INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team AxIS

User-Centered Design, Improvement and Analysis of Information Systems

Sophia Antipolis

THEME COG

Table of contents

1. Team

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

Teal Vice-Leader
Yves Lechevallier [Research Scientist (DR2), Inria Rocquencourt]

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

Research Scientists
Thierry Despeyroux [Research Scientist(CR1), Inria Rocquencourt]
Eric Guichard [Education Nationale, until September 30, Inria Sophia Antipolis]
Florent Masséglia [Research Scientist(CR2), Inria Sophia Antipolis]
Fabrice Rossi [Assistant Professor, Univ. Paris IX Dauphine, Inria Rocquencourt]
Anne-Marie Vercoustre [Research Scientist (DR2), 75 %, from Aug., Inria Rocq.]

Research Scientists (partners)
Mireille Arnoux [Assistant Prof., Univ. Bretagne Occidentale, Inria Sophia Antipolis]
Marc Csernel [Assistant Prof., Univ. Paris IX Dauphine, Inria Rocquencourt]

Technical Staff
Mihai Jurca [EPIA project, Inria Sophia Antipolis]
Aicha El Golli [EPIA project, from June 1st, Inria Rocquencourt]

Post-doctoral Fellows
Brieuc Conan-Guez [Lecturer, Univ. Paris IX Dauphine, until August 31, Inria Rocq.]

Ph. D. Students
Abdourahamane Baldé [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]
Aicha El Golli [Univ. Paris IX Dauphine, until May 31, Inria Rocquencourt]
Doru Tanasa [Univ. Nice Sophia Antipolis (UNSA-STIC), Inria Sophia Antipolis]

Visiting Scientists
Francesco de Carvalho [Federal Univ. of Pernambuco, Brazil, September-October, Inria Rocq.]
Bel Mufti Ghazi [Ecole Supérieure des Sciences Economiques et Commerciales, Tunis, Tunisia, September, Inria Rocq.]
Rosanna Verde [Prof., University of Napoli, Italy, March-May, Inria Rocq.]
Alzennyr Da Silva [Univ. of Pernambuco, Brazil, October, Inria Sophia Antipolis]
Renata De Souza [Assistant Professor, University of Pernambuco, Brazil, October, Inria Rocq.]
Abdelaziz Marzark [Univ. of Casablanca, Nov., Inria Sophia-Antipolis]

Student Interns
Luc Baubois [Univ. Paris IX Dauphine, March-August, Inria Rocquencourt]
Calin Garboni [Univ. of Nice Sophia Antipolis, January-July, Inria Sophia Antipolis]

2. Overall Objectives

2.1. Objectives

Keywords: *KDD, RDF knowledge discovery, WWW, Web mining, Web usage mining, XML, adaptive interface, artificial intelligence, case-based reasoning, classification, clustering, content mining, data mining, document*

mining, *evaluation process*, *experience management*, *hierarchical clustering*, *information retrieval*, *information system*, *knowledge management*, *multimedia*, *neuronal network*, *ontology*, *recommender system*, *semantic Web*, *semantic checking*, *sequential pattern*, *statistics*, *structure mining*, *transportation*, *user-centered design*, *viewpoint management*.

The AxIS project-team was created on July 1st, 2003. AxIS is a multi-disciplinary team (Artificial Intelligence, Data Mining & Analysis and Software Engineering). It leads researches in the area of Information Systems (ISs) (particularly Web-based information systems like web sites) and aims to develop methods and tools for supporting user-centered design, analysis and improvement of ISs. The objectives of our research are twofold. The first one is to assist both designers and users involved with ISs. The second objective is to anticipate from the design step and/or support the following tasks : 1) Usage analysis for the evaluation process and 2) Maintenance and integration of frequent site evolutions
Our approach to IS formalisation for improved quality is inspired from works in semantics in programming languages, making a parallel between the syntax of programming languages and the structure of a web-based information system (or semi-structured documents).

Our research program (cf. fig. 1) distinguishes between the static and the dynamic aspects of ISs (cf. section 3.1) to achieve the following goals:

- improvement of an IS by confronting static and dynamic aspects,
- and knowledge and experience capitalization in multi-view evaluation of IS.

In the area of software development, AxIS aims at defining specific languages based on ontologies relating to specific activities, and proposes software platforms for the assistance of the specification and evaluation of IS.

3. Scientific Foundations

3.1. Semantics and Design of Hypertext Information Systems

Keywords: *formal semantics*, *semantic Web*, *semantic checking*, *semantics*.

Designing and maintaining hypertext information systems, such as Web sites, are a real challenge. On the Web, it is much easier to found inconsistent pieces of information than a well structured site. Our goal is to study and build tools that are necessary to design, develop and maintain complex but coherent sites. We use a multi-disciplinary approach, 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 some of the ideas which have been developed in the software engineering world.
Most of the efforts deployed in the Web domain were related to languages for documents presentation (HTML, CSS, XSL) and structure (XML), to Web sites modeling and Web services (UML), but not to the formal semantics of Web sites to support quality and evolution of Web sites. The initiative led by the W3C consortium on Web Semantic (XML, RDF, RDF-schema) and ontologies aims at a different objective related to resource discovery.

The term "semantics" has at least two significations:

- the scientific study of the meaning of words and texts,
- the study of propositions in a deductive theory.

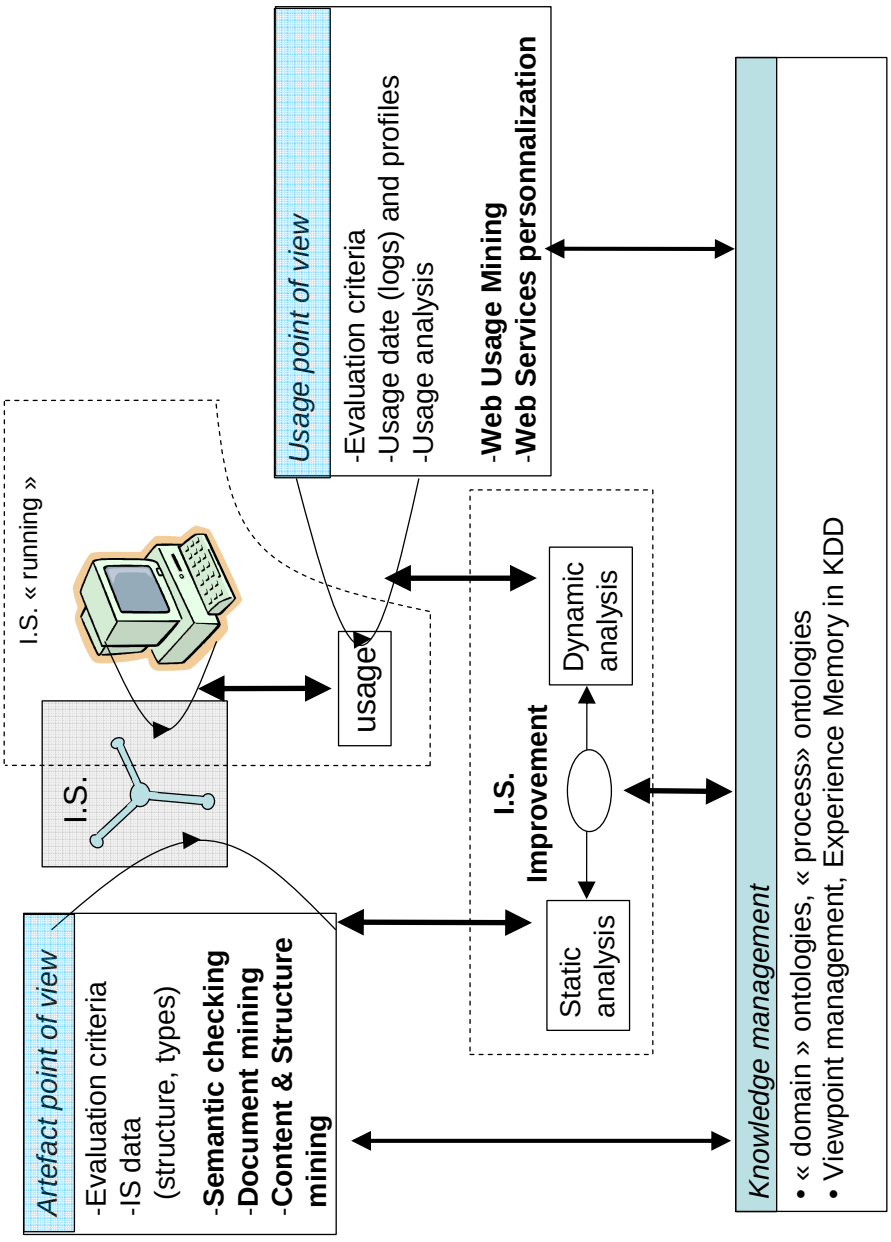

Figure 1. Global View of AxIS Research Topics

We will use this last definition when 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 inside a Web side which also needs to be specified and validated (cf. the execution of a program).
Our approach is different but related to the Semantic Web. The main goal of the Semantic Web is to ease computer-based data mining, formalizing data that is mostly textual, for further discovery. We are concerned by the way Web sites are constructed in the first place, taking into account their development and their semantics. In this respect we are closer to what is called content management.
We use approaches and techniques imported from logic programming and formal semantics of programming languages, in particular operational semantics.

3.2. Usage Mining : Applying KDD to Usage Data

Keywords: *clustering*, *data mining*, *data warehouse*, *sequential patterns*, *usage mining*, *user behaviour*, *web usage mining*.

Let us consider the KDD process represented by Fig. 2. This process is made of four main steps:

1. **Data selection** aims at extraction from the database or datawarehouse the information needed by the data mining step.
2. **Data transformation** will then use parsers in order to create data tables which can be used by the data mining algorithms.
3. **Data mining** techniques range from sequential patterns to association rules or cluster discovery.
4. finally the last step will allow the **re-use of the obtained** results into a usage **analysis** process.

The studies conducted over KDD applied to usage data have two goals: improving the usage of the IS and/or enhance the IS by comparing the structure information about the IS with the results of the usage analysis.

Let us zoom on the five following research topics:

3.2.1. Data selection and transformation

The considered KDD methods will rely on the notion of session, represented through a tabular model (items), an association rules model (itemsets) or a graph model. This notion of session enables us to act in the good level during the process of knowledge extraction 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.

Actually, as the analysis methods come from various research fields (data analysis, statistics, data mining, A.I., ...), a data transformation from input to output is needed and will be managed by the parsers. The input data will come from databases or from standard formatted file (XML) or a private format.

We insist on the importance of this step in the KDD process.

3.2.2. Extracting association rules

Our preprocessing tools (or generalization operators) given in the previous part were designed to build summaries and also 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 presented 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 work coming from the field of generalization operators and data aggregation. These summaries can be integrated in a recommendation mechanism for the user help. We propose to adapt

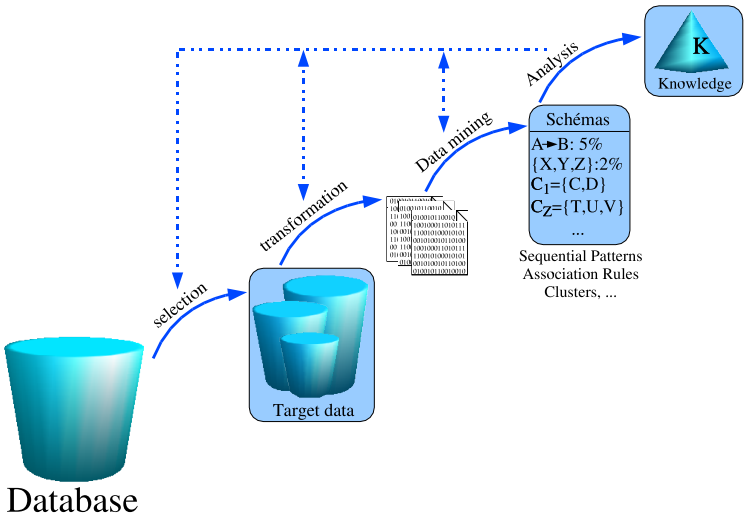

Figure 2. Steps of the KDD Process

frequent itemset research methods or association rules discovery methods to the Web Usage Mining problem. We may get inspired by methods coming from the genomist 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.

3.2.3. Discovering sequential patterns

Knowing the user can be based on sequential pattern (which are inter transactions patterns) discovery. Sequential patterns offer a strong correlation with Web Usage Mining (and more generally with usage analyzes problems) purposes. 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 enhance the sequential pattern extraction methods by taking into account the context where those methods are involved. This can be done:

- first of all by analyzing the causes of a 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.
- It is also necessary to incorporate all the available information related to the usage. Taking into account several information sources in a single sequential pattern extraction process is a challenge and can lead to numerous opportunities.
- Finally, sequential pattern mining methods will have to get adapted to a new and growing domain: data streams. In fact, in numerous practical cases, data cannot be stored more than a specified time (and even not at all). Data mining methods will have to provide solution in order to respect the specific constraints related to this domain (no multiple scan over the data, no blocking actions, etc.).

3.2.4. Clustering approach for reducing the volume of data in data warehouses

Clustering is one of the most popular techhnique in knowledge acquisition and it is applied in various fields including data mining and statistical data analysis. This task organizes 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 as to the number of individuals they can comfortably handle.

Cluster analysis may be divided into hierarchical and partitioning methods. Hierarchical methods yields complete hierarchy, i.e., a nested sequence of partitions of the input data. Hierarchical methods can be agglomerative or divisive. Agglomerative methods yields 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.

3.2.5. Reusing usage analysis experiences

This topic aims at re-using previous analysis results into current analysis: in the short run we will work on an incremental approach of the discovery of sequential motives; in the longer run our approach will be based upon case-based reasoning. Nowadays very fast algorithms have been developed which efficiently search for dependences between attributes (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, and depending on the size of the database, it can sometimes take up to several days to retrieve relevant and useful information. Furthermore, the variation of parameters provided 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 problematic is currently considered as a research problem open within the framework of Data Mining; and even though a few solutions exist, they are not quite satisfactory because they only provide a partial solution to the problem.

3.3. Adaptive Recommender Systems

Keywords: *CBR, KDD, collaborative filtering, hypermedia, personalization, recommender system, user profile, user support Web.*

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 [78][76][74].

A recommender system can be divided into three basic entities (e.g. 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:

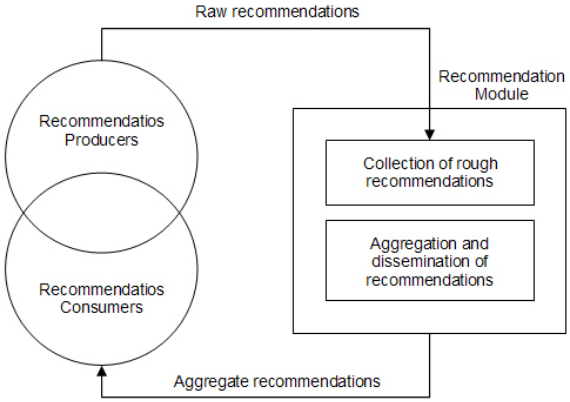

Figure 3. Architecture of a Recommender System

How to produce adaptive recommendations of high quality
minimizing the effort of producers and the consumers?

Two main complementary approaches are proposed in the literature: 1) approaches based on the content and the machine learning of user profiles and 2) approaches known as a collaborative filtering based on data mining techniques. The user profile is a structure of data that describes user's centers of interests 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 to a user something that satisfied previous users with a similar profile (collaborative filtering) [76].
In the Axis project-team, we continue the development of a hybrid approach of calculation of recommendations based on the analysis of visited content and centered on data mining, where the past behaviours of a user group are used to calculate the recommendations (collaborative filtering).

The vast majority of approaches based on data mining are mainly statistical approaches where the order of occurrence of events in the history is not taken into account for the calculation of recommendations. Here are some first examples in the field of navigation assistance on the Web: the FootPrints system [81] and the system of Yan et al [79].
The implementation difficulties of our approach relate to the following aspects:

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

We explore all three problems above by using case-based reasoning (CBR) techniques and more generally KDD techniques.

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). We focus on two types of recommender systems:

- systems where the calculation of recommendations is based on the re-use of experiences of a users group that search for information on an hypertext information system like the Web or on an Internet/intranet site. These systems aim at an adaptive assistance to the search for information activity ;
- systems where the calculation of recommendations is based on the re-use of past experiences of experts, in order to provide an assistance to design process.

3.4. Case-Based Reasoning

Keywords: *case-based reasoning, experience management, indexing, reuse of past experiences, sequential patterns, use of an information system indexing, user behaviour*.

Case-Based Reasoning (CBR) It 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 [75] usually breaks up into four principal phases [66][73]:

1. a << retrieve >> phase for cases having similarities (i.e. similar indices) with the current problem,
2. a << re-use >> phase where a solution to the current problem is built, based on cases identified in the previous phase,
3. a << revise >> phase where the solution may be refined with an evaluation process,
4. a << retain >> phase, that updates the elements of the reasoning by taking into account the experiment which has been just carried out and which could thus be used for future reasoning.

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 research-adaptation link of case (the best case being the most easily adaptable case), definition of an adaptation strategy starting with the found case(s), training of new indices, etc.
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. Moreover, we study sessions indexing techniques and search algorithms of items sub-sequential patterns for the on-line and off-line analysis of the Web users 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 explores any applicative field on design, evaluation and improvement of a big size hypermedia information systems, for which end-users are of primary concern. 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 services of assistance 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, eg.:

  - 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. the search engines (Yahoo, Voila), the 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.

In summary, our fields of interest are the following:

- Semantic checking of an information system,
- Usage analysis of an information system (internet, intranet),
- Re-designing of an information system bases on usage analysis,
- Adaptive recommender systems for supporting information retrieval, Collaborative search of Information on the internet.

Ultimately, it should be noted that other fields (Life Science, health, transports, etc.) may be subject to the 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 badly solved in automatic of control type (eg nutrition of plants under greenhouses, controls in robotics).

5. Software

5.1. Introduction

AxIS softwares are mostly designed using the Rational Rose environment and developed using Java or C++ programming languages. These software are described at the following URL http://www-sop.inria.fr/axis/software.html

5.2. SODAS 2 Software

Participants: Yves Lechevallier [correspondant], Marc Csernel, Aicha El Golli, Brieuc Conan-Guez.

The SODAS 2 Software [60] is the result of the European project called “ASSO”(Analysis System of Symbolic Official data), that started in January 2001 for 36 months (cf. section 8.3.2). 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.
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 library SOM
This software is now in the registration process at APP. The latest executive version (version 2.50) of the SODAS 2 software, with its the user manual (PDF format), can be downloaded at
http://www.info.fundp.ac.be/asso/sodaslink.htm

5.3. Clustering Toolbox Online

Participants: Marc Csernel, Sergiu Chelcea, Francesco de Carvalho, Mihai Jurca, Yves Lechevallier [co-correspondant], Brigitte Trousse [co-correspondant].

The clustering toolbox, written in C++and Java, groups all tools and classification methods developed by the team over time, and uses the SOM library developed by M. Csernel. This library proposes a common data interface to every algorithm. This toolbox supports developers in integrating various classification methods and testing and comparing with other methods. Now it integrates the four methods :

- SCluster and Div (in C++) from AxIS Rocquencourt ;
- CDis (in C++) issued from a collaboration between by AxIS Rocquencourt team and Recife University, Brazil ;
- CCClust (in C++) issued from a collaboration between by AxIS Rocquencourt team and Recife University, Brazil;
- 2-3 AHC (in Java) from AxIS Sophia Antipolis.

We developed an Web interface for this clustering toolbox for our own needs. The aim of this Clustering Tools Online interface is in a short term to allow other team members and Internet users to use these classification methods to processing their own data via the Web. The Web interface is developed in C++, run on our Apache internal Web server.

5.4. CBR*Tools - Object-oriented Framework for Case-Based Reasoning

Participants: Sergiu Chelcea, Mihai Jurca, Brigitte Trousse [correspondante].

CBR*Tools is an object-oriented framework [71][67] for Case-Based Reasoning, that offers a set of abstract classes to models the main concepts necessary to develop applications integrating case-base 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 [80], prototypes indexing [70], 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 aims 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, CASA and RA2001). We showed that, for each application, the thorough expertise necessary to use CBR*Tools relates to only 20% to 40% of the hot spots thus validating the assistance brought by our platform as well on design as on the implementation, thanks to the re-use of its abstract architecture and its components (index, similarity).
CBR*Tools is used in two current contracts: EPIA and MobiVip.
URL= http://www-sop.inria.fr/axis/cbrtools/manual/.

5.5. Broadway*Tools - Generator of Adaptative Recommender Systems

Participants: Mihai Jurca, Brigitte Trousse [correspondante].

Broadway*Tools is a toolbox used to facilitate the creation of adaptive recommendation systems for information retrieval on the Web or in a Internet/intranet information system. This toolbox offers different servers, including a server that calculates 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: Broadway-Web, educaid (France Telecom Lannion - Inria contract), Be-CBKB (XRCE-Inria contract), e-behaviour (Ocolors Action, use of the mouse and eye-tracking events), etc.

Broadway*Tools is used in two current contracts: EPIA and MobiVip.

6. New Results

6.1. Data Transformation and Knowledge Representation

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

6.1.1. ARFF Format Library

Keywords: *ARFF library, PCA, Self Organizing Map, database.*

Participants: Luc Baubois, Aicha El Golli, Yves Lechevallier.

During his trainee period Luc Baubois [53] has proposed the ARFF library. This library generates an ARFF file from database. An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of the University of Waikato for use with the Weka machine learning software(http://www.cs.waikato.ac.nz/~ml/). Weka is a collection of machine learning algorithms for data mining tasks and marketed as open source released under the GNU General Public License. The library proposed also visualization interfaces for the ARFF files using the PCA (Principal Component Analysis) projection, as well as the resulting ARFF files of the Kohonen maps, extended to mixed data.

6.1.2. Structure Mining in Preprocessing

Keywords: *long patterns, sequential patterns, structure mining, web usage mining.*

Participants: Calin Garboni, Florent Masséglia.

In this work, we focus on the preprocessing step of knowledge discovery. Actually, considering the actual representation of the KDD process, given in Figure 4, we are working on the automation of the preprocessing step (within the dash lines). Indeed, in order to perform a knowledge discovery process on any kind of data, one has to transform the data from the original raw format to a specific format that will be understood by the data mining algorithm. This transformation is usually designed by an expert who has the required knowledge about the data and about the data mining algorithm. Our goal is to help and even replace the expert by providing an intelligent tool, able first to “understand” the structure of the data to prepare, and second to propose a parsing over the data (based on the discovered structure). We started using an access log file and tried to discover automatically the structure of this file. As this structure is already well known, we use it as a bench for our proposal. The method is based on the sequential pattern mining principle. Actually, mining sequential patterns aims at extracting the frequent sequences hidden in the data. The structure in a file organizing the records line by line can be considered as a frequent sequence repeated in (almost) each record.

Sequential patterns for structure discovery

Our contribution is based on a comparison between sequential patterns and structures. The structure is expected to be common to all records in the datasets. Nevertheless, in numerous types of records, such as log file entries, data can be altered (due to errors while recording the entry, system crash, etc.). The structure can thus be considered as a sequential pattern having a very high support over the dataset. For an Apache log file, we expect to find a frequent pattern such as :

... - - [/Mar/2003::: +000] "GET / HTTP/1." 0 "" ""

Then, filling this pattern would give the following rule:

[0-9]+”.”[0-9]+”.”[0-9]+”.”[0-9]+ - - [”[0-9]+”/Mar/2003” (etc.)

With this type of rule, it will be possible to parse the original data file and to obtain the transformed data file. So far, we developed the SINTHES (Structure is IN THE Sequence) method [35],[56]. The main contribution of this method is to provide a top-down generating-pruning principle for frequent sequences extraction. Actually, the structure to discover is usually very long, so any existing method based on the apriori principle will fail

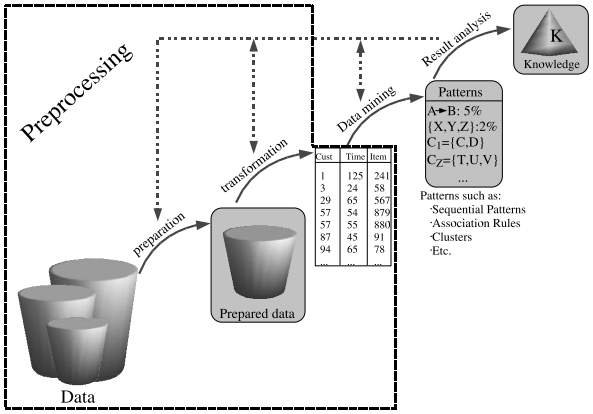

Figure 4. The usual KDD process

(because of the large number of candidates to generate and test). SINTHES uses a sample of the data file to process and applies several filters to the sequences in order to delete infrequent combinations of characters in the considered candidates. When the filters have been applied to the candidates, the most frequent is used in the top-down generating-pruning principle. This candidate will be the seed for all the subsequences of size $k-1$. For each one we determine the frequency and the most frequent subsequence will be chosen for the next iteration. The algorithm will continue until the support of one subsequence is higher than the minimum support and that will be the candidate representing the structure.

Objective: a new schema for KDD

The aim of this work is to provide a new schema for the KDD process. Once the structure is discovered and the parser is generated (thanks to the rules inferred from the data file), it is possible to apply this parser to the original data file (as illustrated in figure 5). At this time, only the structure discovery is possible (thanks to the SINTHES method). A further step will aim at generating the parsing rules by comparing the structure to the data file and extract the missing information (nature of the embedded characters in the structure). Then, based on these rules, the parser will be generated.

6.1.3. Metadata Extraction for Supporting the Interpretation of Clusters

Keywords: *Dublin Core, RDF, classification, metadata.*

Participants: Abdourahamane Baldé, Yves Lechevallier, Brigitte Trousse.

The huge volume of data produced by different information sources requires to develop tools to retrieve pertinent data. Metadata, defined as information about data, is a challenging way to add semantic to data and a way to manage a considerable volume of data without accessing their content. Nowadays, a huge volume of unstructured data needs also to be managed. These data are usually badly structured and difficult to access. In our context, metadata will be used to share information and resources without any access to their content in respect of privacy issues. Data from different sources can be compared using those metadata.

We propose a new methodology in collaboration with M-A Aufaure (Supelec) to extract metadata during the classification process [23], [22]. Metadata will give information about clusters like clusters contents, variables describing the clusters, classification method, set of criteria used, general information. Standards like Dublin Core and RDF have been used to model metadata. We applied this structure to one algorithm called Clustering Algorithm on Symbolic Data Table: our goals are to offer descriptions to facilitate the interpretation of the resulting clusters. Today metadata offer a true means of capitalization of knowledge and know-how.

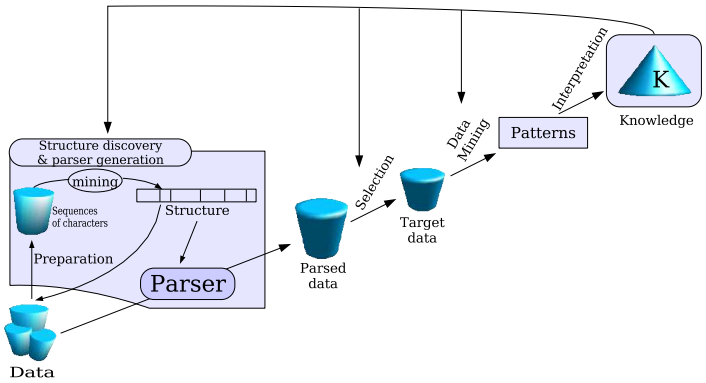

Figure 5. The user driven KDD process

6.1.4. Viewpoint Managment for Annotating a KDD Process

Keywords: *KDD, RDF, annotation, knowledge, metadata, ontology, reusability, viewpoint, weka.*

Participants: Hicham Behja, Brigitte Trousse.

This research is mainly related to Behja's doctoral thesis in the context of France-Morocco Cooperation (Software Engineering Network). Our goal is to make more explicit the notion of " viewpoint " [10] from analysts during their activty and to propose a viewpoint-based KDD model 1) for annotating the underlying goals of KDD activities and 2) for encapsulating existing KDD algorithms or methods and then offering more flexibility and adaptability.

In 2004, in collaboration with Abdelaziz Marzark (Univ of Casablanca, Morocco), we propose a new approach for applying the viewpoint notion in Knowledge Discovery from Data Base (KDD) multiviews analysis. We defined viewpoint as the perception of an expert on a KDD process, perception referred by his/her own knowledge. We propose to structure this knowledge into two different types: 1) domain knowledge that will give information primarily about attributes and data from the database and 2) task knowledge from the analyst field, that will relate to the tasks carried out by the analyst during the KDD process. This classification aims, on the one hand, at integrating the role of the processing expert, and on the other hand, at reducing the size of the handled data. The goals are to facilitate both reusability and adaptability of a KDD process, and to reduce his complexity with maintaining the trace of the past analysis viewpoints. The KDD process will be regarded as generating and transformation views annotated by metadata to store the discovery knowledge. We started with a 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) modeling KDD process adopting a Semantic web based approach [54] and 3) KDD process annotation. Then we designed and implementing an object platform for the KDD processes including the viewpoint notion (design patterns and UML using Rational Rose). The current platform is based on the Weka library. We are now applying our model to analyzing the use of web sites, specially Inria Sophia-Antipolis site (according to the "reliability" and "ergonomic" viewpoints).

6.2. Data Mining Methods

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

6.2.1. Symbolic Data Extraction and Self-Organizing Maps

Keywords: *Self Organizing Map, dissimilarity, relational database, symbolic data analysis, unsupervised clustering.*

Participants: Aicha El Golli, Yves Lechevallier, Brieuc Conan-Guez, Fabrice Rossi.

The aim of symbolic data analysis is to provide a better representation of the variations and imprecision contained in real data. As such data express a higher level of knowledge, the representation must offer a richer formalism than the one provided by classical data analysis.

A generalization process exists that allows data to be synthesized and represented by means of an assertion formalism that was defined in symbolic data analysis. This generalization process is supervised and often sensitive to virtual and atypical individuals. When the data to be generalized is heterogeneous, some assertions include virtual individuals. Faced with this new formalism and the resulting semantic extension that symbolic data analysis offers, a new approach to processing and interpreting data is required.

The original contributions of our work concern new approaches to representing and clustering complex data.

First, we propose a decomposition step, based on a divisive clustering algorithm, that improves the generalization process while offering the symbolic formalism [37]. We also propose a unsupervised generalization process based on the self-organizing map. The advantage of this method is that it enables the data to be reduced in an unsupervised way and allows the resulting homogeneous clusters to be represented by symbolic formalism.

The second contribution of our work is a development of a clustering method to handle complex data [33]. The method is an adaptation of the batch-learning algorithm of the self-organizing map to dissimilarity tables. Only the definition of an adequate dissimilarity is required for the method to operate efficiently. This adaptation can handle both numerical data and complex data. The experiments showed the usefulness of the method and that it can be applied to a wide variety of complex data once we can define a dissimilarity for those data. This method has also given good results for real applications [12][36] with different types of data (meteorological data of China [13], functional data [47] and Web Usage Mining [34].

6.2.2. Functional Data Analysis

Keywords: *curves classification, functional data, neural networks.*

Participants: Fabrice Rossi, Brieuc Conan-Guez, Aicha El Golli, Yves Lechevallier.

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 earlier works, we proposed the extension of MLPs (Multi-Layer Perceptrons) to functional inputs: the Functional Multi-Layer Perceptron (FMLP) [51]. We demonstrated two important properties: this model is a universal approximator, and the parameter estimation is consistent when we only know a finite number of functions known on a finite number of evaluation points.

In 2004, we studied the advantages of a functional pre-processing of input functions before processing by functional neural models:

- We applied FMLPs to a phoneme recognition problem [28]. The goal is to classify 5 different phonemes (TIMIT database). We used a functional PCA in order to smooth noisy spectra, and to reduce the input space dimensionality (each spectrum, described by a vector of 256 components, is projected on 10 eigenfunctions thanks to the Functional PCA). This approach based on Functional PCA gives better results than those obtained by previous studied approaches (for example, representation of input functions thanks to B-spline bases).

- We showed that FMLPs and more generally functional approaches are not very sensitive to missing data when data are functions [48]: we compared functional approach to other traditional approaches (mean value approach, K-NN approach) and best performances were obtained by Functional neural models.
- We extended the Radial Basis Function Networks to functional inputs (FRBFN): compare to FMLPS, the learning stage of this new model can be conducted very quickly, as it involves only algebraic calculus. This allows to explore a wide range of learning parameters (number of neurons, pre-processing applied to functional inputs,...). We apply FRBFNs to a spectrometric application from food industry : results are satisfactory [31].
- Finally for non supervised classification problems (clustering), we extended the self-organizing map method (SOM) to the functional framework [47]. Thanks to the input space dimensionality reduction, this new algorithm is efficient. Moreover, the use of functional transformation (derivative calculation for instance) allows to compare very different clustering of the same data and therefore provides new exploratory representations of functional data.

6.2.3. Partitioning Method : a Clustering Approach for Reducing the Size of Data

Keywords: *Large Data Base, Self Organizing Map, unsupervised clustering.*

Participants: Yves Lechevallier, Luc Baudois.

Clustering methods reduce the volume of data in data warehouses, preserving the possibility to perform needed analyzes [38]. An important issue in databases and data warehouses is that they describe several entities (populations) which are linked together by relationships. In this situation compressed data has no interpretation and cannot be used unless decompressing them. Our work made in cooperation with Antonio Ciampi (Univ of McGill, Canada) and Georges Hébrail (ENST, Paris) differs from this work in the sense that our compression technique has a semantic basis.

Our approach is based on two key ideas [39]:

- A preliminary data reduction using a Kohonen Self Organizing Map (SOM) is performed. As result, the individual measurements are replaced by the means of the individual measurements over a relatively small number of micro-clusters corresponding to Kohonen neurons. The micro-clusters can now be treated as new 'cases' and the means of the original variables over micro-clusters as new variables. This 'reduced' data set is now small enough to be treated by classical clustering algorithms. A further advantage of the Kohonen reduction is that the vector of means over the micro-clusters can safely be treated as multivariate normal, owing to the central limit theorem. This is a key property, in particular because it permits the definition of an appropriate dissimilarity measure between micro-clusters.
- The multilevel feature of the problem is treated by a statistical model which assumes a mixture of distributions, each distribution representing a cluster or group. Although more complex dependencies can be modeled, for example we will assume that the group only affects the mixing coefficients, and not the parameters of the distributions.

6.2.4. Partitioning method : Clustering of Quantitative Data

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

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

We proposed an approach [30], which generalizes easily the dynamic cluster method for the case of the adaptive and non-adaptive Lr distances. This approach can be used with numerical data alone, interval data alone or numerical and interval data together. We did a theoretical study for r = 1 and r = 2: in that case we rediscovered the usual exemplars (median (r = 1) and mean (r = 2), respectively) of the clusters. In the case r>2, the difficulty is to find a realistic interpretation for the cluster representatives. For the future, we would like to continue to study the mathematical properties of these distances and to implement the corresponding algorithms and an empirical framework to their evaluation.

We worked on adaptive and non-adaptive dynamic cluster methods for interval data [49], which generates a partition of the input data, and a corresponding prototype (a vector of intervals) for each class by optimizing an adequacy criterion that is based on Mahalanobis distances between vectors of intervals. In a first approach we used a particular Mahalanobis family of distances but there are other possibilities, which we intend to explore in the near future. We would like also to implement the corresponding algorithms for these others distances and an empirical framework to their evaluation.

We proposed an approach to cluster constrained symbolic data using the dynamic clustering algorithm applied to a dissimilarity table [29]. The clustering criterion is based on the sum of dissimilarities between the objects belonging to the same class. We introduced a suitable dissimilarity function between symbolic data constrained by rules. To be able to compute dissimilarities between constrained symbolic data in a polynomial time, we used a method, called Normal Symbolic Form, which decomposes the data according to the rules in such a way that only the valid parts of the description are represented. For the future, we would like to implement the corresponding algorithms and an empirical framework to their evaluation.

6.2.5. Partitioning Method : Crossed Clustering Method for WUM

Keywords: *WUM, contingence table, dynamic clustering algorithm, unsupervised clustering.*

Participants: Yves Lechevallier, Rosanna Verde.

We proposed a crossed clustering algorithm in order to partition a set of objects in a predefined number of classes and to determine, at the same time, a structure (taxonomy) on the categories of the object descriptors [40]. This procedure is a simultaneous clustering algorithm on contingency tables [41]. The convergence of the algorithm is guaranteed at the best partitions of the objects in r classes and of the categories of the descriptors in c groups, respectively.

In our context we extent the crossed clustering algorithm to look for the partition P of the set E in r classes of objects and the partitions Q in c column-groups of V, according to the Φ^2 criterion on set-valued variables. In this perspective, we generalize a crossed clustering algorithm ([68], [69]).

It is worth to notice that the criterion optimized in such algorithm is additive:

$$\Delta(P,(Q^1,...,Q^p)) = \sum_{v=1}^{p} \Phi^2(P,Q^v|Q)$$

where Q^v is the partition associated to the modal variable y_v and $Q = (Q_1,...,Q_c) = (\bigcup_{v=1}^{p} Q_1^v,...,\bigcup_{v=1}^{p} Q_k^v,...,\bigcup_{v=1}^{p} Q_c^v,)$.

The cells of the crossed tables can be modeled by marginal distributions (or profiles) summarizing the classes descriptions of the rows and columns.

6.2.6. Agglomerative 2-3 Hierarchical Clustering: study and visualization

Keywords: *2-3 hierarchies, aggregation index, clustering, hierarchies, visualization.*

Participants: Sergiu Chelcea, Mihai Jurca, Brigitte Trousse.

Improvement of the 2-3 AHC algorithm

In the context of Chelcea's thesis concerning clustering methods for usage analysis and more particularly the agglomerative hierarchical methods, we have continued in 2004 our work on the Agglomerative 2-3 Hierarchical Clustering (2-3 AHC). We have proposed a new version of the 2-3 AHC algorithm [25] with the same $\Theta(n^2 \log n)$ algorithmic complexity as the classical AHC. Comparative tests between the classical AHC and our 2-3 AHC algorithm were performed on simulated data and proved the richer quality of the 2-3 AHC structures against the classical AHC ones.

We also studied the influence of the aggregation index (single-link and complete-link) on the created 2-3 hierarchy when clusters properly intersect between themselves and improved its quality.

Hierarchies visualization toolbox

To better visualize and compare the created hierarchies and 2-3 hierarchies on same data sets, we developed (in Java) the HIERARCHIES VISUALIZATION TOOLBOX.

Using it, the input data can be randomly generated, loaded from files (e.g. xml, sds, text) or extracted via SQL queries from a specified database server. Next, different methods (AHC, 2-3 AHC with integrated refinement, 2-3 AHC without integrated refinement) and aggregation indexes (single-link and complete-link) can be chosen and executed successively. The results can be then compared based on the number of created clusters, on the induced dissimilarities, on the execution time, etc. for quality based analyses.

The toolbox was also made available via the axis Web server[1] to the other team members for testing purposes. In 2004, it has been integrated in our Clustering Toolbox .

6.3. Web Mining and Web applications

Keywords: *2-3 hierarchies, AHC, HTTP logs, XML, adaptative services, clustering, data analysis, document mining, metrology, neural network, personalization, preprocessing, recommender system, semantics, sequential pattern, user behavior, user profile, web usage mining.*

6.3.1. Site Semantic Checking

Keywords: *CLF, Web Semantics, Web sites, adaptative services, formal approaches, natural semantics, semantics, typing.*

Participant: Thierry Despeyroux.

The main goal of the Semantic Web is to ease a computer-based data mining and discovery, formalizing data that is mostly textual. Our approach is different as we are concerned in the way Web sites are constructed, taking into account their development and their semantics. In this respect we are closer to what is called content management.

Our formal approach is based on the analogy between Web sites and programs when there are represented as terms, although differences between Web sites and programs can be pointed out :

- Web sites may be spread along a great number of files. This is also the case for programs, but these files are usually all located on the same file system. With Web sites we will have to take into account that we may need to access different servers. Currently, a program such as the "make" program cannot handle URLs, only directories.
- Information is scattered, with many forward references. A forward reference describes an object (or a piece of information) that is used before it has been defined or declared. In programs, forward references exist but are most often limited to single files so the compiler can compile one file at a time. This is not the case for Web sites, and as it is not possible to load a complete site at the same time, we need to use other techniques.
- We may need to use external resources to define the static semantics (for example one may need to use a thesaurus, ontologies or an image analysis program). In one of our example, we call the wget program to check the validity of URLs in an activity report.

[1] http://axis.inria.fr:8002

We are developing a specification language to express global constraints in Web sites. The compiler is written in Prolog and produces Prolog code to take advantage of a fast XML parser developed previously and the ease of term manipulation of prolog.

As a real sized test application, we have used the scientific part of the activity reports published by Inria for the year 2001 and 2002 that can be found at the following URLs:
http://www.inria.fr/rapportsactivite/RA2001/index.html and
http://www.inria.fr/rapportsactivite/RA2002/index.html.

The XML versions of these documents contain respectively 108 files and 125 files, a total of 215 000 and 240 000 lines, more than 12.9 and 15.2 Mbytes of data. Our system reported respectively 1372 and 1432 messages.

This work has been presented at the Word Wide Web conference in May 2004 [32].

We started to new applications. The first one concerns XML document mining and is presented in the following section. Our specification language is used to select subpart of XML documents and to interface with an external natural language tagger. The second is to study the way a site or an XML document containing external URLs are corrupted during their life.

6.3.2. XML Document Mining

Keywords: *Document mining, XML classification, XML clustering.*

Participants: Thierry Despeyroux, Yves Lechevallier, Brigitte Trousse, Anne-Marie Vercoustre, Mihai Jurca.

With the increasing amount of available information, there is a need for more sophisticated tools for supporting users in finding useful information. In addition to tools for retrieving relevant documents, there is a need for tools that synthesise and exhibit information that is not explicitly contained in the document collection, using document mining techniques. Document mining objectives include extracting structured information from rough text, as well as document classification and clustering.

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.

We study the impact of selecting (different) parts of documents for a specific clustering task. The idea is that different parts of XML documents correspond to different dimensions of the collection that may play different roles in the classification task. We carried some experiments in clustering homogeneous XML documents to validate an existing classification or more generally an organisational structure.

Our approach integrates techniques for extracting knowledge from documents with unsupervised classification of documents. The goal of unsupervised classification (or clustering) is to identify emerging classes that are not known in advance. We focus on the feature selection used for clustering and its impact on the emerging classification. This approach differs from other ones in two respects : first we mix the selection of structured features with the selection of textual features, second this last selection is based on syntactic typing be means of a tagger. We use TreeTagger, a tool for annotating text with part-of-speech and lemma information that has been developed at the Institute for Computational Linguistics of the University of Stuttgart [77].

Based on the selected features the documents are then clustered using a dynamical classification algorithm that builds a prototype of each cluster as the union of all the features (words) of the documents belonging to this cluster. 6 gives an example of discriminating keywords that are generated for each cluster. They can act as summaries for the clusters.

We illustrate and evaluate this approach with a collection of 139 XML activity reports written by Inria research teams for year 2003. The objective is to cluster projects into larger groups (Themes), based on the keywords or different chapters of these activity reports. We then compare the results of clustering using different feature selections, with the official theme structure used by Inria between 1985 and 2003, and with the new one proposed officially in 2004.

Cluster_1	3d approximation , computer , differential , environment , modeling , processing , programming , vision
Cluster_2	computing , equation , grid , problem , transformation
Cluster_3	code , design , event , network , processor , time , traffic
Cluster_4	calculus , database , datum , image , indexing , information , integration , knowledge , logic , mining , pattern , recognition , user , web

Figure 6. Experiment K-F-a: list of keywords by clusters

The results that will be published in the EGC 2005 conference show that the quality of clustering strongly depends on the selected document features. In our collection of research reports, clustering using *foundation* sections always outperforms clustering using *keywords*.

as an indication for the organization that some parts of the Activity Report do not appropriately describe the research domains and that the choice of keywords and research presentation could be improved to carry a stronger message.

Although the analysis is closely related to our specific collection, we believe that the approach can be used in other contexts, for other XML collections (such as the Inex collection of IEEE articles) where some knowledge of the semantic of the DTD is available.

6.3.3. A Complete Methodology for InterSites Web Usage Mining

Keywords: *HTTP logs*, *pre-processing*, *usage mining*.

Participants: Doru Tanasa, Brigitte Trousse.

In the recent years, the Web Usage Mining (WUM) emerged as a new field of Data Mining and gained an increasing attention from both the business and research communities. Following a survey of the main WUM techniques, we propose in earlier works, a complete methodology for data preprocessing in inter-sites WUM which was published in [19] and [20]. Our first objective is to reduce in a significant but pertinent manner, the size of the Web servers log files. The second objective is to increase the quality of data obtained after the classical preprocessing step by means of an original advanced data preprocessing step. To validate the efficiency of our method, we have conducted an experiment using the log files of Inria's Web sites: we joined and analyzed together the log files collected from four of Inria's Web servers.

In 2004, we develop AxISLogMiner which supports our methodology for preprocessing Web logs for inter-sites Web Usage Mining and integrates various algorithms for extraction sequential patterns developed in the team.

AxISLogMiner URL= http://www-sop.inria.fr/axis/axislogminer/

6.3.4. Hybrid Methods for Web Usage Mining: Improvements

Keywords: *low support*, *neural network*, *sequential pattern*, *user behaviour*, *web usage mining*.

Participants: Florent Masséglia, Doru Tanasa, Brigitte Trousse.

In 2004, we made some improvements of our two hybrid methods for extracting sequential patterns with a low support (cf. 2003 AxIS activity report).

Cluster & Divide : automatic mining for sequential patterns

Related to our Cluster & Discover method implemented in the C& D application [18], we added a new feature, "Mixed Mode", that allows automatic mining for sequential patterns in Web Logs. This feature allows the mining process to run independently (non-interactive) so the user only needs to specify the parameters for clustering and the minimum support. The minimum support is then recalculated according to the size of the

cluster and the sequential patterns are extracted from the cluster. Once the sequential pattern mining is done for all the clusters the user can visualize and explore the results.

Divide & Discover : improved performance in the number of divisions

We observed that the divide and discover method ([43],[16]) implemented in the D&D application needed numerous divisions before obtaining significant results. Actually, the clustering process was based on the extracted sequential patterns. Sequential patterns are not numerous and the number of clients not being classified was large. Even if the support was lowered, the number of sequential patterns did raise, but their length grew up. So the sequential patterns can be either not numerous or too long. In both of these cases, the clustering based on the sequential patterns extracted is not very efficient at the beginning.

In order to solve this problem we added the "just items" feature to the D& D application, which allows to extract only frequent items for the first division. There can be a large number of frequent items with a low support and they are easy to extract. Based on these frequent items we provided a new clustering process which is more efficient because items are not as specific as sequential patterns (because there is no combination between items as there can be within sequential patterns). The clustering is thus stronger and there are less non classified clients.

As a global consequence the number of divisions has been significantly reduced.

6.3.5. Applying our Data Mining Methods on Inria Web Data

Keywords: *2-3 hierarchies*, *WUM data analysis*, *Web data*, *clustering*, *contingence table*, *dissimilarity*, *dynamic clustering algorithm*, *preprocessing*, *self-organizing map*, *unsupervised clustering*, *unsupervised clustering*.

Participants: Sergiu Chelcea, Aicha El Golli, Mihai Jurca, Yves Lechevallier, Fabrice Rossi, Brieuc Conan-Guez, Yves Lechevallier, Doru Tanasa, Brigitte Trousse, Rosanna Verde.

In 2004, we tested our different Data Mining methods (presented in the section 6.2) on Inria logs as reported below:

Applying Crossed Clustering Method on Web Data

An application ([40],[41]) on the web log data from Inria web server allows to validate the proposed procedure and to suggest it as an useful tool in the Web Usage. The log file has been processed in order to record the navigations on both URL: *www.inria.fr* and *www-sop.inria.fr*.

This study aims to detect the behavior of the users and, in the same time, to check the efficacy of the structure of the site. Behind the research of typologies of users, we have defined a hierarchical structure (taxonomy) over the pages at different levels of the directories. The analyzed data set has concerned the set of *page views* by visitors that were connected to the Inria site from the 1st to the 15th of January, 2003. Globally, the database contained 673.389 clicks (like *page views* in an user session), whichhave been already filtered from robot/spider entries and accesses of graphic files.

The data are collected in two tables where each row contains the descriptions of a symbolic object (navigation), that is the distribution of the visited topics on the two websites. Following our aim to study the behavior of the Inria web users, we have performed crossed clustering analysis to identify an homogeneous typology of users according to the sequence of the visited web pages, or better, according to the occurrences of the visited pages of the several semantic topics.

The results of the navigation set partition in 12 classes and of the topics one in 8 classes, constituted by the two partitions Q^1 and Q^2, are shown in the Table 7.

Results in Table 2 is a example of automatic clustering procedure to structure complex data that performs simultaneously typologies of navigation and groups of topics, homogenous from a semantic point of view.

Applying Kohonen maps on Web data

Topic_1	Topic_2	Topic_3	Topic_4
/www/partenaires	/www/projets	/www/presse	/www/dias
/www/agos-sophia	/www/rrrt	/www/actualites-siege	/sop/dias
/www/modeles	/www/w3c	/www/multimedia	
/sop/partenaires	/www/manifestations	/www/icons	
/sop/agos-sophia	/sop/projets	/www/fonctions	
/sop/color	/sop/sophia	/sop/chir	
/sop/interne-sophia	/sop/site-eng	/sop/direction	
/sop/wiki	/sop/externe		
/sop/modeles	/sop/colloquium	Topic_7	Topic_8
/sop/sapr	/sop/horde		
/sop/didacticiel	/sop/manifestations	/www/recherche	/www/sophia
/sop/ctime	/sop/international	/www/accueil-siege	/www/site-old
/sop/freesoft		/www/personnel	/sop/cgi-bin
		/www/intro-inria	/sop/commun
		/www/publications	/sop/accueil-sophia
Topic_5	Topic_6	/www/cgi-bin	/sop/intro-sophia
		/www/ra	/sop/actualites-sophia
/www/travailler	/www/rapports	/www/interne-siege	/sop/rev
/www/formation	/www/semir	/www/international	/sop/intech
/www/valorisation	/sop/rapports	/www/site-beta	/sop/services
/sop/formation	/sop/semir	/www/sophia-antipolis	/sop/challengeTV
/sop/recherche	/sop/rmi	/www/thesauria	/sop/xml

Figure 7. Topic descriptions groups

	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8	pages
Navigation_ 1	222	1470	587	34	611	80	18757	143	21904
Navigation_ 2	78	2381	254	7	3094	80	2055	249	8198
Navigation_ 3	8578	7767	425	309	448	2749	2091	1386	23753
Navigation_ 4	29	280	115	7	3387	7	347	91	4263
Navigation_ 5	209	242	9	26	23	2544	221	55	3329
Navigation_ 6	29	1185	3204	28	1247	19	2670	82	8464
Navigation_ 7	43	140	22	795	39	47	218	636	1940
Navigation_ 8	288	35742	920	90	594	308	2174	1101	41217
Navigation_ 9	186	1040	136	106	283	72	370	3739	5932
Navigation_10	24	39	6	2786	2	25	49	210	3141
Navigation_11	175	7630	606	87	574	326	10227	257	19882
Navigation_12	4	231	3088	4	96	8	179	10	3620
Total pages	9865	58147	9372	4279	10398	6265	39358	7959	145643

Figure 8. Contingence table of the navigation classes and topic groups

We have applied our adaptation of the Self organizing map (SOM) to complex data and specially to dissimilarity data [33] (cf. 6.2.1) on real data by clustering data issued from the same two Inria's Web servers [34].

The first analysis concerns the users navigations that are composed by a set of first level visited topics. To clusters the navigations we used the affinity coefficient between two navigations. For more details, see [12].

The second analysis concerns the clustering of the first syntactic topics in order to find an association between them. We have also applied our method to the same data set of the 2-3 AHC described in [27]. We used also the Jaccard index to cluster Inria's visited first level topics (in particular the research teams).

In the final map, shown in Figure 9, the neurons have been labeled according to the semantic topic of the individual referent. For the semantic topic "research team" we also represent the theme to which it belongs. It showed that the research teams of theme 1 were mapped to neighboring neurons in the map, as well as the research teams of theme 4 and the scientific events.

scientific events	research team(Theme 1)	research team(Theme 3)	Inria
scientific events	research team(Theme 1)	research team(Theme 4)	research team(Theme 2)
research team(Theme 2)	research team(Theme 4)	research team(Theme 4)	research team(Theme 4)

Figure 9. A 4-by-3-unit rectangular map.

The SOM was constructed using the 196 first syntactical topics. Each neuron contains the semantic topic of the referent topic. For the research teams we also represent the theme which they belong to.

Applying 2-3 AHC on Web data

We have applied our 2-3 AHC algorithm [24] (cf. 6.2.6) on real data by clustering data issued from two Inria Web servers [27]. More concretely, we have clustered Inria visited topics (in particular the research teams) based on its Web sites' visitors behavior. Knowing that Inria scientific organization has changed on 1st April 2004, our goal was to analyze the impact of the Web site structure on users navigations before and after this change (two 15 days periods).

To cluster the topics from the visited URLs, we used the Jaccard index on users navigations (sets of URLs) during the advanced data preprocessing of Web data. Our analyses revealed:

- The *global* impact of the Web site on users navigations. For example, in when analyzing the first level visited topics, 16 out of 19 formed clusters contained research teams from same research theme [27] (cf. Table 10).
- The impact of the *scientific organization* : we found that the research teams clustering was different for the two analyzed periods and was highly influenced by Inria former and new scientific organization into research themes (cf. Figures 11a and 11b).

robotvis SOP **3B**, robotvis **3B**, epidaure SOP **3B**, odyssee SOP **3B**, epidaure **3B**, ariana SOP **3B**, ariana **3B**	comore SOP **4A**, icare SOP **4A**, icare **4A**, miaou SOP **4A**, reves SOP **3B**, miaou **4A**, chir SOP **4A**, comore **4A**, caiman SOP **4B**	orion SOP **3A**, axis SOP **3A**, orion **3A**
prisme SOP **2B**, prisme **2B**	koala SOP **2A**, koala **2A**, croap SOP **2A**, croap **2A**	odyssee **3B**, dream SOP **3A**, lemme **2A**, opale SOP **4B**, opale **4B**, certilab **2A**, pastis **3B**
orion SOP **3A**, acacia SOP **3A**, acacia **3A**, axis SOP **3A**, orion **3A**, aid SOP **3A**, aid **3A**	coprin SOP **2B**, saga SOP **2B**, saga **2B**	sinus SOP **4B**, sinus **4B**, smash SOP **4B**
robotvis SOP **3B**, robotvis **3B**, odyssee SOP **3B**	mimosa SOP **1C**, mimosa **1C**, tick SOP **1C**, tick **1C**	sloop SOP **1A**, sloop **1A**, oasis SOP **2A**, oasis **2A**
rodeo SOP **1B**, rodeo **1B**, planete SOP **1B**, planete **1B**	lemme SOP **2A**, tropics SOP **1A**, mascotte SOP **1B**, omega SOP **4B**, galaad SOP **2B**, cafe SOP **2B**, certilab SOP **2A**	mistral SOP **1B**, mistral **1B**
mefisto SOP **4B**, mefisto **4B**	mascotte SOP **1B**, mascotte **1B**	safir SOP **2B**, safir **2B**
meije SOP **1C**, meije **1C**		

Figure 10. Inria's Web site topics clustering using 2-3 AHC

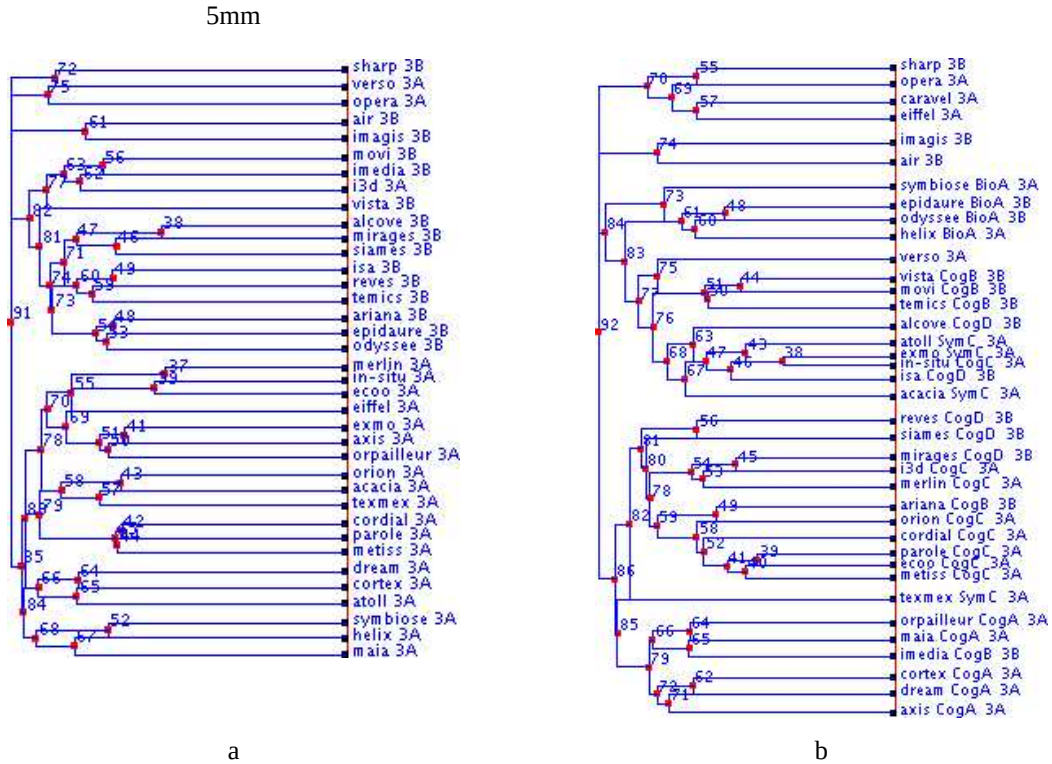

Figure 11. a) Theme 3 projects on first period. b) Theme 3 projects on second period

Figure 12. Be-TRIP recommender system

6.3.6. Personalized Recommendations for Mobility Information Retrieval

Keywords: *personalization, recommender system, trip, user profile.*

Participants: Sergiu Chelcea, Brigitte Trousse.

As members of the MobiVIP project (PREDIT 3), we studied an emerging research field related to mobility in the transport domain which is related to travel information retrieval.

To facilitate such retrieval, in collaboration with Georges gallais (Visa Action, Inria) we propose in [26] and [21] the use of recommender systems in a mobility context: these systems facilitate information retrieval, and support the preparation of the user's journey ("pre-trip": choice of the transport mode, schedule, route, time of the trip, ...) and to carry it out ("on-trip": interactive guidance, way visualization, destination planning). Compared to the state of the art, the originality of our approach lies in:

- its recommendation calculation based on pre-trip and on-line pretrip logs,
- its capacities to adapt the recommendations to the user's behavior during his information retrieval correlated to his own movement,
- the on-line learning capabilities for supporting information retrieval.

Based on such an approach (and our first recommender Broadway-Web), we specify in 2004 the Be-TRIP recommender system and start the implementation with the pre-trip mode. Figure 12 shows a recommandation related to the bus schedule from Valbonne to Antibes for a tourist browsing inside the CASA site (we developed) and particularly in the Valbonne and Antibes sites.

6.3.7. Multi-disciplinary Approach of Internet Measures

Keywords: *internauts practices, internet, metrology.*

Participants: Eric Guichard, Brigitte Trousse, Florent Masséglia, Doru Tanasa, Yves Lechevallier.

Based on the main contributions of the workshop "Mesures de l'internet" organised at Nice in may 2003, E. Guichard supervises a collective book [11] published by "Les Canadiens en Europe" in April 2004.

This book describes various approaches issued from mathematics and computer science (linked to the internet metrology), linguistics, geography and human sciences related to the internet practices. Let us note two AxIS contributions ([14], [17]) in this book.

This work is related to the pragmatic and pluri-disciplinary approach adopted by the team from its creation in order to have a better understanding of the user practices of internet: the benefits of such an approach concern the definition of relevant evaluation criteria of web-based information systems (or relevant usage analysis variables) and also relevant specifications in the (re)-design of such systems.

6.4. Other Applications

Other applications were developed mostly for validating our algorithms (as for [42], [50], [13]).

6.4.1. Comparison of Sanskrit Documents

Keywords: *Sanskrit, text comparison, transliteration.*

Participants: Marc Csernel, Yves Lechevallier.

This research is carried out in the context of the CNRS Action "Histoire des savoirs" (History of Knowledge).

The goal of the projects is to produce software tools that support the construction of critical edition of Sanskrit texts. A critical edition is a document that shows all different versions of a text found in different manuscripts. Generally the critical edition of a text is the base for all further studies.

Sanskrit texts contain some unique features that make inefficient standard tools dedicated to Indo-European languages. First, Sanskrit uses a specific 46 letters alphabet. Sanskrit scripts must be transliterated into roman

scripts to be used on a computer; usual software tools compare the roman characters of the transliteration and do not use the Sanskrit alphabet directly.

Second, Sanskrit texts (especially in ancient manuscripts) can be written without spaces between words. This made comparisons between texts quite complex since it is hard to separate words. To avoid this difficulty we use two kinds of text: a lemmatized master text (called the padapatha) and the text to be compared. This will greatly improve the algorithmic complexity, but will introduce some new difficulties. Indeed, when spaces between two words are suppressed, the two words are not simply glued. They are modified according to some special rules call Sandhi. Taking Sandhi into account is not a trivial task.

During 2004, we developed the comparison software. The approach was to adapt the velhuis transliteration system in order to compare the words according to the Sanskrit alphabet and not with the roman alphabet. The implementation is using mostly Lex.

The second step was to allow comparison between text of different structures, the "normal text" and the lemmatised text: before comparison, the lemmatised text is transformed according to Sanskrit Rules. This transformation needs to be done for each comparison, because only the lemmatised text can indicate in which word a difference occurs. We have implemented most of the Sandhi rules, although there is still some work left. Once the implementation of the Sandhi rules is completed, we will test them on real examples.

The software is now able to make comparison between the "master lemmatised text" and another text, taking into account the Sanskrit alphabet as well as sandhi rules, but the algorithm needs some optimisation. We plan to increase the its performance by adapting the DIFF algorithm to deal with Sanskrit characteristics.

6.4.2. Using GrepMiner on Gene Regulatory Expression Profiles

Keywords: *Apriori, DNA chip, DNA microarray, affymetrix GeneChip, differential expression, expression analysis, gene expression, sequential pattern mining, suffix tree.*

Participants: Doru Tanasa, Brigitte Trousse.

Given the advent of microarray technology, it is now possible to analyze the expression of a large number of genes simultaneously. Microarray experiments can be classified according to the nature of the samples, i.e. time of collection, location, type of tissue, class of tumor, etc. In the paper [50] we are interested in exploring our computational methods when applied to time series microarray experiments. In particular we report results applied to gene expression time series associated to mouse cerebellum development [72].

Biological motivation and gene expression data generation

The time-series gene expression data was generated by Kagami et al [72]. The data is publicly available through GEO [2], http://www.ncbi.nlm.nih.gov/geo/. In such study Kagami et al [72] investigated differentially expressed genes during the development of mouse cerebellum. Their biological interest was focused to further understanding the molecular basis of mouse cerebellum development. The mouse cerebellum is not entirely developed until post-natal day 21, therefore their experiment was an ideal framework for the understanding of the genetic foundations and mechanisms of neural development.

Sequential Patterns Discovery in microarray Data.

We propose APRIORI-GST, an APRIORI-like algorithm that uses a Generalized Suffix Tree (GST) index for discovering sequential patterns from microarray data. The microarray data is transformed into sequences of three possible levels of exposure (e^+, e^0 or e^-). These sequences are indexed using a GST index. A microarray sequential pattern may be seen, in this case, as a sub-sequence of levels of exposures that frequently occur.
From the extracted patterns we outlined the hypothesis that there is a lot of gene activity between the prenatal stage E18 and postnatal stage P7, which needs to be further investigated.

GREPminer Software Tool

To support our methodology, we designed and implemented in Java, the GREPminer[3] tool presented in Fig. 13. The user chooses a dataset file and extracts sequential patterns having the support superior to a specified

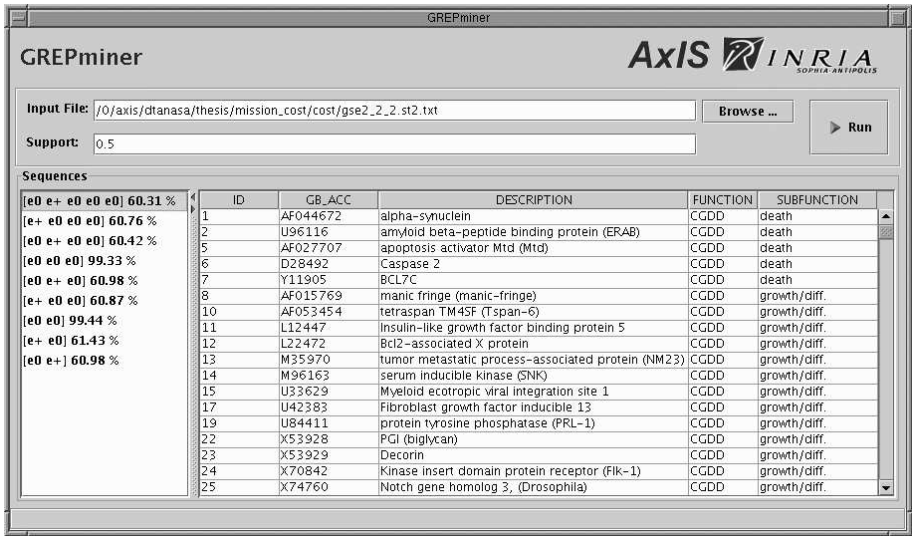

Figure 13. The GREPminer Software Tool Implementing the Apriori-GST Algorithm

threshold. The extracted frequent sequential patterns are listed on the left side and the details (list of genes) for the selected pattern is displayed on the right side.

7. Contracts and Grants with Industry

7.1. Industrial Contracts

7.1.1. EPIA : a RNTL Project (2003-2005)

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 for two years and half. Partners are Dalkia, Mediapps and Inria. We continued the work on this contract in the following directions:

- Supporting users of MediappsNet via clustering clients. This task has started this year and some generic algorithms and pre-treatment tools were developed [53] and some log analysis.
- Understanding the user needs for NetPortal. To achieve this job, two main actions were made: a) some meetings were organised with eDalkia users and project managers at Dalkia Ile de France and Dalkia headquarter; b) some log analysis was made from a clone of the database used behind the eDalkia system (solution Net.Portal). The purpose of this analysis was to identify the needs of eDalkia users but also to have a good understanding of the user actions currently logged in the database.

[2] *Gene Expression Omnibus*
[3] Gene Regulatory Expression Profiles Miner

- Specifying the trace engine of NetPortal: a draft version of the deliverable D3: "Experimental context and trace engine in Net.Portal". The D3 deliverable formalizes the user profile and the users' actions in Net.Portal and provides the detailed specification of the trace engine. To achieve it, we had a reverse engineering task on the Net.Portal to understand the internal system organisation and the data storage in the database. The result of this work is the description of the Net.Portal relational database schema and the data organization included in the deliverable D3.

We held various AxIS project meetings in Paris during this year (the last one was on 20 October 2004 involved Brigitte Trousse, Mihai Jurca, Y. Lechevallier, Aicha El Golli and Luc Beaubois).

The Epia project was presented by the project manager P. Snieg (Mediapps) via a poster at the national RNTL workshop at Rennes in october.

7.1.2. MobiVIP : a PREDIT Project (2004-2006)

Inria Contract Reference: 2 03 A2005 00 00MP5 01 1

MobiVIP, Individual Public Vehicles for Mobility in town centres, 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 centres. This service is made up of small urban vehicles integrated into 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 and 3° the valuation of possible technology transfer.

URL= http://www-sop.inria.fr/mobivip/

This year we worked in collaboration with B. Senach (Ergomatics Consultants) mainly on the delivrable 5.1 which we co-coordonnate with Georges Gallais (Visa Action, INria Sophia Antipolis). This delivrable aims to define a common generic evaluation scenario and proposes a framework to facilitate the identification of the main evaluation dimensions for each planned test or experimentation. The user point of view was mainly developped in such a delivrable [64]. To achieve this deliverable, we design a questionnary to be fullfilled by each partner concerning by the future expriments in order to capture the main evaluation dimensions manipuled by each of them and facilitate the design of a common framework for evaluation purposes.

7.1.3. Industrial Contacts

In the context of industrial visits organised by Inria, we present AxIS researches

- at Inria Sophia Antipolis, AxIS presentation by B. Trousse to Hitachi (June 1st), Ricoh - Japan (February 25) and Accenture (nov. 10, M. Gershmann, Research Mananger and others members from the center of Sophia Antipolis).
- at Inria Futurs, AxIS presentation by Y. Lechevallier to SAP (june 16)

Some others contacts during this year:

- ATOS Origin: a proposal related to CBR was made together for DGA (B. Trousse)
- CIMPA related to CBR*Tools (B. Trousse)
- Mondeca and VisioLab (2 meetings at Lip6)(Y. Lechevallier)
- Contacts with others industrial partners such as France Télécom Sophia Antipolis, Cnes Toulouse (B. Trousse)

8. Other Grants and Activities

8.1. Regional Initiatives

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

- University South Toulon (LePont laboratory) and LIRMM (IHMH team): a proposal was submitted and accepted to the program COLOR 2005, related to defining and evaluating new Web pages ranking criteria based on page presentation.
- Inria VISA Action: 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”. The MobiVIP proposal, involving 22 partners, started January.
- Laboratoire des Usages, CNRT Télius Sophia Antipolis. B. Trousse is a member of the scientific committee and a substitute member of the management committee.
- Supelec: Research collaboration with Marie-Aude Aufaure on methods for building automatically ontologies [23] and more precisely one travel ontology from Web sites.

8.2. National Initiatives

AxIS is involved in several national working groups.

8.2.1. CNRS RTP 12: << information et connaissance: découvrir et résumer >>

In the context of the pluri-disciplinary thematic network 12, H. Behja, F.Masséglia and B. Trousse participated to the CNRS Specific Action (AS 120) “Disco Challenge” animated by J.F. Boulicaut and B. Crémilleux. They attended the worshop held at EGC04 in January where H. Behja presented the work done in his PhD thesis [55]).

8.2.2. CNRS RTP 15: << économie, organisation & STIC >>

In the context of the pluri-disciplinary thematic network35, our team participated in the CNRS Specific Action (AS 140) "Données dynamiques et mesures des flux sur internet", animated by L. Lebart (CNRS & ENST Paris) and V. Beaudoin (FT R&D). M. Csernel, E. Guichard, Y. Lechevallier, F. Rossi, A-M. Vercoustre participated to the seminar held in Paris (September, 23). Doru Tanasa maintains the web server for this action.

8.2.3. CNRS RTP 33: << DOC >>

This CNRS STIC - SHS departments network called “RTP DOC” (http://rtp-doc.enssib.fr/) started the 31 may 2002 under the animation of Jean-Michel Salaün (Enssib, Lyon). E. Guichard participates related to the reflexion of internet measures, initiated during the workshop “Mesures de l’internet we organized in Nice [11].

8.2.4. CNRS << Action Concertée : Histoire des savoirs >>

This initiative 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.

8.2.5. EGC << National Group on Mining Complex Data >>

URL: http://morgon.univ-lyon2.fr/GT-FDC/

AxIS members participated actively this year to the Working Group “Fouille de données complexes” created by D.A Zighed in June 2003 in the context of the EGC association:

1. B. Trousse with P. Gancarski (LSIIT, Strasbourg) co-organised and co-chaired the first workshop “Fouille de données complexes dans un processus d’extraction de connaissances” (January 23-24, 2004) [10]. Y. Lechevallier and F. Masséglia were members of the program committee.

2. F. Masséglia with O. Boussaid co-animated the topic "organisation and structuration of complex data" of the national group on mining complex data. He co-organised a working day on March 16, 2004 at ENSAM, Paris. B. Trousse presented AxIS works related to the notion of complexity of a Web usage Mining process. Other AxIS participants: Y. Lechevallier, A. El Golli, F. Rossi, B. Conan-Guez, M. Csernel.
3. F. Masséglia and B. Trousse participated in the main meeting of the working group held in Lyon (ERIC) on September 17, 2004.

8.2.6. GDR-I3

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

- GRACQ (*Groupe de Recherche en Acquisition des Connaissances*) (GRACQ): B. Trousse.
- Working Group 3.4 (GT) on Data Mining animated by P. Poncelet and J.M. Petit. F. Masséglia participated to the meeting in at LIRIS (Lyon), on 15th November 15.
- Working Group 3.1 "Sécurité des Systèmes d'Information" animated by D. Boulanger and A. Gabillon: F. Masséglia and B. Trousse.

8.2.7. Other Collaborations

- LIRMM (Montpelier) and Ecole des mines d'Alès (LGI2P) : F. Masséglia with M. Tesseire (LIRMM) and P. Poncelet (LGI2P) proposed two surveys, one related to sequential pattern mining method and issues [15], the other related to the management of time constraints in the generalized sequential pattern extraction process [44].
- ENST: Y. Lechevallier collaborated with Georges Hébrail (ENST) on [38]
- via two ACI proposals :1) ACI Masse de données "Mining Complex Data" with D. Zighed (ERIC, Univ Lyon 2), Mohand-Said Hacid (LIRIS, UNiv Lyon 1), H. Briand (LINA, Nantes) et Y. Kodratoff (LRI Univ Paris Sud); 2) ACI "FOUINE" with arie-Aude Aufaure (Supelec), K. Zeitouni (PRISM, Versailles) and C. Claramunt (Irenav Ecole Navale).
- Institut de recherche pour le développement (IRD)(Agathe Petit, anthropologue, post-doctoral position): E. Guichard
- E. Guichard has various contacts this year. Let us cite researchers in cybergeography: Henri Desbois (U. Paris-X), Loïc Grasland (U. Avignon), Jean-Claude Moissinac (ENST, Paris), Hervé Théry (Mappemonde, CNRS), Dany Vandromme (RENATER, Paris).

8.3. European Initiatives

8.3.1. IST European Network : Ontoweb

AxIS participated in the European network Ontoweb (Ontology-based Information Exchange for Multilingual Electronic Commerce and Information Integration) proposed in 2000 by Dieter Fensel (Division of Mathematics & Computer Science, Vrije Universiteit Amsterdam).
The project ended in May 2004 with an excellent evaluation.

8.3.2. IST European Project: ASSO

URL=http://www.info.fundp.ac.be/asso/index.html

ASSO was a project of the European Union Fifth framework Research and Development program in the Information Society Technology strand, number IST-2000-25161. ASSO offered methods, methodology and software tools for the analysis of multidimensional complex data (numerical or non numerical) coming from databases in statistical offices and administrations using Symbolic Data Analysis. ASSO ended this year.

ASSO was based around nine working groups ("Workpackages") with the participation of fifteen partners ("Consortium").

AxIS contributions to program

- **Benchmark and prototype evaluation** The first part in this task consists in the benchmark definition: choice of data, variables, population, encoding, weights, rules and taxonomies with a special interest to data on unemployment, social insertion as well as business registers.
- **Data Format SOM library** Every algorithm within the project uses the SOM library, which interfaces the different algorithms with the data. They will take into account the metadata, introduce a new representation in order to provide to the different algorithms a way to use the knowledge provide by the rules.
- **Metadata Model Design included in SOM** The mission of this task is to design a semantically rich metadata model that will hold additional information for the assertions that construct the SO. Such a model should be carefully designed in order to ensure that the data consumers needs are not underestimated (missing information) or that data providers are not forced to allocate resources in capturing metainformation of little value (overestimation).
- **DB2SO**, which creates SO from a database, has a first version in the previous project but many improvements have to be done. In the new version every groups of individuals is not generalized by one Symbolic Description but by many disjunctive Symbolic Descriptions.
- **DIV, SCLUST Partitional clustering** In this task we propose clustering algorithms in order to partition a set of SO into a predefined number of classes. The classes are interpreted and represented by suitably generalized class prototypes (again in the form of symbolic objects).
- **Interpretation of partition, robustness of methods and cluster validation included in SCLUSTER** We select a cluster and interpret it for a selected set of variables. We will measure the robustness of a cluster and a symbolic object related to extent, intent dissimilarity and algorithm. We will measure the degree of isolation and compactness of each cluster and its symbolic intent.

8.3.3. COST Action 282

In 2004, we participed actively in the COST Action 282 (2001-2005): “Knowledge Exploration in Science and Technology”.

URL=http://www.mpa-garching.mpg.de/~opmolsrv/COST282/

- Organisation of the KELSI 2004 workshop in Nice in January 2004 : S. Honnorat and B. Trousse URL= http://www-sop.inria.fr/axis/cost282/kelsi04/
- Short Term Cost missions to the Bioinformatics Research Group at the University of Ulster (Nth. Ireland) July 2004: D. Tanasa and B. Trousse
- Writing a common paper with Jesus Lopez accepted to the KELSI04 Symposium (Milano) [50].
- Participation in two Management Committees (January in Nice, November in Milan): B. Trousse is one of the two French representatives in the Management Committee.

8.3.4. EuropAid project: Sanskrit

An Asia-Information Technology and Communications project (EuropeAid) has been submitted and accepted by the EEC. This project aims to design a most advanced IT tool for an "archaeology of ancient Asian texts" (mainly written in Sankrit). It will start in 2005 for one year.

Consortium members of this projet are: 1) Inria, AxIS team 2) Bhandrakar Institute, India, 3) University Sapienza di Roma, Italy and 4) Mahendra Sanskrit University, Nepal.

8.3.5. Other Collaborations

- Italy, University of Napoli II (Profs C. Lauro and R. Verde) [41][40] : Y. Lechevallier, A. El Golli, M.Csernel
- Belgium, Facultés Universitaires Notre-Dame de la Paix à Namur (Profs A. Hardy, M. Noirhomme and J.-P. Rasson) [60]: Y. Lechevallier.
- Belgium, Université Catholique de Louvain-la-Neuve, DICE Laboratory (N Delannay, M Verleysen) [31][42] : B. Conan-Guez, F. Rossi

8.4. International Initiatives

8.4.1. Australia

A-M. Vercoustre collaborates with RMIT, Computer Sciences department, Melbourne, Australia, where she is co-supervising a student working on XML search (Jovan Pehcevski) [45]. In this context, she participated in the Initiative for the Evaluation of XML Retrieval (INEX-2004), DELOS network of Excellence. [46]

A-M. Vercoustre collaborated with CSIRO-ICT, Melbourne, on the application of Documents technologies for reusing educational material.

A-M. Vercoustre organised two seminars with australian invited speakers (cf. section 9.1.4).

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 (Récife) and his team. We welcomed F.A.T. de Carvalho, in March-April and September. During this year some analysis of Recife Log files are carried out by students from Recife University, Napoly University and by AxIS members.

8.4.3. Canada

Y. Lechevallier pursues his collaboration with A. Ciampi (Univ of McGIll, Montréal) [39]. E. Guichard has various contacts with Jacques Lajoie (UQAM, Canada), David Olson (U. Toronto), Serge Proulx (UQAM).

8.4.4. China

A workshop on Complex Data Analysis was organised by Pr. Wang, Huien and Pr. Siwei, Cheng (Beijang University, Beijing, October 27-29) with the financial support of the National Nature Science Foundation of China., the Beijing Univ. of Aeronautic and Astronautic, Inria and the Univ. of Naples (Frederico II). E. Diday (Dauphine University), G. Hébrail (ENST), J.-P. Nakache (Inserm), G. Saporta (CNAM) and Y. Lechevallier (Inria) were the French speakers.

8.4.5. India

Via the CNRS action "History of Knowledge" and also via the consortium members of EuropeAid projet of Asia-Information Technology and Communications: Bhandrakar Institute, India and Mahendra Sanskrit University, Nepal.

8.4.6. Morocco

AxIS is involved in a France-Morocco thematic network in software engineering and, in this context, B. Trousse co-supervises with Abdelaziz Marzark (UNiversity of Casablanca) a Ph-D student: H. Behja (ENSAM, Meknès, Morocco). The Third Joint Meeting of France-Morocco Cooperation Program (<< Atelier STIC >>) between the three current thematic networks was help in Rabat, Dec 13-15 (B. Trousse and H. Behja). H. Behja presented his thesis work to the scientific committee.

8.4.7. Romania

We maintained our contacts with the Computer Science department of the West University of Timisoara (Prof Viorel Negru). S. Chelcea and C. Garboni visited V. Negru during one week during the period of the SYNASC Conference (Timisoara). For the future, we planned via the Brancusi PAI to initiate a collaboration on "High performance algorithms using grid computing for Data Mining".

8.4.8. Tunisia

Y. Lechevallier was invited in january by ENIT (professor Ben Ahmed) at Tunis for a tutorial on "Data Mining and neural Methods". Possible cooperations were studied via a co-supervision of future doctoral students.

9. Dissemination

9.1. Promotion of the Scientific Community

9.1.1. Journals

AxIS members are members of the editorial boards of two international journals and two national journals:

- 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, revue électronique http://www.jsda.unina2.it): Y. Lechevallier, F. Rossi and B. Trousse.
- the RIA journal (<< Revue d'Intelligence Artificielle >>) (Hermès publisher; editor in chief: M. Pomerol): B. Trousse.
- the I3 electronic journal of the GDR-I3 (editor-in-chief: C. Garbay et H. Prade) http://www.Revue-I3.org/: B. Trousse.

F. Masséglia and B. Trousse were invited editors (with O. Boussaid and P. Gancarski) of a Special issue of the RNTI (Revue des Nouvelles Technologies de l'Information) on "Complex Data Mining". Y. Lechevallier was a member of the editorial board and A-M. Vercoustre was an additional reviewer.
E. Guichard is a correspondant of the Mappemonde electronic journal
http://mappemonde.mgm.fr/.
AxIS members were reviewers for four international journals and one national journal:

- the KAIS journal in 2004 (Knowledge and Information Systems, for a special issue KDD MDM 2004 : B. Trousse.
(http://www.emba.uvm.edu/~xwu/kais/)
- the IEEE Journal of Transactions on Data and Knowledge Engineering (TDKE: F. Masséglia.
(http://www.computer.org/tkde/)
- the International Journal of Computer, Applications in Technology (IJCAT): F. Rossi.
(https://www.inderscience.com/browse/index.php?journalID=5)
- ACM Transactions on Information Systems, June 2004: A-M. Vercoustre.
- a Special issue of the RNTI (Revue des Nouvelles Technologies de l'Information) dedicated to selected and extended papers of the national conference SFC04 : B. Conan-Guez, Y. Lechevallier, F. Rossi and B. Trousse.

9.1.2. Program Committees

Several AxIS members were involved in Program Committees at national or international levels:

9.1.2.1. National Conferences/Workshops

- EGC'04 (Clermont-Ferrand, January) Extraction et Gestion des Connaissances: Y. Lechevallier and B. Trousse. (http://www.isima.fr/~egc2004/)
- Atelier Fouille de Données Complexes (at EGC'04) : B. Trousse (co-chair), F. Masséglia, Y. Lechevallier (http://www-sop.inria.fr/axis/fdc-egc04/). F. Rossi was an additional reviewer.
- INFORSID 2004 (Biarritz, May): B. Trousse.
(http://inforsid2004.univ-pau.fr/index.htm)
- Mobilité et Ubiquité (ESSI, June): B. Trousse (http://www.essi.fr/UbiMob/)
- SFC04 (Bordeaux, Sept 8-10): Y. Lechevallier. B. Trousse was an additional reviewer.
- BDA 2004 (Montpellier, October 19-22): F. Masseglia. (http://www.lirmm.fr/BDA2004/)

9.1.2.2. International Conferences/Workshops

- MCO 2004 (Metz, July) Modelling, Computation and Optimization in Information Systems and Management Sciences: Y.Lechevallier
- IFCS 2004 (Chicago, July) Meeting of the International Federation of Classification Societies: Y.Lechevallier
- SIGKDD/MDM 2004 (Seattle, August) International Conference on Knoweldge Discovery and Data Mining: B.Trousse
(http://www.acm.org/sigs/sigkdd/kdd2004/index.html)
- SFC 2004 (Bordeaux, September) 11èmes Rencontres de la Société Francophone de Classification: Y. Lechevallier
- ECML 2004 (Pisa, September) 15th European Conference on Machine Learning: F. Rossi was an additional reviewer.
- ECCBR 2004 (Madrid, August-September) 7th European Conference on Case-Based Reasoning: B. Trousse (http://www.idt.mdh.se/eccbr/)
- DocEng 2004 (Milwaukee, USA, October), The ACM Symposium on Document Engineering): A-M. Vercoustre (http://www.sdml.info/doceng2004/)
- INTELLCOMM 2004 (Bangkok, Thailand, Nov.), the 2004 IFIP International Conference on Intelligence in Communication Systems: A-M. Vercoustre
- KELSI'04 Symposium (Milano, Nov.) Knowledge Exploration and Life Science Informatics: B.Trousse

9.1.3. Invited Seminars

- ENIT Tunisia,Séminaire RAIDI : Y. Lechevallier ("Classification automatique d'objets décrits par un vecteur d'intervalles"), january.
- CEA, January 2004 : F. ("Méthodes neuronales et données fonctionnelles")
- GT Fouille de données complexes, journée nationale du thème 2, Paris, March 16: B. Trousse (Prétraitement des données Web pour l'analyse des usages").
- International workshop on symbolic data analysis, Univ. PARIS-IX Dauphine(May 6-7), three invited presentations: A.Balde, Y. Lechevallier, A. El Golli
- National Group on "Mining Complexe data", Data preprocessing Workshop : B. Trousse (" Notion de complexité relativement au prétraitement de logs Web"), 16 March, Paris.
- Journées CNRS Histoire des Savoirs": M. Csernel in cooperation with Gerdi Gerschheimer and Pascale Haag and François Patte, ("Grammaire et mathématiques dans le monde indien"), May.
- Journée "XML et Statistique", July 17, organised by the French Society of Statistics (SFdS), groupe InfoStat, Data Mining et Logiciels: Th. Despeyroux ("Analyse sémantique de sites Web et de documents XML").
- Smash group (GRIMM team), University of Toulouse II Le Mirail, Nov. 2004 : F. Rossi ("Algorithmes neuronaux pour le traitement des données fonctionnelles").
- M. Csernel, "Normal Symbolic Form", Invited Conference, CIN UFPE Recife Pe Brazil, november.
- Machine Learning Group (DICE Laboratory), Université Catholique de Louvain-la-Neuve (Belgium), Nov. 2004 : F. Rossi ("Web Usage Mining with the Median Self Organizing Map").
- Dauphine Lise-CEREMADE, December 2004 : A. El Golli and Y. Lechevallier ("Application des méthodes de classification dans le cadre du Web Usage mining").
- F. Rossi, Invited Conference, CIN UFPE Recife Pe Brazil, december.

9.1.4. Organization of Conferences or workshops

- Organisation of the KELSI workshop: B. Trousse and S. Honnorat (in collaboration with W. Dubitsky (Ulster University, Ireland) and M. Simonetti (Inria).
  http://www-sop.inria.fr/axis/cost282/kelsi04/
- Co-organisation of a workshp on " Mining Complex Data" at the EGC04 conference : B. Trousse (in collaboration with P. Gancarski).
  http://www-sop.inria.fr/axis/fdc-egc04/
- Organisation of an half day June 24 (ENS Paris) related to the presentation of the book « Mesures de l'internet 2004 » [11][14]: E. Guichard.
- Organisation of two seminars at Inria Rocquencourt: A-M. Vercoustre
  - "Towards more effective enterprise search", by David Hawking, CSIRO- ICT center, Canberra, Ausralia, 4th August 2004, Rocquencourt.
  - "Identifying influences, ideas and opportunities by data mining citations using formal concept analysis", by Pr. Peter Eklund, University of Wollongong, Australia, 1st. December 2004, Rocqencourt.
- Organisation of our annual AxIS workshop at Inria Sophia Antipolis (22-24 November) : S. Honnorat and B. Trousse at . From April 2004, monthly team meetings were organised by videoconferecne between AxIS Sophia Antipolis and AxIS Rocquencourt.

9.1.5. AxIS Web Server

AxIS maintains an external and an internal Web sites 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.
http://www-sop.inria.fr/axis/.

9.1.6. Activities of General Interest

- T. Despeyroux is 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.
- Th. Despeyroux is participating in the project for redesigning the intranet Web site of Inria-Rocquencourt.
- E. Guichard was a member of the evaluation committee for the call for Proposals related to "Usages d'Internet" (Research Ministry, France).
- B. Trousse is a member of the scientific committee and also a substitute member of the decision committee of the "Laboratoire des Usages des NTIC" of Sophia Antipolis.
- B. Trousse is a member of the RSTI scientific committee related to the << ISI, L'OBJET, RIA, TSI >> journals (Hermès publisher).
- 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 an Open Archive in cooperation with CNRS.

9.2. Formation

9.2.1. University Teaching

The team members are teaching in various university curriculum, and AxIS is an associated team for the "STIC Doctoral school" at the University of Nice Sophie Antipolis (UNSA):

- "DEA Informatique" (resp. Mr Kounalis) at UNSA Sophia Antipolis: Optional tutorial on "Web usage Mining" (F. Masséglia, 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, 10 students, 30h supervised) on *Mining HTTP Logs From Inria's Web Sites* : S. Chelcea, D. Tanasa, B. Trousse.
- DEA of Social Sciences ENS-EHESS, *"Lectures de l'internet"*: E. Guichard (resp.).
- "DEA Modélisation et traitement des données et des connaissances" (resp: S. Pinson) of the University Paris IX-Dauphine (4h): Tutorial on *"Analyse des connaissances numériques et Symboliques"*: Y. Lechevallier.
- "DESS Mathématiques appliquées et sciences économiques (MASE)" of the University Paris IX-Dauphine: Tutorial (18h) on *"Méthodes neuronales en classification"*: Y. Lechevallier.
- "ISUP" of the University of Paris 6: Tutorial on *"Méthodes de classification et de classement"* (30h) : Y. Lechevallier.
- "ENSAE": Tutorial on *"Data Mining"* (12h): Y. Lechevallier.
- "DECISIA" : Tutorial on *"Neural Networks"* (7h) : B. Conan-Guez.
- "ENIT /Tunis": Tutorial on *"Data Mining et méthodes neuronales"* (12h): Y. Lechevallier.

9.2.2. PhD Thesis

Ph.Ds. defended in 2004 :

1. **Aicha El Golli** (starting date: end of 2001), "Cartes topologiques et modèles statistiques : application à la classification de données symboliques", University of Paris IX Dauphine (directors : E. Diday and Y. Lechevallier). June 1., 2004.

Ph.D. in progress :

1. **Doru Tanasa**, (start: end of 2001), “Trace et analyse de l’usage pour l’aide à la reconception d’un site Web”(Trace and Usage Analysis for assisting Web site re-design), University of Nice-Sophia Antipolis (director: B. Trousse).
2. **Sergiu Chelcea**, (start: end of 2002), “Classification de profils utilisateurs d’un site Web”(Classification of Web site users), Université de Nice-Sophia Antipolis (directors: J. Lemaire and B. Trousse with the support of P. Bertrand on 2-3 AHC).
3. **Hicham 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).
4. **Abdourahmane Balde**, (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).
5. **Nicolas Delannay**, (start: October 2003), “Méthodes neuronales pour les données structurées”, Université Catholique de Louvain-la-Neuve, Belgium (director: Michel Verleysen). F. Rossi is a member of the thesis committee.

AxIS researchers were members of Ph.D. committees in 2004 :

- Gename Youness “Une méthodologie pour la Comparaison de Partitions” in June at CNAM of Parisé : Y. Lechevallier
- Allou Badara Same “Modèles de mélange et classification de données acoustiques en temps réel” in December at the Technology University of Compiègne : Y. Lechevallier
- Gaëlle Legrand “Approche méthodologique de sélection et construction de variables pour l’amélioration du processus d’extraction des connaissances à partir de grandes bases de données” in december at University Lumière Lyon 2 : Y. Lechevallier
- Aicha El Golli “Cartes topologiques et modèles statistiques : application à la classification de données symboliques”, in June at the Dauphine University of Paris: Y. Lechevallier and F. Rossi.

A-M Vercoustre was involved as a Ph.D reviewer in the area of Content Management Systems (name confidential), for the School of Information Technology, University of Sydney, August 2004.

9.2.3. Internships

1. **C. Garboni** (University of Nice Sophia Antipolis (UNSA-DEA), January 13 to July 31, Inria Sophia Antipolis) [35].
2. **L. Baubois** University of Paris XIII, (University of Paris IX Dauphine, UFR Informatique de Gestion, Diplome d’Ingénieur, “Outils de visualisation des Cartes de Kohonen” [53].

9.2.4. Vulgarization

F. Rossi is responsible of the artificial intelligence vulgarization rubric of the GNU/Linux Magazine (in French, http://www.linuxmag-france.org/). The magazine published four Rossi's articles on various artificial intelligence topics in 2004 ([61][62][59][63]).

A.-M Vercoustre wrote an article for Tribunix (AFUU online magazine), Special issue on XML, October 2004, ([65]).

9.3. Participation to Workshops, Conferences, Seminars, Invitations

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

- INFORSID 2004 (Biarritz, May 24-26): B.Trousse http://inforsid2004.univ-pau.fr/
- Participation in the presentation of Cybercars, in the context of the CyberMOVE project (Antibes, June 3-13): B.Trousse, S. Schelcea
- EGC-FDC 2004 (Clermont-Ferrand, June 20-23): F. Masseglia, B. Trousse
- "Semaine du Document Numérique", in conjunction with the 7th international workshop on Numeric Doucment (CIDE), 21-25 juin, La Rochelle: A-M Vercoustre.
- Invited seminar by E. Guichard related to « Fonctionnement des moteurs de recherche » (in particular google), the cultural challenges of the search engines, Bibliothèque Publique d Information (musée Georges Pompidou), Paris, 29 march.
- Participation at " Réseaux Sociaux de l'Internet" thematic day organised by the institute of complexity sciences (ISCP) at ESPCI, Paris, 2 june: E. Guichard. http://www.liafa.jussieu.fr/~latapy/RSI.

F. Rossi was invited professor at the Université Catholique de Louvain-la-Neuve (Belgium) for one week in November

10. Bibliography

Major publications by the team in recent years

[1] H. H. BOCK, E. DIDAY (editors). *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*, Springer-Verlag, Studies in Classification, Data Analysis and Knowledge Organisation, 1999.

[2] P. BERTRAND, M. JANOWITZ. *The k-weak Hierarchies: An Extension of the Weak Hierarchical Clustering Structure*, in "Discrete Applied Maths", North-Holland, 1999.

[3] M. CHAVENT. *A monothetic clustering method*, in "Pattern Recognition Letters", 1999, p. 989-996.

[4] M. CSERNEL. *On the complexity of computation with symbolic objects using domain knowlege*, in "New Advances in Data Science and Classification", Springer-Verlag, 1998, p. 85-90.

[5] T. DESPEYROUX, B. TROUSSE. *Web sites and Semantics*, in "HYPERTEXT'01, the twelfth ACM Conference on Hypertext and Hypermedia, Aarhus, Danemark", août 2001, p. 239-240.

[6] M. JACZYNSKI. *Modèle et plate-forme à objets pour l'indexation des cas par situation comportementale: application à l'assistance à la navigation sur le Web*, Ph. D. Thesis, Université de Nice Sophia-Antipolis, Sophia-Antipolis, December 1998.

[7] F. MASSEGLIA. *Algorithmes et Applications Pour l'Extraction de Motifs Se'quentiels Dans le Domaine de la Fouille de Donne'es : de l'Incre'mental au Temps Re'el*, Ph. D. Thesis, Universite de Versailles St-Quentin en Yvelines, France, January 2002.

[8] B. TROUSSE. *Vers des outils d'aide à la conception coopérative: "Design Groupware"*, in "Connaissances et savoir-faire en entreprise - Intégration et capitalisation", J.-M. FOUET (editor)., chap. 17, Hermes, Paris, 1997, p. 317-341.

[9] B. TROUSSE. *Viewpoint Management for Cooperative Design*, in "Proceedings of the IEEE Computationnal Engineering in Systems Applications (CESA'98)", M. K. P. BORNE, A. E. KAMEL (editors)., UCIS - Ecole Centrale de Lille - CD-Rom, april 1998.

Books and Monographs

[10] P. GANÇARSKI, B. TROUSSE (editors). *Premier atelier sur la "Fouille de données complexes dans un processus d'extraction des connaissances", EGC'04*, 20 Janvier 2004, http://www-sop.inria.fr/axis/fdc-egc04/.

[11] 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.

Doctoral dissertations and Habilitation theses

[12] A. E. GOLLI. *Extraction de données symboliques et cartes topologiques: Application aux données ayant une structure complexe*, Thèse de doctorat, Université Paris-IX Dauphine, France, 2004.

Articles in referred journals and book chapters

[13] A. EL GOLLI, B. CONAN-Y GUEZ, F. ROSSI. *Self Organizing Map and Symbolic Data*, in "Journal of Symbolic Data Analysis", vol. 2, n° 1, November 2004.

[14] E. GUICHARD. *Mesures de l'internet*, E. GUICHARD (editor)., chap. L'internet, une technique intellectuelle, Les Canadiens en Europe, 2004, p. 19-49.

[15] F. MASSEGLIA, P. PONCELET, M. TEISSEIRE. *Extraction de motifs séquentiels : problèmes et méthodes*, in "Ingénierie des Systèmes d'Information (ISI)", vol. 9, n° 3/4, 2004, p. 183-210.

[16] F. MASSEGLIA, D. TANASA, B. TROUSSE. *Diviser pour découvrir. Une méthode d'analyse du comportement de tous les utilisateurs d'un site Web*, in "RSTI - Ingénierie des systèmes d'information (ISI)", vol. 9, n° 1, 2004, p. 61-83.

[17] D. TANASA, B. TROUSSE, F. MASSEGLIA. *Mesures de l'internet*, Sous la direction d'Eric Guichard, chap. Fouille de données appliquée aux logs web : état de l'art sur le Web Usage Mining, Les Canadiens en Europe, 2004, p. 126-143.

[18] D. TANASA, B. TROUSSE, F. MASSEGLIA. *Classer pour Découvrir : une nouvelle méthode d'analyse du comportement de tous les utilisateurs d'un site Web*, in "Revue des Nouvelles Technologies de l'Information (RNTI), numéro spécial EGC'2004", vol. 2, January 2004, p. 549–560, http://www.cepadues.com/.

[19] D. TANASA, B. TROUSSE. *Advanced Data Preprocessing for Intersites Web Usage Mining*, in "IEEE Intelligent Systems", vol. 19, n° 2, March-April 2004, p. 59–65, http://csdl.computer.org/comp/mags/ex/2004/02/x2toc.htm.

[20] D. TANASA, B. TROUSSE. *Data Preprocessing for WUM*, in "IEEE Potentials", vol. 23, n° 3, 2004, p. 22–25, http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?puNumber=45.

[21] B. TROUSSE, S. CHELCEA, G. GALLAIS. *Faciliter les déplacements par des recommandations personnalisées à la recherche d'information*, in "Revue Génie Logiciel, rubrique Systèmes d'informations et transports", n° 70, September 2004, p. 48 - 57.

Publications in Conferences and Workshops

[22] A. BALDÉ. *Extraction of Metatada on the Prototypes Resulting from Objects Clustering*, in "International workshop on symbolic data analysis, Paris, France", University PARIS-IX Dauphine., May 2004.

[23] A. BALDÉ, Y. LECHEVALLIER, M.-A. AUFAURE. *Extraction de métadonnées sur les prototypes issus de la classification d'objets*, in "11èmes Rencontre de la Société Francophone de Classification, Bordeaux", SFC'2004, septembre 2004, p. 95-98, http://www-sop.inria.fr/axis/papers/04sfc/balde-sfc2004.pdf.

[24] S. CHELCEA, P. BERTRAND, B. TROUSSE. *A New Agglomerative 2-3 Hierarchical Clustering Algorithm*, in "Innovations in Classification, Data Science, and Information Systems. Proc. 27th Annual GfKl Conference, University of Cottbus, March 12 - 14, 2003, Heidelberg-Berlin, Germany", D. BAIER, K.-D. WERNECKE (editors)., Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, ISBN 3-540-23221-4, 2004, p. 3-10.

[25] 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.

[26] S. CHELCEA, G. GALLAIS, B. TROUSSE. *Recommandations personnalisées pour la recherche d'information facilitant les déplacements*, in "Premières Journées Francophones : Mobilité et Ubiquité 2004, ESSI, Nice, Sophia-Antipolis, France", Cepadues - ISBN : 2-85428-653-7 / ACM Digital Library - ISBN : 1-58113-915-2, 1-3 Juin 2004, p. 143 - 150.

[27] S. CHELCEA, B. TROUSSE. *Application of the 2-3 Agglomerative Hierarchical Classification on Web usage data*, in "Proceedings of SYNASC 2004, 6th International Workshop on Symbolic and Numeric Algorithms for Scientific Computing, Timisoara, Romania", P. D., N. V., Z. D., J. T. (editors)., Mirton Publisher, ISBN 973-661-441-7, 26-30 September 2004, p. 107-118.

[28] B. CONAN-GUEZ, F. ROSSI. *Phoneme Discrimination with Functional Multilayer Perceptron*, in "Classification, Clustering, and Data Mining Applications (Proceedings of IFCS 2004), Chicago, Illinois", D. BANKS, L. HOUSE, F. R. MCMORRIS, P. ARABIE, W. GAUL (editors)., Springer, IFCS, July 2004, p. 157–165.

[29] M. CSERNEL, F. A. T. DECARVALHO, Y. LECHEVALLIER. *Partitioning of Constrained Symbolic Data*, in "9th Conference of the International Federation of Classification Societies, Chicago, USA", IFCS2004, juillet 2004.

[30] F. A. T. DE CARVALHO, Y. LECHEVALLIER, R. M. C. SOUZA. *Dynamic cluster algorithm based on adaptive Lr distances for quantitative data*, in "Proceedings of the 9th Conference of the International Federation of Classification Societies, Chicago, USA", Springer-Verlag, juillet 2004, p. 33-42.

[31] N. DELANNAY, F. ROSSI, B. CONAN-GUEZ, M. VERLEYSEN. *Functional Radial Basis Function Network*, in "Proceedings of ESANN 2004, Bruges, Belgium", April 2004, p. 313–318.

[32] 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.

[33] A. EL GOLLI, B. CONAN-GUEZ, F. ROSSI. *A Self Organizing Map for dissimilarity data*, in "Classification, Clustering, and Data Mining Applications (Proceedings of IFCS 2004), Chicago, Illinois", D. BANKS, L. HOUSE, F. R. MCMORRIS, P. ARABIE, W. GAUL (editors)., Springer, IFCS, July 2004, p. 61–68.

[34] 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 SFC, Bordeaux, France", Septembre 2004, p. 181–184.

[35] C. GARBONI, F. MASSEGLIA. *Structure Mining - A Sequential Pattern based Approach*, in "6th International Workshop on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2004, Timisoara, Romania", 26-30 September 2004.

[36] A. E. GOLLI. *Symbolic Data and Self Organizing Map*, in "International workshop on symbolic data analysis, Paris, France", University PARIS-IX Dauphine, May 2004.

[37] A. E. GOLLI, Y. LECHEVALLIER. *Extraction de classes homogènes et données symboliques*, in "Atelier N°6: Fouille de données complexes dans un processus d'extraction de connaissances, Clermont-Ferrand, France", EGC'2004, Janvier 2004.

[38] G. HÉBRAIL, Y. LECHEVALLIER. *Building small scale models of multi-entity databases by clustering*, in "Proceedings of the 9th Conference of the International Federation of Classification Societies, Chicago, USA", Springer-Verlag, juillet 2004.

[39] Y. LECHEVALLIER, A. CIAMPI. *Clustering large and Multi-levels Data Sets*, in "International Conference on Statistics in Heath Sciences, Nantes, France", ICSHS2004, juin 2004.

[40] Y. LECHEVALLIER, R. VERDE. *Classification croisée d'un tableau de données symboliques*, in "11èmes Rencontres de la Socie'te' Francophone de Classification (SFC), Bordeaux, France", septembre 2004, p. 245-248.

[41] Y. LECHEVALLIER, R. VERDE. *Crossed Clustering method: An efficient Clustering Method for Web Usage Mining*, in "Complex Data Analysis, Pekin, Chine", CDA'2004, octobre 2004.

[42] 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 (à paraître), Paris, France", Decembre 2004.

[43] F. MASSEGLIA, D. TANASA, B. TROUSSE. *Web Usage Mining: Sequential Pattern Extraction with a Very Low Suppor*, 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.

[44] F. MASSEGLIA, M. TEISSEIRE, P. PONCELET. *Pre-Processing Time Constraints for Efficiently Mining Generalized Sequential Patterns*, in "11th International Symposium on Temporal Representation and Reasoning (TIME'04), Tatihou, France", 1-3 July 2004.

[45] J. PEHCEVSKI, J. THOM, A.-M. VERCOUSTRE. *Enhancing Content-And-Structure Information Retrieval using a Native XML Database*, in "Proc of The First Twente Data Management Workshop on XML Databases and Information Retrieval, Enschede, The Netherlands", TDM'04, June 2004, http://www-rocq.inria.fr/~vercoust/PAPERS/jovanp-TDM04.pdf.

[46] J. PEHCEVSKI, J. THOM, A.-M. VERCOUSTRE. *RMIT INEX experiments: XML Retrieval using Lucy/exist*, in "Proceedings of the 2nd Initiative on the Evaluation of XML Retrieval, Dagsthul, Germany", ERCIM Workshop Proceedings, January 2004.

[47] F. ROSSI, B. CONAN-GUEZ, A. EL GOLLI. *Clustering Functional Data with the SOM algorithm*, in "Proceedings of ESANN 2004, Bruges, Belgium", April 2004, p. 305–312.

[48] F. ROSSI, B. CONAN-GUEZ. *Functional Preprocessing for Multilayer Perceptrons*, in "Proceedings of ESANN 2004, Bruges, Belgium", April 2004, p. 319–324.

[49] R. M. C. SOUZA, F. A. T. DE CARVALHO, Y. LECHEVALLIER. *Dynamic cluster methods for interval data based on Mahalanobis distances*, in "Proceedings of the 9th Conference of the International Federation of Classification Societies, Chicago, USA", Springer-Verlag, juillet 2004, p. 351-360.

[50] D. TANASA, J. LOPEZ, B. TROUSSE. *Extracting Sequential Patterns for Gene Regulatory Expressions Profiles*, in "Knowledge Exploration in Life Science Informatics, KELSI 2004, Milan, Italy, November 2004. Proceedings", LNAI, vol. 3303, Springer-Verlag, 2004, p. 46–57.

Internal Reports

[51] F. ROSSI, B. CONAN-GUEZ. *Estimation consistante des paramètres d'un modèle semi-paramétrique pour des données fonctionnelles discrétisées aléatoirement*, (improved version of the LISE/CEREMADE report number 0334, 2003), Technical report, nº 5228, INRIA Rocquencourt, juin 2004, http://www.inria.fr/rrrt/rr-5228.html.

Miscellaneous

[52] S. AHEHEHINNOU, B. AKROUT, F. BEKKA, G. F. EID. *Représentation Graphique des Cartes de Kohonen*, Internship Report, ISPG, University of Paris-XIII, 2004.

[53] L. BAUBOIS. *Outils de visualisation des Cartes de Kohonen*, Rapport de stage, Université Paris-Dauphine, 2004.

[54] H. BEHJA, B. TROUSSE, A. MARZAK. J. B. ET B. CRE'MILLEUX (editor). *Etat de l'art sur l'utilisation des techniques Web Sémantique en ECDAtelier "scenarios d'extraction de connaissance à partir de bases de données" - (EGC'2004)*, January 2004.

[55] H. BEHJA, B. TROUSSE, A. MARZAK. *Etat de l'art sur l'utilisation des techniques web sémantique en ECD*, Atelier EGC 2004 - Scénarios d'extraction de connaissance à partir de bases de données - (resps : JF Boulicault et B. Crémilleux), 2004.

[56] C. GARBONI. *Amélioration des résultats d'un processus d'extraction de connaissances : les solutions à apporter en amont*, Master's thesis, UNSA - Université de Nice Sophia-Antipolis, July 2004.

[57] M. JURCA, B. TROUSSE, J.-B. ABERT, A. GUIRAL. *Spécifications du moteur de traces de mediapps.net*, INRIA Sophia Antipolis, Delivrable D1 du Projet NRTL EPIA, janvier 2004.

[58] M. JURCA, B. TROUSSE, A. E. GOLLI, Y. LECHEVALLIER. *Contexte expérimental et moteur de traces de Net.Portal*, INRIA Sophia Antipolis, Draft version, Delivrable D3 (version 1.4 Draft) du Projet NRTL EPIA, december 2004.

[59] J.-M. MARIN, F. ROSSI. *Découvrez les réseaux Bayésiens*, GNU/Linux Magazine France, volume 60, pages 56–65, Avril 2004.

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

[61] F. ROSSI. *Découvrez les algorithmes évolutionnaires*, GNU/Linux Magazine France, volume 57, pages 42–51, Janvier 2004.

[62] F. ROSSI. *L'ordinateur peut-il lire dans votre esprit ?*, GNU/Linux Magazine France, volume 58, pages 34–44, Février 2004.

[63] F. ROSSI. *La reconnaissance de gestes*, GNU/Linux Magazine France, volume 63, pages 58–69, Juillet/Août 2004.

[64] B. SENACH, B. TROUSSE, G. GALLAIS. *Evaluation des apports des NTIC. Définition du scénario générique guidant l'évaluation du service VIP*, INRIA Sophia Antipolis (AxIS-Ergomatics et Visa), Delivrable 5.1 du Projet Mobivip, version 1, December 2004.

[65] A.-M. VERCOUSTRE. *Accéder l'information dans une collection de documents XML : Requêtes et Recherche plein texteTribunix*, AFUU, October 2004, http://www-rocq.inria.fr/~vercoust/PAPERS/XMLSearchTribuneX.ps.

Bibliography in notes

[66] A. AAMODT, E. PLAZA. *Case-Based Reasoning : Foundational Issues, Methodological Variations, and System Approaches*, in "The European Journal oo Artificial Intelligence", vol. 7, nº 1, 1994, p. 39-59.

[67] M. E. FAYAD, D. C. SCHMIDT. *Object-Oriented Application Frameworks*, in "Communication of the ACM", vol. 40, nº 10, 1997, p. 32-38.

[68] G. GOVAERT. *Algorithme de classification d'un tableau de contingence*, in "In Proc. of First International Symposium on data Analysis and Informatics, Versailles", INRIA, 1977, p. 487-500.

[69] G. GOVAERT, M. NADIF. *Clustering with block mixture models.*, in "Pattern Recognition", vol. 36, nº 2, 2003, p. 463-473.

[70] M. JACZYNSKI, B. TROUSSE. *Fuzzy Logic for the Retrieval Step of a Case-Based Reasoner*, in "Second European Workshop on Case-Based Reasoning (EWCBR'94), Chantilly", 1994, p. 313-320.

[71] R. JOHNSON, B. FOOTE. *Designing Reusable Classes*, in "Journal of Object-oriented programming", vol. 1, nº 2, 1988, p. 22–35.

[72] Y. KAGAMI, T. FURUICHI. *Investigation of differentially expressed genes during the development of mouse cerebellum*, in "Gene Expression Patterns", vol. 1, nº 1, August 2001, p. 39–59.

[73] J. KOLODNER. *Case-Based Reasoning*, Morgan Kaufmann Publishers, 1993.

[74] J. A. KONSTANT, 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, nº 3, 1997, p. 77-87.

[75] A. NAPOLI, A. MILLE, M. JACZYNSKI, B. TROUSSE, ALII. *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.

[76] P. RESNICK, H. R. VARIAN. *Recommender systems*, in "Communications of the ACM", vol. 40, nº 3, 1997, p. 56-58.

[77] H. SCHMID. *Probabilistic Part-of-Speech Tagging Using Decision Trees*, in "International Conference on New Methods in Language Processing, Manchester, UK", unknown, 1994.

[78] 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.

[79] H.-M. E. U. D. T.W. YAN. *From User Access Patterns to Dynamic Hypertext Linking*, in "Computer Network and ISDN systems", (proceedings of WWW'5)., vol. 28, May 1996, p. 1007-1014.

[80] S. WESS, K. ALTHOFF, G. DERWAND. *Using K-d Trees to Improve the Retrieval Step in Case-Based*

Reasoning, in "Lecture Notes in Artificial Intelligence, Topics in Case-Based Reasoning", S. WESS, K. ALTHOFF, M. M. RICHTER (editors)., Springer-Verlag, 1994, p. 167-181.

[81] 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.