other one was connected on 01/02/05, (*i.e.* the second group was connected one day later). During the working session, students have to perform the following navigation: First they access URL "www-sop.inria.fr/cr/tp_accueil.html", then "www-sop.inria.fr/cr/ tp1_accueil.html" which will be followed by "www-sop.inria. fr/cr/tp1a.html".

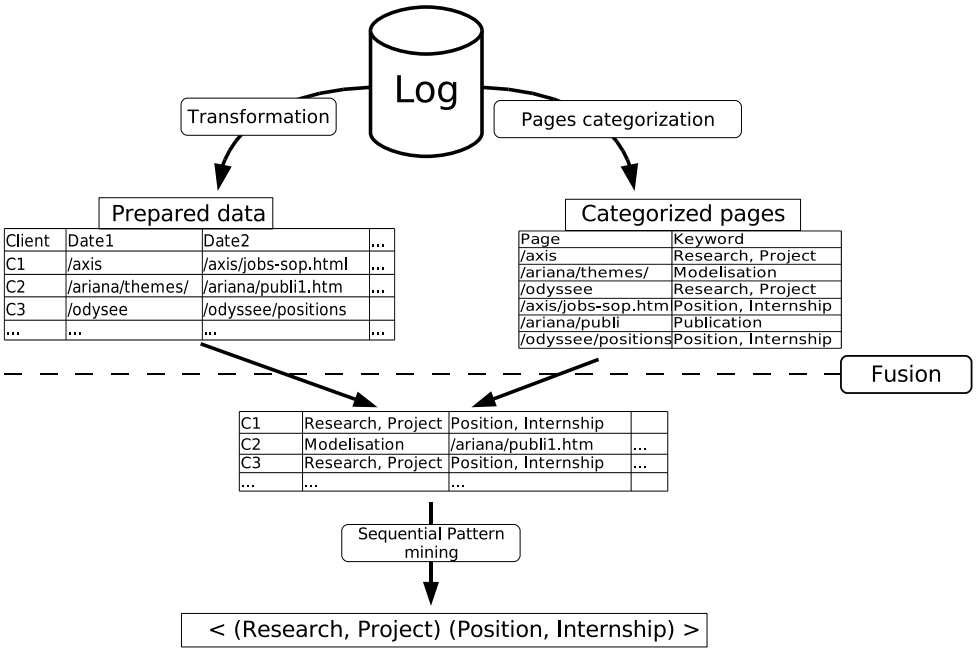

Figure 7. The generalization method in GWUM

Let us consider, as it is usual in traditional approaches, that we analyze access logs per month. During January, we only can extract twenty similar behaviours, among 200 000 navigations on the log, sharing the working session. Furthermore, even when considering a range of one month or of one year, this sequence of navigation does not appear sufficiently on the logs (20/20000) and will not be easy to extract. Let us now consider that we are provided with logs for a very long period (*e.g.* several years). With our method, we can find that it exists at least one dense period in the range [31/01-01/02]. Furthermore, we know that, during this period, 340 users were connected. We are thus provided with the new following knowledge: 11% (*i.e.* 40 on 340 connected users) of users visited consecutively the URLs "tp_accueil.html", "tp1_accueil.html", and finally "tp1a.html". The outline of our method [52] 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). Let us now consider $perioResult$ the result obtained by running the method presented in this paper. We want to minimize $\sum_{i=0}^{size(perioResult)} S_i/S_i \neg realResult$ (with S_i standing for a frequent sequence in $perioResult$), as well as maximize $\sum_{i=0}^{size(realResult)} R_i/R_i \in perioResult$ (with R_i standing for a frequent sequence in $realResult$). In other words, 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).
We have conducted some experiments and 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.

6.4.5. P2P Usage Mining

Keywords: *Peer-2-Peer*, *genetic algorithms*, *sequential patterns*.

Participant: Florent Masseglia.

With the huge number of information sources available on the Internet, Peer-to-Peer (P2P) systems offer a novel kind of system architecture providing the large-scale community with applications for file sharing, distributed file systems, distributed computing, messaging and real-time communication. P2P applications also provide a good infrastructure for data and compute intensive operations such as data mining.
In [53] we have proposed a new approach for improving resource searching in a dynamic and distributed database such as an unstructured P2P system. This approach takes advantage of data mining techniques. By using a genetic-inspired algorithm, we propose to extract patterns or relationships occurring in a large number of nodes. Such a knowledge is very useful for proposing the user with often downloaded or requested files according to a majority of behaviors. It may also be useful in order to avoid extra bandwidth consumption. For instance, it may be discovered, in a P2P file sharing network, such as Gnutella [91], that "Mandriva Linux 2005" distribution is often downloaded as "CD1.iso, then CD2.iso and finally CD3.iso".

We consider that the connected nodes can act with a special peer (a "*meter peer*") in order to provide the end user with a good approximation of patterns embedded in this very large distributed database. To evaluate our approach, we implemented a simulator capable of running simulated unstructured P2P system. Experiments were also conducted by using real datasets.

6.4.6. Web Usage Mining for Ontology Evolution

Keywords: *ontology evolution*, *ontology mangement*, *tourism*, *web usage mining*.

Participants: Brigitte Trousse, Marie-Aude Aufaure, Yves Lechevallier, Florent Masseglia.

This year we propose in collaboration with B. Legrand (LIP6) an original approach for ontology management in the context of Web-based information systems. Our approach relies on the usage analysis of the considered Web site, in complement to the existing approaches based on content analysis of Web pages. Our methodology is based on knowledge discovery techniques mainly from HTTP Web logs and aims at confronting the discovered knowledge in terms of usage with the existing ontology in order to propose new relations between concepts.

We illustrate our approach on a Web site provided by French local tourism authorities (related to Metz city) (cf. section 4.4) with the use of clustering and sequential patterns discovery methods. One major contribution of this work is thus the application of usage analysis to support ontology evolution and/or web site reorganization.
Such a work has been accepted for publication as a chapter of a book [112].

6.4.7. Web Site Analysis based on an Ergonomic and Web Usage Mining Approach

Keywords: *Web site*, *ergonomics*, *evaluation*, *web usage mining*.

Participants: Bernard Senach, Brigitte Trousse.

Web Usage Analysis are often realized from different points of view and with exclusive techniques. For instance, considering web usage, the log analysis of a site is rarely related to the ergonomic analysis of this site (and conversely). The MobiVIP project (cf. section 7.1.2) has been an opportunity to set up a new methodology coupling the ergonomic approach with the technical log analysis. The study [70], [71] has been conducted on a transportation web site used to consult various information about a bus network (lines' structure, geographical information, time tables): URL http://www.envibus.fr.

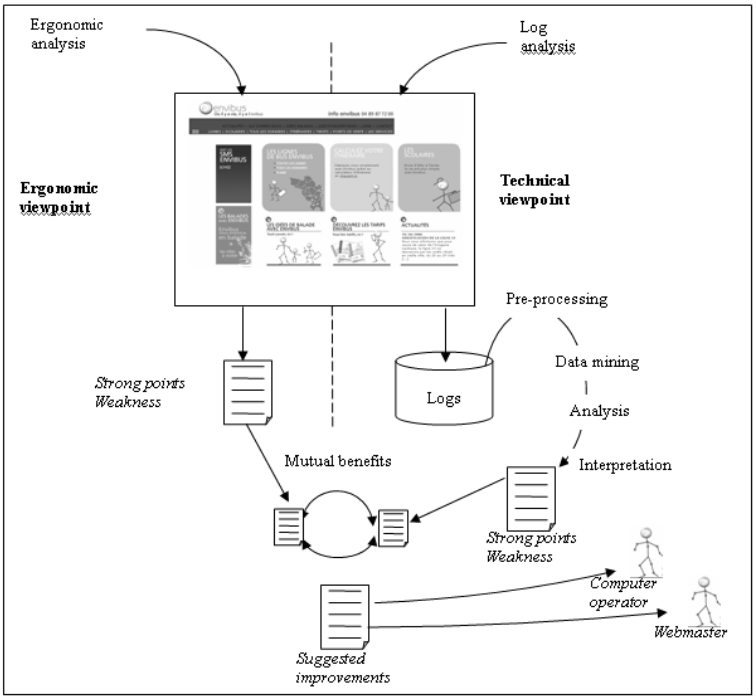

Figure 8. Steps of our Ergonomic and WUM based Web Site Analysis

The illustration in Figure 8 sums up the different steps which have been followed:

A "discount usability" technique has first been used to point out potential users' difficulties linked for instance to a low structural or graphical user interface consistency. The suspected problem drove a specific log analysis and in some case, it was possible to find out in the usage data patterns confirming the hypothesis. For instance, to be used efficiently, some decision aids in the envibus required a topographical knowledge of the area, and it was suspected that this could be a reason to give up during a transaction. The log analysis showed that this assumption was correct as the ratio of interrupted request was very high on the corresponding pages. An important benefit of the coupling is also that suggested improvements given to the user interface designer are much more powerfull as more information can be provided and quantitative data enforces the qualitative analysis.

6.5. Document Mining and Information Retrieval

6.5.1. XML Document Mining

Keywords: *Classification*, *Clustering*, *Data Mining*, *INEX*, *Mining Complex Data*, *XML Document*, *XML mining*.

Participants: Thierry Despeyroux, Yves Lechevallier, Anne-Marie Vercoustre, Sergiu Chelcea, Florent Masseglia, Brigitte Trousse.

XML documents are becoming ubiquitous because of their rich and flexible format that can be used for a variety of applications. Standard methods have been used to classify XML documents, reducing them to their textual parts. These approaches do not take advantage of the structure of XML documents that also carries important information.

First we made in 2005 two types of researches related to XML Document mining which are published this year:

- First we developed a new representation model for clustering XML documents and described our approach working both for structure-based clustering and Structure-and-Content clustering. Results from several experiments using the INEX [3] IEEE collections [60] and INRIA activity reports were published this year [61].
- Second we proposed an original supervised classification technique for XML documents which is based on a linearization of the structural information of XML documents and on a characterization of each cluster in terms of frequent sequential patterns. Experiments on the MovieDB collection validated the efficiency of our approach [37].

Finally we have also contributed to a book chapter on XML document mining that has been accepted and will be part of a book on "Data Mining Patterns: New Methods and Applications" [77].

Moreover as in 2005 [85], we studied the impact of selecting different parts (sub-structures) of XML documents for specific clustering tasks. Our approach integrated techniques for extracting representative words from documents elements with the 2-3 HAC algorithm (cf. section 6.3.6) for classifying documents. We evaluated with the same collection of Inria XML activity reports (year 2003) used previously. Our first objective was to study the impact on the resulted 2-3 hierarchies of selecting different parts (sub-structures) of XML documents and of using two distances then to compare them with INRIA's research organization. As results, we showed that the quality of clustering strongly depends on the selected document features as in [85], and also on the used distance. Indeed with the classical euclidian distance on the words frequency we had a lot of atypical teams as the epidaure team in Figure 9 due, for example, to the heavely usage of the word "imaging" in its presentation. We noted that the Jaccard distance allowed us to reduce the influence of too general frequent words (such as the words "applications", "computer"). Our second objective was to compare the 2-3 AHC algorithms gains with the classical AHC ones on this application: for more details, see section 6.3.6.

[3] Initiative for the Evaluation of XML Retrieval

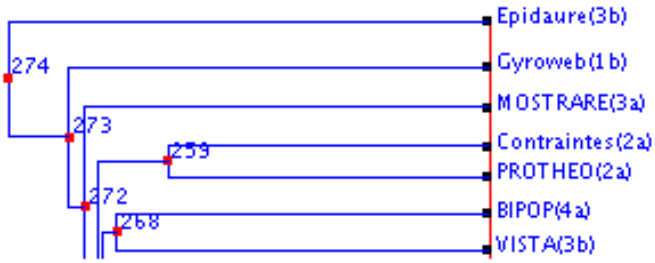

Figure 9. Atypical teams in T-P experiment (i.e. based on their "Overall Objectives" section)

6.5.2. Entity Extraction From XML Documents

Keywords: *Entity Extraction*, *Wrapping*, *named entities*, *semantic annotations*.

Participants: Anne-Marie Vercoustre, Thierry Despeyroux, Eduardo Fraschini.

In order to improve the reliablility and accessibility of document-based information systems, we need to develop tools that involve both the structure of the documents (beyond the DTD model, or the actual tree structure) and the content, i.e. the textual part. We have contributed to the development of http://Ralyx.inria.fr/, a system that uses the structure of the INRIA Activity Reports to provide multiple views on these XML reports and supports the exploitation of this rich information source, both by internal and external parties. However those reports contain mostly text that is difficult to exploit without further document mining. We need to extend their textual content with specific and semantic annotations that can be used both in validation tools and in more specific queries.

This year we experimented with extracting entity names (organisation names) from these activity reports. In this task the XML structure was only used to select specific parts of the reports (e.g. the "contract" section) in order to identify and extract our partners organisations.

Our approach is inspired by the techniques for extracting information from regular web pages (wrapper induction) and does not require natural language ressources, large collections for training, nor manual tuning. The main idea is to use a small list of known organisms (e.g. extracted from a few documents), and to use them as a seed to generate patterns of extraction for new names. We have generated only very generic patterns based on the syntagms of the language as identified by standard taggers. The patterns were trained on a subset of documents for which the entities had been manually extracted. Half of the training set was used to validate the approach with standard recall and precision measures. Using such generic patterns, we could not expect a very high precision, but we were able to extract many new names that were not known in advance. The experiments and the results have been accepted for publication [84]. The next directions to explore will be

- pattern generation without training,
- weighting the patterns both locally (within one document) and globally,
- dynamically improve the initial list of entities by reusing the list provided the previous year.

6.5.3. Document Mining for Scientific and Technical Watch

Keywords: *XML Document*, *clustering*, *mapping*, *scientific watch*.

Participants: Reda Kabbaj, Mustapha Eddahibi, Bernard Senach, Brigitte Trousse.

Research Institutes are more and more involved in applying for grants and supports. As the number of wickets increases, they require more and more resources in watching "calls for tender" to identify current opportunities and route them to the relevant research teams. There is no currently much support tool for this task and calls for tender may be routed to the wrong research teams or, inversely, a competent team may not receive an invitation to tender, therefore missing an opportunity. It is the aim of our "Mapping bidirectionnel AO-ER" system to provide such a tool. Our methodology relies mainly on a bottom up approach to classify documents and qualify research teams, based on text mining. We will explore the use of ontologies to improve our approach. The main contributions are:

- the use of text mining to describe research teams and calls to tender;
- a mapping method based on classification;
- a generic architecture independent of specific data;

The system is based on an automatic classification (supervised and non supervised) of two types of documents: DDAO ("Documents décrivant les Appels d'Offres") are the textual descriptions of the invitation to tender, DDER ("Documents décrivant les Equipes de Recherche") are the textual description of the research team The system, which relies on the K-means algorithm, offers four components:

- a pre-processing module has an information selection mechanism which can represent specific terms of the domain,
- an indexation module provides a data structure to rapidly access to DDAO et DDER,
- a knowledge module represents the knowledge extracted from the documents,
- a mapping module associates research teams and calls to tenders (in both directions) and allows classification.

Initially, the system will be applied to ANR's calls for tender and then extended to other calls for grants.

6.5.4. Web HTML Pages Clustering For Ontology Construction

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

Participant: Marie-Aude Aufaure.

We proposed two approaches for ontology construction from web pages. The first one is based on a contextual and incremental clustering of terms described here, while the second one (cf. section 6.2.4) designs a knowledge base for learning ontologies from the web.

Our first approach defines and evaluates a context-based clustering algorithm for ontology learning (COCE algorithm) included in a global architecture for knowledge discovery for the semantic web [41]. 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. It is deduced from the various analysis included in the pre-processing step (structural and linguistic analysis). 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 cooccurents selection for each word. By performing an incremental process and by recursively dividing each cluster, the COCE algorithm refines the context of each word cluster and improves the conceptual quality of the resulting clusters and consequently of the extracted concepts. The COCE algorithm offers the choice between either an automatic execution or an interactive one with the user. We experiment the contextual clustering algorithm on HTML document corpus related to the tourism domain (in French) and we evaluate the extracted ontological concepts with our contextual algorithm [42]. 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.

6.5.5. Formal Concept Analysis and Semantics for Contextual Information Retrieval

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

Participant: Marie-Aude Aufaure.

In this work, we define an information retrieval methodology which uses Formal Concept Analysis in conjunction with semantics to provide contextual answers to Web queries [44], [96]. 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 pre-treatment providing the global conceptual context and then in an online contextual processing of users requests, associated to an instantaneous context. The pre-treatment consists in computing a conceptual lattice from tourism Web pages in order to build an overall conceptual context. Each concept of the lattice corresponds to a cluster of Web pages with common properties. A matching is performed between the terms describing each page and a thesaurus about tourism, in order to label each concept in a standardized way. Whereas the processing of tourism Web pages is achieved offline, the information retrieval is performed in real-time: users formulate their query with terms from the thesaurus. This cluster of terms is then compared to the concepts labels and the best-matching concepts are returned. Users may then navigate within the lattice by generalizing or on the contrary by refining their query.

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.

6.5.6. Web Pages Mining for Improving Search Engines

Keywords: *Web*, *query formulation*, *ranking criteria*, *search engine*.

Participants: Thierry Despeyroux, Yves Lechevallier, Florent Masseglia, Bernard Senach, Doru Tanasa, Brigitte Trousse, Anne-Marie Vercoustre.

Motivated by our work in 2005 in the context of the e-Mimetic project (confidential status), we pursued some researches in collaboration with E. Boutin from LePont laboratory of the University South Toulon and M. Nanard from the IHMH team of LIRMM. Our goal was to define and evaluate new Web pages ranking criteria based on page presentation. For this we used a test collection of Web pages returned by different search engines in response to a specific set of queries. The pre-processing and clustering tasks allowed us to make some methodological propositions to solve some emerging sub-problems such as:

- criteria to automatically identify the language of a web page (french/english);
- trails to qualify a web page according to its presentation.

7. Contracts and Grants with Industry

7.1. Industrial Contracts

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

Participants: Semi Gaieb, Yves Lechevallier, Bernard Senach, Doru Tanasa, 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, and started on September 2003 until june 2007. Partners are Dalkia, Ever(Mediapps) and Inria.

The objectives of this project are the following:

- Supporting users of Mediapps.Net (tool for selecting canal information of an extranet) via clustering clients. This task started in 2004 and some generic algorithms and pre-processing tools were developed until the beginning of 2005. Some log analysis haven't been done because of the unavailability of real data.
- After understanding the user needs for Net.Portal (construction tool for intranet portals), we finished the specification of the trace of the NetPortal engine (cf. the first version of the deliverable D3: "Experimental context and trace engine in Net.Portal".). The result of this work is the description of the Net.Portal relational database schema and the data organization. The specification of the Net.CanalRecommender was stopped and studied in the new context of the eversuite software.

The project was re-oriented mid 2006 in order to take into account Ever wishes and the future integration of Mediapps.net and Net.portal in the eversuite software. This year we studied the trace engine proposed by Ever (written in Java), its use in the context of Net.Portal and the specification of a recommender system for information sources in this context.

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

Participants: Sergiu Chelcea, Christophe Mangeat, Ghuilaine Clouet, Bernard Senach [co-res.p], 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 involves 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 transport. The MobiVIP project will develop 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, transport modeling, evaluation of service;
2. the demonstrations on 5 complementary experimental sites;
3. the evaluation of possible technology transfer. http://www-sop.inria.fr/mobivip/.

In December 2004, we finalized in collaboration with B. Senach (Ergomatic Consultants) the deliverable 5.1 [108] which we coordinate with Georges Gallais (Visa Action, Inria Sophia Antipolis). This deliverable aimed at defining a common generic evaluation scenario and proposed a framework to facilitate the identification of the main evaluation dimensions for each planned test or experimentation. 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. This continuation will allow us to conduct a large scale experimentation which will start in January 2007 in Antibes.

This year we had two main tasks: 1) the preparation of the deliverables 5.2 [69], 5.3 [71] and 2) the work related to the task 5.4 and the VU Log experimentation (cf. the task 5.5).

- The deliverable 5.2 [69] addresses the definition of evaluation criteria. From a questionnaire distributed to MobiVIP' partners, 7 evaluation dimensions where identified and about 150 criteria were specified.
- The deliverable 5.3 [71] addresses the use of an information web site for trip planning. AxIS used a specific methodology to conduct this analysis (cf section 6.4.7): a human-factors evaluation was realized in parallel to a log analysis. Some results [71], [70] of the ergonomic analysis suggested to give a deeper look at the logs: suspected difficulties were confirmed by the log' analysis. This combined study of a site' use provided a better insight of the man-machine interface weaknesses and allowed powerful recommendations to improve site quality and user interactions.

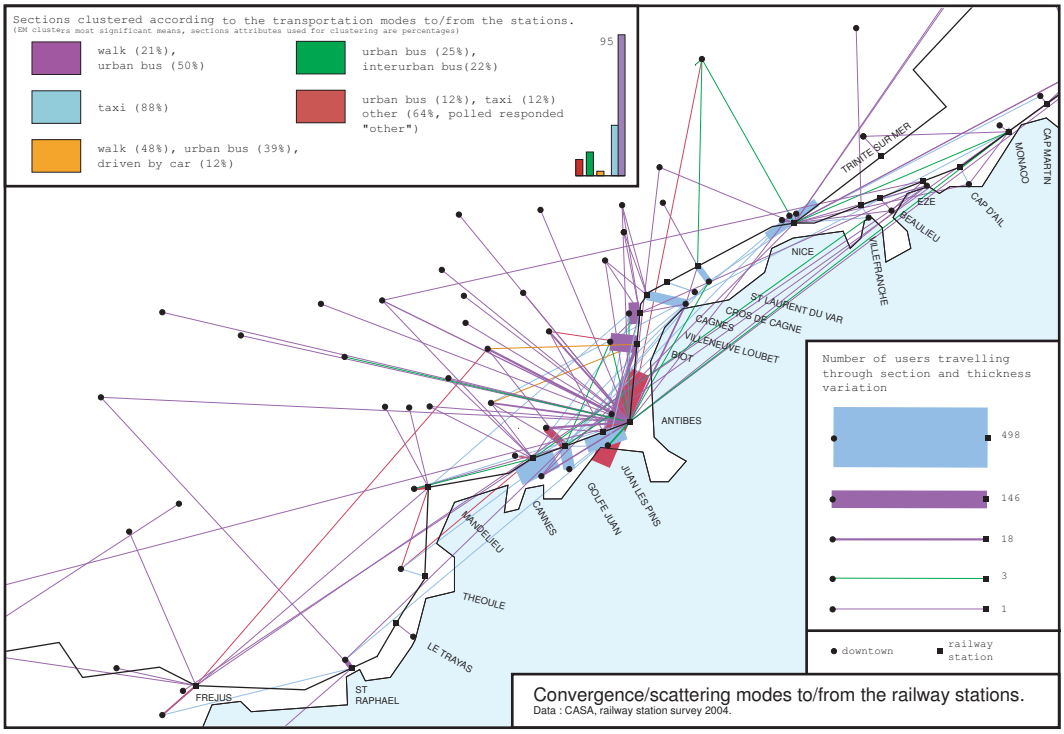

Figure 10. Cartography in CASA

- The task 5.4 addresses the use of data mining technologies to improve travelers' information. Two studies aiming at path classification have been conducted. A first study aiming at defining a semantic distance between paths was conducted in 2005 with simulated data. About thirty routes across a town have been defined. These routes, viewed as potential travels for different professionnal profils (doctors, insurance agent) have been classified with our 2-3AHC algorithm using several similarity distance (Jaccard, Dice) and taking in account attributes such as the number of crossroads, of common parts, proximity of origin. A state of the art on this topic and a review about spatial cognition and mobility, still in progress, will complement the analysis of travelers'behaviours.

  The second one was realized with true data coming from a large survey conducted in railway stations. In the Figure 10, depicted sections represent the traffic of railway users between stations and downtowns. A section thickness represents the amount of users following it and a section color codes the cluster in which the section has been clustered. There are 13 different transportation modes. Section structure is defined by the distribution of transportation mode used by people following the section. Clustering algorithm processes the percentage of each transportation mode used in a section as a section attribute. With EM clustering, each cluster can be described by averaging the different attributes of all its members. With section structures, this averaging process leads to a "mean structure" for each cluster that helps characterizing them.
- We are still working on the design of the Antibes experimentation (cf. the task 5.5) which will have two stages. The first stage will consist of a study of clients' expectations about new transportation services. This study will precise a previous survey (a questionnaire designed by Loria) done during a demonstration which took place in Nancy. In the second phase of the Antibes experimentation, 5 electric cars will be available in the town center, and their use will be precisely analysed from recorded users' actions.

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

Participants: Marie-Aude Aufaure, Zeina Jrad, Cyrille Maurice, Yves Lechevallier [resp.], Bernard Senach, 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. AxIs strongly collaborated with Bénédicte Legrand and Michel Sotto from LIP6.

7.1.4. Industrial Contacts

Some 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).
- Morocco Telecom: In collaboration with Casablanca University and ENSAM (Meknes), AxIS proposed to Morocco Telecom a web usage study. The research goal of the WRUM project (Web Redesign by Usage Mining) is to extend an existing redesign methodology based on site usage mining and to build a plateform with reusable tracking components. In the study, we will consider the use of several sites to trace the users' navigations and to gain inshight of requirements to improve the sites.

- SAP, Sophia Antipolis related to data mining and data streams (security and environnement problems). Contact: B. Trousse and F. Masseglia

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. Senach) participates to the preparation of the ROSCOE project, which aimed at enabling communication between vehicles. Our partners are: Hitachi Europe (leader), VU Log, Nexo, Inria (Mascotte), CNRT Télius, etc.

- "Pôle de compétitivité - Pégase" AxIS (B. Senach and B. Trousse) participates at the preparation of several projects with different partners:
  - CEFH ("Centre d'Etude des Facteurs Humains") aims at building an excellence center of human factors in risk systems. AxIS will be involved in the design and evaluation of decision aids based on usage data mining (partner: Areva TA)
  - In collaboration with Alcatel Alenia Space, AxIS is involved in 3 other projects with a different role in each one
  - Control room design: technical assistance on ergonomic aspects (user's need analysis, man machine interface design and evaluation)
  - Numerical modeling: technical assistance on man machine interface design and evaluation
  - Automatic editing of operational instructions: AxIS provides links between Alcaltel teams and experts in this domain (Areva TA).

  B. Senach and B. Trousse participated to a meeting on December 5th between the leaders of Pégase and different research teams at Inria Sophia Antipolis. AxIS participation to the CEFH project (see above) is required as well as in a large project aiming to design the futur airports. Other participation will be considered according to the forthcoming projects.

- "Pôle de compétitivité" Cap Digital: AxIS has been identified as a potential partner by a special interest group "Numerical Heritage". AxIS could answer to future invitations to tender sent by the Pole "Cap Digital".

- Software: AxIS has began the design of a tool which could facilitate the mapping between Inria teams and thematic invitation to tender sent by Research Institutions (ANR).

8.1.2. Other initiatives

- VISA Action (Inria Sophia Antipolis) collaboration with G. Gallais and P. Rives (VISTA team, Inria Sophia Antipolis), M. Riveill (Rainbow team, I3S UNSA) on the topic "adaptation and evaluation of services in the context of transports" via the MobiVIP project, involving 22 partners (January 2004, December 2006).

- In October, AxIS (B. Senach) participated, as an observer, to the monthly progress review of the ITER project in Cadarache. AxIS captured information about the requirements for audio conference planning when participants are largely scattered around the world (USA, Korean, Japan, India, China, Europe) and about the use of advanced communication tools (shared screens). This action was conducted for Inria "Comorale" workgroup.

- Eurecom, EHESS, Acacia (Inria Sophia-Antipolis): AxIS (B. Senach) is involved in the multidisciplinary workgroup "Usage et Utilisabilité" (B. Conein, EHESS) which stands at Eurecom (social and human sciences, Artificial Intelligence). The aim of this informal workgroup is to discuss research thematics focused on the use of new technologies and to adapt current methodologies to improve their study. The making up of the next seminar (Shared knowledge) is in progress.
- Ilog & Acacia (Inria Sophia Antipolis): AxIS has participated to a meeting of the local workgroup "Usable Intranet" who develops and studies tools for enhancing collaborative work (use of wiki, social tagging).
- Lirmm (Montpellier, M-L Mugnier): AxIS participated to a meeting Inria Sophia Antipolis-Lirmm (december 19) and presented its researches in order to identify potential collaborations.
- Laboratoire des Usages, CNRT Télius, http://www.telius.org, Sophia Antipolis.
- CPER Télius FOCUS: in the framework of a grant between government and regional administration, AxIS proposed the design of a usage mining platform. The FOCUS platform will provide to its "clients" tools and services to get a precise understanding of how people use new technologies. A complete analysis chain is needed in usage mining field to deal with large scale experiments such a ubiquitous computing where a massive flow of temporal data has to be processd on the fly.
- RTRA SISCOM: AxIS participated to a proposal aiming at the creation of a new thematic network (SISCOM) supporting inter-regional cooperation.
- RTRA Digiteo (http://www.digiteo-labs.fr/): AxIS (M-A. Aufaure) participated at the inauguration of Digiteo Labs, the first research pole in information and technology science in Ile de France, October 4th, Polytechnic. The Digiteo teams are involved in two “pôles” System@tique and Cap Digital. Research works in Digiteo are dedicated to intelligent systems design and development.
- Ecole Française d’Extrême Orient (E.F.E.O): In the framework of It Asia contract Francois Patte organized in Pune (india) in collaboration with Marc Csernel (AxIS Rocquencourt) a meeting for the dissemination on the results obtained during the contract.
- Supelec: our partner Marie-Aude Aufaure from Supelec collaborates with others Supelec members [41].

8.2. National Initiatives

AxIS is involved in several national working groups.

8.2.1. 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.

8.2.2. EGC Association: National Group on Mining Complex Data

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:

- B. Trousse with O. Boussaïd (ERIC, Lyon) co-organised and co-chaired the third workshop “Fouille de données complexes dans un processus d’extraction de connaissances” (January 17, 2006) [16]. M-A. Aufaure, F. Masseglia and Y. Lechevallier were members of the program committee.

- F. Masséglia 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”.

8.2.3. SFDS association: InfoStat Group

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

- January 12, Rennes: “Visualisation statistique: du PDA à l'écran géant”.
- March 23, Paris: "Les prévisions ont-elles un avenir ?".
- October 24, Paris: "Classification avec des modèles de mélange: la nouvelle version 2.0 du logiciel libre MIXMOD".[2]

8.2.4. GDR-I3

AxIS participated to three working groups of the GDR-PRC~I3 National Research Group “Information - Interaction - Intelligence” of CNRS:

- Working Group 3.4 (GT) on Data Mining animated by P. Poncelet and J.M. Petit: MA. Aufaure, F. Masseglia, B. Trousse
- GRACQ (*Groupe de Recherche en Acquisition des Connaissances*) (GRACQ): B. Trousse.
- Working Group 3.7 “Sécurité des Systèmes d'Information” animated by D. Boulanger and A. Gabillon: F. Masseglia and B. Trousse.

8.2.5. Other Collaborations

- LePont Laboratory of the South Toulon University (E. Boutin) and LIRMM 5(M. Nanard). Our work was on the pre-processing and clustering of a test collection of Web pages returned by different search engines in response to a specific set of queries (cf. section 6.5.6.
- 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 [44]..
- ENST Paris: Y. Lechevallier collaborated with Georges Hébrail (ENST) [38].
- 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. With P. Bertrand (ENST B) and Annie Morin (TEXMEX, IRISA Rennes) we had a meeting on the possible use of the N-GRAMM for the sanskrit texts comparisons.
- 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 [79],[32].
- Two ARC proposals were submitted: 1) ARC SéSur: “Sécurité et Surveillance dans les data streams” (resp: F. Masseglia) with M.O Cordier (DREAM, IRISA) P. Poncelet (LGI2P, Alès) and M. Teisseire (LIRMM, Montpellier); 2) ARC Valex: “ Vérification et exploitation de collections de documents semi-structurés” (resp: A.-M. Vercoustre) with Annie Morin (TEXMEX, IRISA Rennes), A. Napoli (Orpailleur, INRIA, Nancy), E. de la Clergerie (ATOLL) and B. Sagot (Signes), INRIA Rocq.

- ANR: We have proposed a project in response to the ANR call for proposals on "Masses de Données / Connaissances Ambiantes". The "MIDAS" (MINING DATA STREAMS) project gathered together 6 academic partners: INRIA's team-projects AxIS and DREAM, ENST (Paris), LIRMM (Montpellier), LGI2P (Nimes), GRIMAAG (Martinique); and two industrial partners: EDF R&D and France Télécom R&D. Despite this project has not been accepted, we plan to work on further proposal on this subject.
- GRIMM-SMASH team (Université Toulouse Le Mirail): F. Rossi works with N. Villa on Support Vector Machines and functional data (cf section 6.3.3 and [62], [27]).
- 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) [20], [22], [34] and others topics [24].
- A. Michard (Inria Rocquencourt): we hold different meetings with A. Michard to discuss a service based on modeling the dependency of Information and Organization in order to improve the quality of Entreprise Information Systems. We also discussed the possible application of data mining in anticipating possible breakdown of the system. Such a model was submitted to Bouygues Telecom and France Telecom in the prospect of a future R&D project.

8.3. European Initiatives

8.3.1. EuropeAID Project: For Archaeology of Ancient Asian Texts (AAT)

Participants: Marc Csernel, Sergiu Chelcea, Jean-Nicolas Turlier, Yves Lechevallier, Brigitte Trousse.

Contract Reference: IT ASIA Contract 2004/091-775

In 2005 we started our project called "AAT" in the context of the EuropAid (DG1) projects and more precisely of the Asia Information Technology (I.T. Asia). We collaborated mainly with F. Patte (Ecole Française d'Extrême Orient (E.F.E.O.)) Pascale Haag (EHESS, Centre d'études de l'Inde et de l'Asie du Sud, Paris), M. Le Pouliquen and P. Bertrand (ENST B).

8.3.1.1. The objective of the AAT

Ancient texts, whether religious, scientific or philosophic are known to us due to the patient and vigilant work of scribes who, from centuries to centuries, have copied and copied again successive versions of an original text (usually lost for ever).

So there is a chain of copies starting with the original text and continued by an immense tree of hundreds of copies that has grown more or less like a genealogical tree.

They are never identical to each other, sometimes extremely different. Parts of the original are missing, fragments are not readable anymore, some have been miscopied, and some others have been voluntarily transformed. This is particularly true for the large Indian subcontinent where at least one third of the manuscript existing through the whole world are supposed to exist, mostly unpreserved, unreferenced , and being at mercy of any accidental event. Even during the 20th century manuscripts were copied by hand by armies of scholars.

Still a question remains unsolved as to how to compare hundreds of different copies of a same original ancient text, and to decide which fragments are original and which ones are not in order to re-build the original document.

Specific software has recently been designed for Latin and Greek scripts which open new avenues to study ancient texts from Roman and Hellenistic periods. It is the aim of the present project to design a most advanced IT tool for "archaeology of ancient Asian texts". Such IT Tool will be based strictly on open source.

8.3.1.2. Contributions to program

This project involves AxIS as the applicant of the project and three others partners: University "La Sapienza" in Rome (Facoltà di Studi Orientali), the Bhandarkar Institute of Oriental Studies (BORI) in Poona (India) and the Mahendra Sanskrit University of Kathmandu (Népal).

Our three partners will dedicate their force to the collection of manuscripts of a famous Indian grammatical text: The Kâçikâvritti or “Benares glosses”. This text is the oldest comment (around the 7th century) of the Panini grammar, the world oldest example of generative grammar. It is well known trough hundreds of manuscripts disseminated all around the Indian subcontinent. These manuscripts are dated from the 12th century to the beginning of the 20th century. They are supposed to display the representation of the same text, but because of the time, their completeness is only partially assumed, and they can differ from each other. AxIS is developing all the software to be used in the different steps of the project:

- Providing the software tools necessary to help the creation of critical edition of the Sanskrit texts. As a secondary result, a distance between the texts should be established based on the presence/absence of the different words in each manuscript (cf. section 6.2.5).
- Using the distance established by the first software sets, trying to establish which are the different cluster set of manuscript (for example via a 2-3 AHC clustering), try to establish more or less a phylogeny of the different manuscripts

One could wonder what is the need for a specific project to compare different Sanskrit texts, as tools such as the famous Unix DIFF exist since a long time. The response is given by some of the Sanskrit writing specificities:

- Sanskrit is written according to a 48 letters alphabet, but, on computer, is written using Latin alphabet using a transliteration such as the Velthuis one.
- Sanskrit is written without blank and the blanks are not very significant
- When two words are written without blank separation, the spelling becomes different, it is the sandhi problem.

Four internships were carried out on this project: S. Tandabany (2005), M. Dufresne (2005) S. Kebbache (2005) and J.N. Turlier [72] this year. A closing workshop co-organised by F. Patte and M. Csernel has been done in Pune (India) [68] (cf. section 9.1.3).

8.3.2. Other Collaborations

- Germany: AxIS participated to the project "Core Technology Cluster" of the AII program "QUAERO" in the multimedia domain (ontology construction, perosnalization)
- Italy, University of Napoli II (A. Irpino and R. Verde) [48], [39], [47],
- Italy, University La Sapienza (Roma) Prof Rafaele Torella and Dr Vincenzo Vergiani collaboration in the framework of th IT Asia Project on camparison of Sanskrit manuscripts.
- Belgium, Facultés Universitaires Notre-Dame de la Paix à Namur (Profs A. Hardy, M. Noirhomme and J.-P. Rasson) [103]:Y. Lechevallier.
- Belgium, Université Catholique de Louvain, DICE Laboratory (Prof. Michel Verleysen, Prof. Vincent Wertz, Dr. Amaury Lendasse, Damien François): F. Rossi
- Belgium: a belgium delegation from Namur visited AxIS Sophia Antipolis on december 8th.

8.4. International Initiatives

8.4.1. Australia

In the context of the FEAST program, we collaborated with Ken Redd of the Deakin University on the application of classification techniques to time use data to identify lifestyle differences across societies.

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 Mahalanobis distances suitable to symbolic interval data and they finalized the conception of dynamic clustering models based on Mahalanobis distances suitable to symbolic interval data. A complete paper has been submitted to the "Computational Statistics and Data Analysis" 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 evolves overtime. An application of this approach has been done on data from web usage which evolves on the time. A paper has been accepted for publication at EGC'07 french conference [82].
- Fabrice Rossi visited CIn-UFPE from 04 to 16/12/06. He worked with Renato Correa (a PHD student of prof. Tesesa Ludermir from CIn-UFPE) and Francisco De Carvalho on clustering of documents using SOM based on keywords associated by google and yahoo on the pages of Inria site.
- Francisco de Carvalho and Yves Lechevallier participated to an advanced course on Knowledge Extraction by Interval Data Analysis in Caserta [45] (Belvedere di San Leucio, November 27-29, Italy).

8.4.3. Canada

Y. Lechevallier pursued his collaboration with A. Ciampi (Univ of McGill, Montréal). Guy Cucumel, Professor at the university of Montreal (Canada), visited us during one month at Inria Rocquencourt. He participated to the FEAST projet with Ken Redd.

G. Lévesque, professor at the university of Montréal (UQAM, Canada) visited the team on June 14, Inria Sophia Antipolis.

8.4.4. China

Marie-Aude Aufaure collaborates with Yanwu Yang, Institute of Software Research, Beijing, on user modelling for the semantic web and Yves Lechevallier collaborates with Hueiwen Wang, BUAA, Beijing on clustering methods.

8.4.5. India

M. Csernel collaborated with the Bhandrakar Institute (India) and the Mahendra Sanskrit University (Nepal) (cf. section 8.2.1) and also via the consortium members of EuropeAid projet of Asia-Information Technology and Communications (cf. sections 6.2.5 and 8.3.1).

8.4.6. Morocco

AxIS is involved in a France-Morocco thematic network in software engineering. In this context, B. Trousse co-supervises with Abdelaziz Marzark (University of Casablanca) a Ph.D. student: H. Behja (ENSAM, Meknès, Morocco). H. Behja spent several months with us for his thesis work. Mr. Marzark visited us for a week (july 16 - july 26). A meeting was organised at the University of Mirail on July 20th in order to prepare future France-Morocco collaborations. An R&D cooperation with Casablanca University and ENSAM (Meknes) is in progress (cf. section 7.1.4).

8.4.7. 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. We proposed a collaboration in the context of the Brancusi program on exploiting distribued calculus and grid computing for improving data mining algorithms.

8.4.8. 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.

8.4.9. Other Collaborations

New Zealand: Annika Hinza which leads the Information Systems and Databases Group at the University of Walkato (Hamilton, New Zealand) visited us at Sophia Antipolis on november 9th in order to identify future collaborations related to recommender systems in Tourism and transportation domains.

9. Dissemination

9.1. Promotion of the Scientific Community

9.1.1. Journals

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

- 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.

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

- 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 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.
- the RSTI journal Special Issue on "Viewpoints" (resp: B. Coulette, Univ Mirail): B. Trousse
- RNTI Special Issue "Fouille du Web" (Publisher, Cépaduès Editions): 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 20 international and national journals and for 4 international books:

- JVLC Journal of Visual Language and Computing: M-A. Aufaure
- DKE (Date and Knowledge Engineering): M-A. Aufaure
- AAI (Applied Artificial Intelligence): M-A. Aufaure
- BIT, the International Journal 'Behaviour & Information Technology' (Taylor & Francis Publisher): B. Trousse
- INS: International Journal on Information Sciences: F. Masseglia (http://ees.elsevier.com/ins/)
- ETRI: International Journal on Electronics and Telecommunications Research Institute: F. Masseglia. (http://etrij.etri.re.kr/Cyber/index.html)
- CI: International Journal on Computational Science: F. Masseglia (http://www.blackwellpublishing.com/journal.asp?ref=0824-7935&site=1)
- JIIS: International Journal of Intelligent Information Systems: F. Masseglia (http://www.springerlink.com/content/1573-7675/)
- the Information Systems (IS) Journal: F. Masseglia (http://ees.elsevier.com/is/)
- the Data Mining and Knowledge Discovery (DMKD) Journal: F. Masseglia (https://www.editorialmanager.com/dami/)
- the Journal of Systems and Software (JSS): F. Masseglia (http://ees.elsevier.com/jss/)
- the Data and Knowledge Engineering Journal (DKE): F. Masseglia (http://www.sciencedirect.com/science/journal/0169023X)
- Tehnometrics: F. Rossi (http://www.amstat.org/publications/tech/index.cfm?fuseaction=main)
- IEEE Transactions on Neural Networks: F. Rossi (http://ieee-cis.org/pubs/tnn/)
- Neural Networks (Dec.2005): F. Rossi (http://www.elsevier.com/wps/find/journaldescription.cws_home/841/description#description)
- Transaction on Internet Research: F. Rossi (http://www.internetjournals.net/journals/tir/2006/)
- Computational Statistics: F. Rossi (http://comst.wiwi.hu-berlin.de/)
- Computational Statistics and Data Analysis: F. Rossi (http://www.elsevier.com/locate/csda)
- Neurocomputing: F. Rossi (http://www.elsevier.com/locate/issn/09252312)
- AI Communications: F. Rossi (http://aicom.web.cse.unsw.edu.au/)
- Book - "Web Semantics and Ontology" (Idea Group Publishing): M-A.Aufaure
- Book - Encyclopedia of Data Warehousing and Mining - 2nd edition (idea group): B. Trousse, F. Masseglia (http://frontpage.montclair.edu/wangj/)
- Book - Encyclopedia of Multimedia Technology and Networking (idea group): F. Masseglia (http://www.idea-group.com/reference/details.asp?id=4461)
- Book - Data Mining with Ontologies: Implementations, Findings and Frameworks (idea group): F. Masseglia, B. Trousse

9.1.2. Program Committees

Several AxIS members were involved at national or international conferences/worhshops as member of Program Committee or as additional reviewer. Let us note that we organized two workshops this year (MDM at KDD'06 http://www.fortune.binghamton.edu/MDM2006/, and FDC at EGC'06. (http://www-sop.inria.fr/axis/fdc-egc06/.

9.1.2.1. National Conferences/Workshops

- EGC 2006: Lille, France (Jan. 17-20, 2006) - Y. Lechevallier, B.Trousse (http://www-rech.enic.fr/egc2006/)
- Ateliers EGC 2006: Lille, France (Jan. 17, 2006) (http://www.grappa.univ-lille3.fr/~ppreux/egc2006/ateliers.html)
  - FDC: M-A. Aufaure, Y. Lechevallier, B. Trousse, F. Masseglia (http://www-sop.inria.fr/axis/fdc-egc06/)
  - Visualisation et Extraction de Connaissances: M-A. Aufaure (http://visu.egc.free.fr/EGC06/)
  - Fouille du Web: M-A. Aufaure (http://www.antsearch.univ-tours.fr/fw-egc06/default.asp?FCT=P&ID_PAGE=1)
- SFDS 2006: Clamart, France (May 29 - June 2, 2006) - Y. Lechevallier (http://www.jds2006.fr/)
- SFC 2006: Metz, France (September 5 - September 8, 2006) - Y. Lechevallier (http://www.sfc06.org)
- RàPC 2006: Besançon, France (May 30-31, 2006) - B. Trousse (http://www.lab.cnrs.fr/RaPC2006/organisation.html)
- EDA 2006: Versailles, France (June 19, 2006) - M-A. Aufaure (http://www.prism.uvsq.fr/~eda06/)
- BDA 2006: Lille, France (Oct. 17-20, 2006) - F. Masseglia (http://www2.lifl.fr/BDA2006/)

9.1.2.2. International Conferences/Workshops

- EGC 2006: Lille, France (Jan. 17-20, 2006) - Y. Lechevallier, B. Trousse (http://www-rech.enic.fr/egc2006/)
- ECIR'2006: London, UK (Apr.10-12, 2006) - A-M. Vercoustre (http://ecir2006.soi.city.ac.uk/)
- IEEE ICEIS 2006: Islamabad, Pakistan (Apr. 22-23, 2006) - F. Rossi (http://www.jinnahresearch.net/iceis2006/)
- ESANN 2006: Bruges, Belgium (Apr. 26-28, 2006) - F. Rossi (http://www.dice.ucl.ac.be/esann/)
- CSCWD 2006: Nanjing, China (May 3-5, 2006) - B. Trousse (http://2006.cscwid.org/)
- ECCBR 2006: Fethiye, Turkey (Sept. 4-7, 2006) - B. Trousse ()http://2006.eccbr.org/)
- ICANN 2006: Athens, Greece (Sept. 10-14, 2006) - F. Rossi (http://icann2006.org/chapter1/)
- SYNASC 2006: Timisoara, Romania (Sept. 26-29, 2006) - B. Trousse (http://synasc06.info.uvt.ro/)
- DocEng 2006: Amsterdam, The Netherlands (Oct. 10-13, 2006) - A-M. Vercoustre (http://www.documentengineering.org/)
- ISDA 2006: Jinan, Shandong, China (Oct. 16 - 18, 2006) - D. Tanasa (additional reviewer) (http://isda2006.ujn.edu.cn/)
- WIDM 2006 (ACM): Arlington, Virginia, USA (Nov. 10, 2006) - A-M. Vercoustre (http://workshops.inf.ed.ac.uk/WIDM2006/)
- ICTAI 2006: Washington D.C., USA (Nov. 13-15, 2006) - F. Masseglia (http://www.nvc.cs.vt.edu/ictai06/)
- ADCS 2006: Brisbane, Australia (Dec. 11, 2006) - A-M.Vercoustre (http://www.adcs06.fit.qut.edu.au/)

9.1.3. Organization of Conferences or Workshops

Besides the organization of the workshops FDC/EGC'06 and MDM/KDD'06, we are involved in others organization tasks:

- Organization of a meeting in Pune (India): With the valuable help of Francois Patte (EFEO) a workshop was organized in Pune with the Bhandarkar Institute of Oriental Studies (BORI) on the 17th July for the achievment of the IT Asia project. About fifty people were attending this workshop where the result of the project IT Asia were exposed. This workshop was the occasion for the participants to expose all the work wich has been achieved during the Project. Indian participants speak mostly about the collation of the kachikavritti. M. Csernel made a presentation of the developed software. Speakers: M. Le Pouliquen (ENSTB), S. Bhate(BORI), M. Keskar-Lalit(BORI), P-S Filliozat (French Academy of the humanities), M. Csernel (Inria AxIS), F. Patte (EFEO).
- Organization of a seminar by James Thom, RMIT, Melbourne, Australia, on "Information retrieval evaluation issues" (27th november): Anne-Marie Vercoustre.
- Organisation of our annual AxIS workshop at Inria Rocquencourt (13-15 november): S. Aubin, S. Honnorat, Y. Lechevallier and B. Trousse. Furthermore, monthly team meetings were organised by videoconference between AxIS Sophia Antipolis and 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.
- B. Trousse was expert for evaluating proposals related to the ANR academic research program on Massive Data (ANR “Masse de données - Connaissances ambiantes”) launched in France in 2006.
- 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”. He is involved in three 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 and 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. .
- 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 an associated team for the "STIC Doctoral school" at the University of Nice Sophia Antipolis (UNSA) and the team members are teaching in various university curriculums:

- "Master PMLT" (resp. Mr Kounalis) at UNSA Sophia Antipolis: Tutorial (12h) on *Data Mining and Web Mining*: F. Masseglia, D. Tanasa, B. Trousse.
- "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*: D. Tanasa, B. Trousse.
- "Institut des techniques d'ingénieur de l'industrie" (ITII) (resp. Robert Viani) at ESINSA Sophia Antipolis: Two course modules on *Programming Techniques"* (36h) and on *Algorithmics* (12h): D. Tanasa.
- 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. Tutorial (12h) on *Ontologies and Web Mining*: M-A. Aufaure.
- 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 the 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 recherche Informatique, Paris XI: Tutorial (3h) on *Ontology construction*: M-A. Aufaure.
- "Question Spéciale" on *Introduction au Data Mining* of the Namur University "Facultés Universitaire Notre Dame de la Paix" , Belgium (11-13 June)
- 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

H.D.R defence in 2006:

1. **F. Rossi**, "Contribution à l'analyse des données complexes", November 23, University of Paris-Dauphine

Ph.D. in progress:

1. **S. Chelcea**, (start: end of 2002), "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) (defence planned in 2007)
2. **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).
3. **A. Baldé**, (start: end of 2003), "Extraction de méta-données à partir de prototypes issus d'une classification" (Metadata Extraction from classification prototypes), University of Paris IX Dauphine, (directors: E. Diday and Y. Lechevallier) with the participation of B. Trousse and M.-A. Aufaure (Supelec).
4. **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).
5. **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).

F. Rossi is a member of the thesis committees of **N. Delannay** (start: October 2003) on “Méthodes neuronales pour les données structurées”, and also 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 co-supervised with Mohammed Ben Ahmed the thesis of R. Djedidi (start: end of 2005): “Towards a generic approach for ontology construction from heterogeneous sources”, University Paris XI and University La Manouba (Tunisia).

AxIS researchers were members of H.D.R or Ph.D. committees in 2006:

- Fabrice Rossi, (H.D.R), “Contribution à l’analyse des données complexes”, November 23, Y. Lechevallier.
- Amadou Boubacar Habiboulaye, “Classification dynamique de données non-stationnaires. Apprentissage et suivi de classes évolutives”, June 28, Y. Lechevallier
- Yanwu Yang, “Towards spatial web personalization”, July, M-A. Aufaure
- Alexandre Blansché, “Classification non supervisée avec pondération d’attributs par méthodes évolutionnaires ”, September 28, Y. Lechevallier
- Sylvain Ferrandiz, “Apprentissage supervisé à partir de données séquentielles ”, October 23, Y. Lechevallier
- Marie Plasse, “Utilisation conjointe des méthodes de recherche de règles d’association et de classification. Contribution à l’amélioration de la qualité des véhicules en production grâce à l’exploitation des systèmes d’information”, October 26, Y. Lechevallier
- Nicolas Bonnel, “Génération dynamique de présentations interactives en multimédia 3D, de données, pour des applications en ligne ”, December 4, Y. Lechevallier
- Aliou Boly, “Fonctions d’oubli et résumés dans les entrepôts de données ”, December 15, Y. Lechevallier

9.2.3. Internships

We welcomed four students in AXIS this year:

1. **R. Kabba** (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”.
2. **M. Eddahibi** (supervisor B. Trousse), Ph-D (3rd year), Faculté des Sciences Semlalia, Marrakech, Morocco, “Prétraitement de documents XML relatifs aux rapports annuels des équipes de recherche Inria”.
3. **N. Lopes** (supervisor F. Rossi), Federal University of Pernambuco, Brazil, until March, Inria Rocquencourt. “Participation to the Somlib development”.
4. **C. Maurice** (supervisor Y. Lechevallier), Master 2, (Université de Namur, Belgique),“ Navigation et visualisations graphiques sur PDA”.
5. **J.-N. Turlier** (supervisor M. Csernel) Institut Informatique d’Entreprise (IIE) Evry, "Comparaison de manuscrits Sanskrit en vue d’un édition critique".

The following one was in the context of the Inria international internship program:

1. **E. Fraschini** (supervisors A-M. Vercoustre, T. Despeyroux), University of the Republic, Uruguay), May-Sept.

9.3. Participation to Workshops, Conferences, Seminars, Invitations

Readers are kindly asked to report to the publication references for the participation to conferences with a submission process. Furthermore we attended the following conferences or workshops:

- Gemo seminars from GEMO team (INRIA Futurs, Orsay): Marie-Aude Aufaure
- IC 2006 (Ingéniérie des connaissances), Nantes, June 2006: M-A. Aufaure
- EDA 2006 conference (datawarehouse and on-line analysis), Versailles, June 2006: M-A. Aufaure
- DigiteoLabs inauguration day, October the 4th, Polytechnic, Palaiseau: M-A. Aufaure
- DOCEng'06 (Document Engineering), October 11-13, Amsterdam: A-M. Vercoustre
- RNTL - RIAM - RNRT days, November 15-16, Lyon: M-A. Aufaure
- TM'06 - On the Move to meaningful Internet Systems and Ubiquitous Computing, Montpellier, October 29- November 3: B. Trousse
- INEX'06 Workshop (Initiative for the Evaluation of XML Retrieval), Schloss Dagstuhl, Germany, December 18-20: A-M. Vercoustre, J. Pehceski
- Seminars Jean-Pierre Fénelon on "Analyse des données": Y. Lechevallier
- Workshop on Computer Assisted Critical Edition of Sanskrit Texts Pune July 2006: M. Csernel

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

M. Csernel made an invited seminar in July at the Center for Development of Advanced Computing (C-DAC, Pune) on "Sanskrit Manuscripts Comparison for Critical Edition".

Y. Lechevallier with R. Verde (in the context of the IFCS Transversal Group on Symbolic Data Analysis) organised an invited session on "Symbolic data analysis" at IFCS 2006 Conference - Data Science and Classification (Ljubljana, Slovenia, July 25 - 29).

Y. Lechevallier gave with F. De Carvalho and R. Verde a turotial on "Clustering methods in symbolic data analysis" [67].

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] M. CSERNEL, F. A. T. DE CARVALHO. *Usual Operations With Symbolic Data Under Normal Symbolic Form*, in "Applied Stochastic Models in Business and Industry", vol. 15, 1999, p. 241–257.

[7] 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.

[8] 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.

[9] 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.

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

[11] M. JACZYNSKI, B. TROUSSE. *Patrons de conception dans la modélisation d'une plateforme à objets pour le raisonnement à partir de cas. Design patterns for modelling a case-based reasoning tool*, in "Revue L'objet - logiciel, bases de données, réseaux", vol. 5, n^{o} 2, 1999, p. 203-232.

[12] 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.

[13] 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.

[14] 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.

[15] 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

[16] O. BOUSSAID, B. TROUSSE (editors). *Actes de FDC'06, le troisième atelier sur la « Fouille de données complexes dans un processus d'extraction de connaissances»*, (Atelier de la conférence EGC'06), 2006, http://www-sop.inria.fr/axis/fdc-egc06.

Doctoral dissertations and Habilitation theses

[17] F. ROSSI. *Contribution à l'analyse de données complexes*, HDR, Université de Paris Dauphine, November 2006.

Articles in refereed journals and book chapters

[18] A. AWASTHI, Y. LECHEVALLIER, M. PARENT, J.-M. PROTH. *Using Hybrid Clustering to Approximate Fastest Paths on Urban Networks*, in "Journal od Data Science", vol. 4, n^{o} 1, January 2006, p. 39-66.

[19] 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.

[20] 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.

[21] 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, 2006, p. 167-179, http://www-sop.inria.fr/axis/Publications/uploads/pdf/SouzaChaventDeCarvalhoLechevallier_Versio00.pdf.

[22] A. EL GOLLI, F. ROSSI, B. CONAN-GUEZ, Y. LECHEVALLIER. *Une adaptation des cartes auto-organisatrices pour des données décrites par un tableau de dissimilaritées*, in "Revue de Statistique Appliquée", vol. LIV, n^{o} 3, 2006, p. 33–64.

[23] 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.

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

[25] 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^{o} 2, February 2006, p. 215–226, http://dx.doi.org/10.1016/j.chemolab.2005.06.010.

[26] F. ROSSI, N. VILLA. *Support Vector Machine For Functional Data Classification*, in "Neurocomputing", vol. 69, n^{o} 7–9, March 2006, p. 730–742, http://dx.doi.org/10.1016/j.neucom.2005.12.010.

[27] N. VILLA, F. ROSSI. *Un résultat de consistance pour des SVM fonctionnels par interpolation spline*, in "Comptes Rendus Mathématiques", vol. 343, n^{o} 8, October 2006, p. 555–560, http://dx.doi.org/10.1016/j.crma.2006.09.025.

Publications in Conferences and Workshops

[28] A. BALDÉ, Y. LECHEVALLIER, B. TROUSSE, M.-A. AUFAURE. *Utilisation de métadonnées pour l'aide à l'interprétation de classes et de partitions*, in "Actes des 6ème journées Extraction et Gestion des Connais-

sances (EGC 2006), Revue des Nouvelles Technologies de l'Information (RNTI-E-6), Lille, France", 17-20 January 2006, http://www-sop.inria.fr/axis/Publications/uploads/pdf/article.pdf.

[29] N. BEN MUSTAPHA, M.-A. AUFAURE, H. BAAZAOUI-ZGHAL. *Towards and Architecture of Ontological Components for the Semantic Web*, in "Web Information Systems Modeling Workhop (WISM), in conjunction with CAISE'2006, Luxembourg", 6 June 2006.

[30] N. BEN MUSTAPHA, M.-A. AUFAURE, H. BAAZAOUI-ZGHAL. *Vers une approche de construction de composants ontologiques pour le Web sémantique: synthèse et discussion*, in "Troisième atelier sur la « Fouille de données complexes dans un processus d'extraction des connaissances », 6èmes journées francophones « Extraction et Gestion des Connaissances » (EGC), Lille, France", 17 January 2006, http://www-sop.inria.fr/axis/Publications/uploads/pdf/FDC06-BenMustapha&al.pdf.

[31] M. CHAVENT, O. BRIANT, Y. LECHEVALLIER. *Comparaison d'une méthode de classification descendante hiérarchique monothétique avec Ward et les centres mobiles*, in "13èmes Rencontres de la Société Francophone de Classification (SFC'06), Metz", Société Francophone de Classification, september 2006, p. 60-63, http://www-sop.inria.fr/axis/Publications/uploads/pdf/chaventEtAl_SFC06.pdf.

[32] M. CHAVENT, Y. LECHEVALLIER. *Empirical Comparaison of a Monothetic Divisive Clustering Method with Ward and k-means Clustering Methods*, in "Data Science and Classification (Proceedings of IFCS 2006), Ljubljana, Slovenia", V. BATAGELJ, H.-H. BOCK, A. FERLIGOJ, A. VZIBERNA (editors). , Studies in Classification, Data Analysis, and Knowledge Organization, Springer, IFCS, July 2006, p. 83-90, http://www-sop.inria.fr/axis/Publications/uploads/pdf/Chavent_ifcs2006-Revised.pdf.

[33] S. CHELCEA, A. DA SILVA, Y. LECHEVALLIER, D. TANASA, B. TROUSSE. *Prétraitement et classification de données complexes dans le domaine du commerce électronique*, in "In Atelier N°6: Fouille de Données Complexes dans un processus d'extraction de connaissances, EGC 2006, Lille", O. BOUSSAID, B. TROUSSE (editors). , 17 January 2006, p. 51 - 64.

[34] B. CONAN-GUEZ, F. ROSSI, A. EL GOLLI. *Un algorithme efficace pour les cartes auto-organisatrices de Kohonen appliquées aux tableaux de dissimilarités*, in "Actes des treizièmes rencontres de la Société Francophone de Classification, Metz, France", M. NADIF, F.-X. JOLLOIS (editors). , September 2006, p. 73–76.

[35] A. DA SILVA, F. DE CARVALHO, Y. LECHEVALLIER, B. TROUSSE. *Characterizing Visitor Groups from Web Data Streams*, in "2nd IEEE International Conference on Granular Computing (GrC 2006), Atlanta, USA", 10-12 May 2006.

[36] 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.

[37] C. GARBONI, F. MASSEGLIA, B. TROUSSE. *A Flexible Structured-based Representation for XML Document Mining*, in "Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Dagstuhl Cstle,Germany", LNCS, DOI 10.1007/11766278_35, vol. 3977/2006, Springer Berlin / Heidelberg, 28 June 2006, p. 458-468, http://www-sop.inria.fr/axis/papers/06Lncs-inex05/lncs06-inex05.pdf.

[38] B. HUGENEY, G. HÉBRAIL, Y. LECHEVALLIER. *Réduction de séries temporelles par classification et segmentation*, in "Journées de la Société Francaise de Statistique", June 2006, http://www-sop.inria.fr/axis/Publications/uploads/ps/SFDS-SynthG.ps.

[39] A. IRPINO, R. VERDE, Y. LECHEVALLIER. *Dynamic clustering of histograms using Wasserstein metric*, in "17th COMPSTAT Symposium of the IASC", 2006, http://www-sop.inria.fr/axis/Publications/uploads/pdf/IRP_VER_LECH_COMPSTAT06_V3.pdf.

[40] T. KARNA, F. ROSSI, A. LENDASSE. *LS-SVM functional network for time series prediction*, in "Proceedings of XIVth European Symposium on Artificial Neural Networks (ESANN 2006), Bruges (Belgium)", April 2006, p. 473–478.

[41] 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.

[42] L. KAROUI, M.-A. AUFAURE. *Ontological Concepts Evaluation Based on Context*, in "Poster and Demo Proceedings of the 15th int. Conf. on Knowledge Engineering and Knowledge Management Managing Knowledge in a World of Networks (EKAW'06), Podebrady, Czech Republic", 2-6 October 2006, p. 16-16.

[43] L. KAROUI, N. BENNACER, M.-A. AUFAURE. *Extraction de concepts guidée par le contexte*, in "13èmes rencontres de la Société Francophone de Classification (SFC), Metz, France", 6-8 September 2006, p. 119-123.

[44] B. LE GRAND, M.-A. AUFAURE, M. SOTO. *Semantic and Conceptual Context-Aware Information Retrieval*, in " , the IEEE/ACM International Conference on Signal-Image Technology & Internet-Based Systems (SITIS'2006), Hammamet, Tunisie", 17-21 December 2006.

[45] Y. LECHEVALLIER. *Partitioning Methods*, in "Knowledge Extraction by Interval Data Analysis, Belvedere di San Leucio Caserta", Computational Statistics, Nova Universitas, University of Napoli II, Faculty Jean Monnet, November 2006.

[46] Y. LECHEVALLIER, F. MASSEGLIA, D. TANASA, B. TROUSSE. *Techniques de généralisation des URLs pour l'analyse des usages du Web*, in "In Atelier No 6: Fouille de Données Complexes dans un processus d'extraction de connaissances, EGC 2006, Lille", O. BOUSSAID, B. TROUSSE (editors). , 17 January 2006, p. 141 - 154.

[47] Y. LECHEVALLIER, R. VERDE, F. DE CARVALHO. *Symbolic Clustering of Large Datasets*, in "Data Science and Classification (Proceedings of IFCS 2006), Ljubljana, Slovenia", V. BATAGELJ, H.-H. BOCK, A. FERLIGOJ, A. VZIBERNA (editors). , Studies in Classification, Data Analysis, and Knowledge Organization, Springer, IFCS, July 2006, p. 193-202, http://www-sop.inria.fr/axis/Publications/uploads/pdf/LechevallierVerdeDeCarvalhoIfcs2006After-revision.pdf.

[48] Y. LECHEVALLIER, R. VERDE, A. IRPINO. *Mesures de proximité entre objets décrits par des histogrammes*, in "13ème Journées de la Société Francophone de Classification", Université de Metz, 2006, p. 145-148, http://www-sop.inria.fr/axis/Publications/uploads/pdf/SFC_Lec_Ver_Irp_07.pdf.

[49] A. MARASCU, F. MASSEGLIA. *Classification de flots de séquences basée sur une approche centroïde*, in "Fouille de données complexes dans un processus d'extraction des connaissances (FDC), Lille, France", 17 January 2006, p. 131-139.

[50] A. MARASCU, F. MASSEGLIA. *Classification de flots de séquences basée sur une approche centroïde*, in "XXIVème Congrès INFORSID Informatique des organisations et systèmes d'information et de décision, Hammamet, Tunisie", 2 June 2006, p. 751-765.

[51] A. MARASCU, F. MASSEGLIA. *Extraction de motifs séquentiels dans les flots de données d'usage du Web*, in "Extraction et Gestion des Connaissances (EGC), Lille, France", 20 January 2006, p. 627-638.

[52] F. MASSEGLIA, P. PONCELET, M. TEISSEIRE, A. MARASCU. *Usage Mining : extraction de périodes denses à partir des logs Web*, in "Extraction et Gestion des Connaissances (EGC), Lille, France", 18 January 2006, p. 403-408.

[53] F. MASSEGLIA, P. PONCELET, M. TEISSEIRE. *Peer-to-Peer Usage Analysis: a Distributed Mining Approach*, in "Proceedings of the IEEE 20th International Conference on Advanced Information Networking and Applications (AINA 2006), Vienna, Austria", April 2006, p. 993-998.

[54] F. ROSSI, F. DE CARVALHO, Y. LECHEVALLIER, A. DA SILVA. *Comparaison de dissimilarités pour l'analyse de l'usage d'un site web*, in "Actes des 6ème journées Extraction et Gestion des Connaissances (EGC 2006), Revue des Nouvelles Technologies de l'Information (RNTI-E-6), Villeneuve d'Ascq, France", G. RITSCHARD, C. DJERABA (editors). , vol. II, January 2006, p. 409–414, http://www-sop.inria.fr/axis/Publications/uploads/pdf/egc2006.pdf.

[55] F. ROSSI, F. DE CARVALHO, Y. LECHEVALLIER, A. DA SILVA. *Dissimilarities for Web Usage Mining*, in "Data Science and Classification (Proceedings of IFCS 2006), Ljubljana, Slovenia", V. BATAGELJ, H.-H. BOCK, A. FERLIGOJ, A. VZIBERNA (editors). , Studies in Classification, Data Analysis, and Knowledge Organization, Springer, July 2006, p. 39-46, http://apiacoa.org/publications/2006/wum-diss-ifcs06.pdf.

[56] F. ROSSI, D. FRANÇOIS, V. WERTZ, M. VERLEYSEN. *A functional approach to variable selection in spectrometric problems*, in "Artificial Neural Networks (Proceednigs 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.

[57] F. ROSSI. *Visual Data Mining and Machine Learning*, in "Proceedings of XIVth European Symposium on Artificial Neural Networks (ESANN 2006), Bruges (Belgium)", April 2006, p. 251–264.

[58] F. ROSSI. *Visualization Methods for Metric Studies*, in "Proceedings of the International Workshop on Webometrics, Informetrics and Scientometrics, Nancy, France", May 2006, p. 356–366, http://eprints.rclis.org/archive/00006047/.

[59] D. TANASA, F. MASSEGLIA, B. TROUSSE. *GWUM : une généralisation des pages Web guidée par les usages*, in "Actes de INFORSID 2006, HAMMAMET, Juin 2006", 2006, p. 783-798, http://www-sop.inria.fr/axis/Publications/uploads/pdf/gwum_inforsid.pdf.

[60] A.-M. VERCOUSTRE, M. FEGAS, S. GUL, Y. LECHEVALLIER. *A Flexible Structured-based Representation for XML Document Mining*, in "Advances in XML Information Retrieval and Evaluation, The Fourth International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2005), Schloss Dagstuhl, Germany", N. FUHR, M. LALMAS, S. MALIK, G. KAZAI (editors). , Lecture Notes in Computer Science (LNCS), Volume 3977, N° 3-540-34962-6, vol. Volume 3977 / 2006, n^{O} 3-540-34962-6, Springer, 2006, p. pp. 443 - 457, http://hal.inria.fr/inria-00000839.

[61] A.-M. VERCOUSTRE, M. FEGAS, Y. LECHEVALLIER, T. DESPEYROUX. *Classification de documents XML à partir d'une représentation linéaire des arbres de ces documents*, in "Actes des 6ème journées Extraction et Gestion des Connaissances (EGC 2006), Revue des Nouvelles Technologies de l'Information (RNTI-E-6), Lille, France", January 2006, http://hal.inria.fr/inria-00000840.

[62] N. VILLA, F. ROSSI. *SVM fonctionnels par interpolation spline*, in "Actes des 38ièmes Journées de Statistique de la SFDS, Clamart, France", June 2006.

Internal Reports

[63] T. DESPEYROUX. *Developing efficient parsers in Prolog: the CLF manual (v1.0)*, Technical Report, n^{o} 0310, INRIA, december 2006, https://hal.inria.fr/inria-00120518.

Miscellaneous

[64] S. CHELCEA. *Agglomerative 2-3 Hierarchical Classification: Theoretical and Applicative Study*, Draft version of the PH-D theis document (for reviewers), University of Nice Sophia Antipolis, 13 december 2006.

[65] M. CSERNEL. *Sankrit Manuscript Comparison for Critical Edition*, IT ASIA Constract 2004/091-775, July 2006, Closing workshop of Archeology of Ancient Texts.

[66] A. DA SILVA. *Classification automatique en Web Usage Mining*, 20-21 March 2006, http://www-sop.inria.fr/axis/Publications/uploads/pdf/resume_DA_SILVA.pdf, 2èmes Rencontres Inter-Associations (RIAS 2006).

[67] Y. LECHEVALLIER, F. DE CARVALHO, R. VERDE. *Clustering methods in symbolic data analysis*, July 2006, http://www-sop.inria.fr/axis/Publications/uploads/slides/Clustering Methods in SDA ifcs tutorial.pdf, Workshop Symbolic Data analysis.

[68] F. PATTE, M. CSERNEL. *Computer assisted critical edition of Sankrit texts*, IT ASIA Constract 2004/091-775, July 2006, Closing workshop of Archeology of Ancient Texts.

[69] B. SENACH, B. TROUSSE. *Evaluation des apports des NTIC à la mobilité urbaine. Spécification de critères d'évaluation*, INRIA Sophia Antipolis, Delivrable D5.2. (Version finale) du Projet PREDIT MOBIVIP, 67 pages, april 2006.

[70] B. TROUSSE, B. SENACH, C. MANGEAT, G. CLOUET. *Analyse du site envibus*, Projet AxIS Report, INRIA Sophia Antipolis, 78 pages, november 2006.

[71] B. TROUSSE, B. SENACH, C. MANGEAT, G. CLOUET. *Evaluation des apports des NTIC à la mobilité urbaine: analyse de l'usage d'un site d'information voyageur*, INRIA Sophia Antipolis, Delivrable D5.3 (Version 1) du Projet PREDIT MOBIVIP, 85 pages, november 2006.

[72] J.-N. TURLIER. *Comparaison des manuscrits Sankrit en vue d'une édition critique*, Institut d'Informatique d'entreprise, June 2006, IIE Evry & Inria Internship.

References in notes

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

[74] P. BESSE, H. CARDOT, D. STEPHENSON. *Autoregressive forecasting of some functional climatic variations*, in "Scandinavian Journal of Statistics", vol. 4, 2000, p. 673-688.

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

[76] P. BORRAS, D. CLEMENT, T. DESPEYROUX, J. INCERPI, G. KAHN, B. LANG, V. PASCUAL. *Centaur: the system*, in "Proceedings of the 3rd Symp. on Software Development Environments, Boston, USA", Rapport de Recherche INRIA 777, Inria-Sophia-Antipolis, France, December 1987, November 1988.

[77] L. CANDILLIER, L. DENOYER, P. GALLINARI, M.-C. ROUSSET, A. TERMIER, A.-M. VERCOUSTRE. *Mining XML documents*, F. MASSEGLIA, P. PONCELET, M. TEISSEIRE (editors). , to appear, Idea Group Inc., 2007.

[78] M. CHAVENT. *A monothetic clustering method*, in "Pattern Recognition Letters", vol. 19, n^{o} 11, September 1998, p. 989-996.

[79] M. CHAVENT, Y. LECHEVALLIER. *Evaluation d'une approche de classification conceptuelle*, in "7ièmes Journées francophones Extraction et Gestion des Connaissances EGC 2007, 23/01/2007, Namur, Belgique", to appear, 2007.

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

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

[82] A. DA SILVA, Y. LECHEVALLIER, F. ROSSI, F. DE CARVALHO. *Construction et analyse des résumes de données évolutives : application aux données d'usage du Web*, in "7ièmes Journées francophones Extraction et Gestion des Connaissances EGC 2007, 23/01/2007, Namur, Belgique", to appear, 2007.

[83] 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.

[84] 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, 23/01/2007, Namur, Belgique", Preprint version; to appear, 2007, http://hal.inria.fr/inria-00116910/en/.

[85] T. DESPEYROUX, Y. LECHEVALLIER, B. TROUSSE, A.-M. VERCOUSTRE. *Expériences de classification de documents XML homogènes*, in "Actes des 5ème journées Extraction et Gestion des Connaissances (EGC 2005), Revue des Nouvelles Technologies de l'Information (RNTI-E-3), Paris, France", N. VINCENT, S. PINSON (editors). , vol. 1, Cépaduès-Editions, January 2005, p. 183-188, http://hal.inria.fr/docs/00/03/49/10/PDF/raclass.pdf.

[86] A. EL GOLLI, B. CONAN-GUEZ, F. ROSSI. *Self Organizing Map and Symbolic Data*, in "Journal of Symbolic Data Analysis", vol. 2, n^{o} 1, November 2004.

[87] 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.

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

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

[90] 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.

[91] GNUTELLA. *Gnutella.com*, http://www.gnutella.com.

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

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

[94] 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.

[95] A. KRASKOV, H. STÖGBAUER, P. GRASSBERGER. *Estimating mutual information*, in "Physical Review E", vol. 69, n^{o} 2, 2004, p. 066138.1-066138.16.

[96] B. LE GRAND, M.-A. AUFAURE, M. SOTO. *Sémantique et contextes conceptuels pour la recherche d'information*, in "7ièmes Journées francophones Extraction et Gestion des Connaissances EGC 2007, 23/01/2007, Namur, Belgique", to appear, 2007.

[97] A. LENDASSE, D. FRANÇOIS, F. ROSSI, V. WERTZ, M. VERLEYSEN. *Sélection de variables spectrales par information mutuelle multivariée pour la construction de modèles non-linéaires*, in "Actes de la conférence Chimiométrie 2004, Paris, France", Décembre 2004, p. 44–47.

[98] 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.

[99] A. MARASCU, F. MASSEGLIA. *Mining Sequential Patterns from Temporal Streaming Data*, in "Proceedings of the first ECML/PKDD Workshop on Mining Spatio-Temporal Data (MSTD'05), held in conjunction with the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05), Porto, Portugal", 3 October 2005, http://www-sop.inria.fr/axis/Publications/uploads/pdf/MSTD_05.pdf.

[100] 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.

[101] 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.

[102] 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.

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

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

[105] P. RESNICK, H. R. VARIAN. *Recommender systems*, in "Communications of the ACM", vol. 40, n^{o} 3, 1997, p. 56-58.

[106] 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.

[107] 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.

[108] B. SENACH, B. TROUSSE. *Définition du scénario générique guidant l'évaluation du service VIP*, INRIA Sophia Antipolis, Delivrable D5.1. du Projet PREDIT MOBIVIP, December 2004.

[109] 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.

[110] D. TANASA, F. MASSEGLIA, B. TROUSSE. *Mining Generalized Web Data for Discovering Usage Patterns*, to appear, Idea, 2008.

[111] D. TANASA, B. TROUSSE. *Data Preprocessing for WUM*, in "IEEE Potentials", vol. 23, n^{o} 3, August 2004, p. 22–25.

[112] B. TROUSSE, M.-A. AUFAURE, B. LE GRAND, Y. LECHEVALLIER, F. MASSEGLIA. *Web Usage Mining for Ontology Management*, to appear, Idea Group, 2008.

[113] 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.

[114] 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.

[115] 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.