INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team AXIS

User-Centered Design, Improvement and Analysis of Information Systems

Sophia Antipolis

THEME 3A

Table of contents

1. Team

Head of Project-Team

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

Vice-head of Project-Team

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

Administrative Assistants

Stéphanie Aubin [TR Inria, since April 1, INRIA Rocquencourt]
Christiane Demars [AI Intia, partial time,until March 31, INRIA Rocquencourt]
Sophie Honnorat [AI Inria, partial time,INRIA Sophia Antipolis]

Staff members

Thierry Despeyroux [Research Scientist(CR1), INRIA Rocquencourt]
Florent Masseglia [Research Scientist(CR2), INRIA Sophia Antipolis]

Research Scientists (secondment)

Marie-Aude Aufaure [IUT Lyon, May 1 to August 31, INRIA Rocquencourt]
Eric Guichard [Education Nationale, INRIA Sophia Antipolis]

Research Scientists (partners)

Mireille Arnoux [Professor Assistant, University of Bretagne Occidentale, INRIA Sophia Antipolis]
Patrice Bertrand [Professor Assistant, ENST Bretagne, INRIA Rocquencourt]
Marc Csernel [Professor Assistant, University of Paris IX Dauphine, INRIA Rocquencourt]
Fabrice Rossi [Professor Assistant, University of Paris IX Dauphine, since October 1st, INRIA Rocquencourt]

Project Technical Staff

Mihai Jurca [EPIA project since Nov. 1, INRIA Sophia Antipolis]

Junior Technical Staff

Sébastien Simard [until August 31, INRIA Sophia Antipolis]

Post-doctoral Fellows

Brieuc Conan-Guez [Student Assistant, University of Paris IX Dauphine, INRIA Rocquencourt]
Mohamed Semi Gaieb [Until February 28, INRIA Sophia Antipolis]

Doctoral Students

Abdourahamane Balde [Univ.of Paris IX Dauphine, from October 1st, INRIA Rocquencourt]
Hicham Behja [France-Morocco Cooperation (STIC-GL network), Univ.Hassan II Ben M'Sik, Casablanca, Morocco, INRIA Sophia Antipolis]
Sergiu Chelcea [University of Nice Sophia Antipolis (UNSA-STIC), INRIA Sophia Antipolis]
Brieuc Conan-Guez [Student Assistant, University of Paris IX Dauphine, INRIA Rocquencourt]
Aicha El Golli [University of Paris IX Dauphine, INRIA Rocquencourt]
Doru Tanasa [University of Nice Sophia Antipolis (UNSA-STIC), INRIA Sophia Antipolis]

Visiting Scientists

Rocio Alaiz [University of Leon, Spain, July, INRIA Rocquencourt]
Francesco de Carvalho [Federal University of Pernambuco, Brazil, March-April and September, INRIA Rocquencourt]
Antonio Ciampi [Mc Gill Univerity, QC, Canada, July, INRIA Rocquencourt]
Manuel Castejon Limas [University of Leon, Spain, July, INRIA Rocquencourt]
Bel Mufti Ghazi [Ecole Supérieure des Sciences Economiques et Commerciales, Tunis, Tunisia, October, INRIA Rocquencourt]
Abdelaziz Marzark [University of casablanca, september, INRIA Sophia-Antipolis]
Rosanna Verde [Professor, University of Napoli, Italy, March, July and September, INRIA Rocquencourt]

Student Interns

Abdourahamane Balde [Univ.of Paris IX Dauphine, April 6 to Sept. 30, INRIA Rocquencourt]
Luc Baubois [Univ.of Paris IX Dauphine, May 1 to July 31, INRIA Rocquencourt]
Fabien Benoit [ESSI (UNSA), until April 24, INRIA Sophia Antipolis]
Gérome Bernon [Télécoms 3, ISPG, until February 28, INRIA Rocquencourt]
Luc Baubois [Univ.of Paris IX Dauphine, May 1 to July 31, INRIA Rocquencourt]
Karine Deletre [DESS ErgoNTIC (UNSA), until february, INRIA Sophia Antipolis]
Olivier De Maeyer [DESS ErgoNTIC (UNSA), until february, INRIA Sophia Antipolis]
Olivier Famin [Télécoms3, ISPG, until February 28, INRIA Rocquencourt]
Patrick Guirchowski [DESS ErgoNTIC (UNSA), June 1 to Sept. 9, INRIA Sophia Antipolis]
Sigrid Famin [Télécoms3, ISPG, until February 28, INRIA Rocquencourt]
Mihai Jurca [Western University of Timisoara, Romania, until Oct. 31, INRIA Sophia Antipolis]

2. Overall Objectives

2.1. Objectives

Key words: *information system*, *design management*, *user-centered design*, *evaluation process*, *usage mining*, *knowledge management*, *experience management*, *knowledge extraction from Data*, *KDD*, *Web*, *utilisability*, *personalization*, *cooperative work artificial intelligence*, *statistics*, *clustering case based reasoning*, *adaptive service*, *adaptive interface*, *recommender system*, *semantic Web*, *semantic checking*.

The AxIS project-team was created on July 1st, 2003 from the ex AID, ADOPT and CROAP teams. AxIS is a pluri-disciplinary team (Artificial Intelligence, Data Mining, Data Analysis, Software Engineering). It aims at conceiving methods and tools, directed by usage, for assistance with the design and the analysis of knowledge and/or information systems (IS). Although in the short run the project is directed mainly towards the sites or Web services, we place ourselves in a global point of view of design and evaluation of adaptive information systems based on the W3C standards. The word « adaptive » represents both the ability to adjust to the user (personalization), and the ability to learn from usage analysis.

More precisely, the objective of the project is:

- to privilege and anticipate, starting from the design, the problems involved in the evolution of IS contents (architecture and documents) and those related to usage, as well as its evolution.
- to help IS designers (e.g. editors, Web-Masters or administrators) to better take into account the end-user.

Axis proposes to work according to two central points of view which are that of the artefact (designer, editor or webmaster of the IS) and that of usage. More generally we aim at facilitating the management of various points of view. In order to do so, we must initially understand and formalize the concept of multi-views information system. The implemented techniques are at the junction of different and complementary disciplines such as artificial intelligence (AI), information retrieval from data (KDD was introduced by Piatetsky-Shapiro in 1989 during one of the IJCAI'89 conference workshops), and software engineering. Our research program (cf. fig. 1 revolves around the static and dynamic aspects of IS and two transverse and federator topics: a) improvement of IS by confronting static and dynamic aspects, and b) capitalization of the resulting knowledge.

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

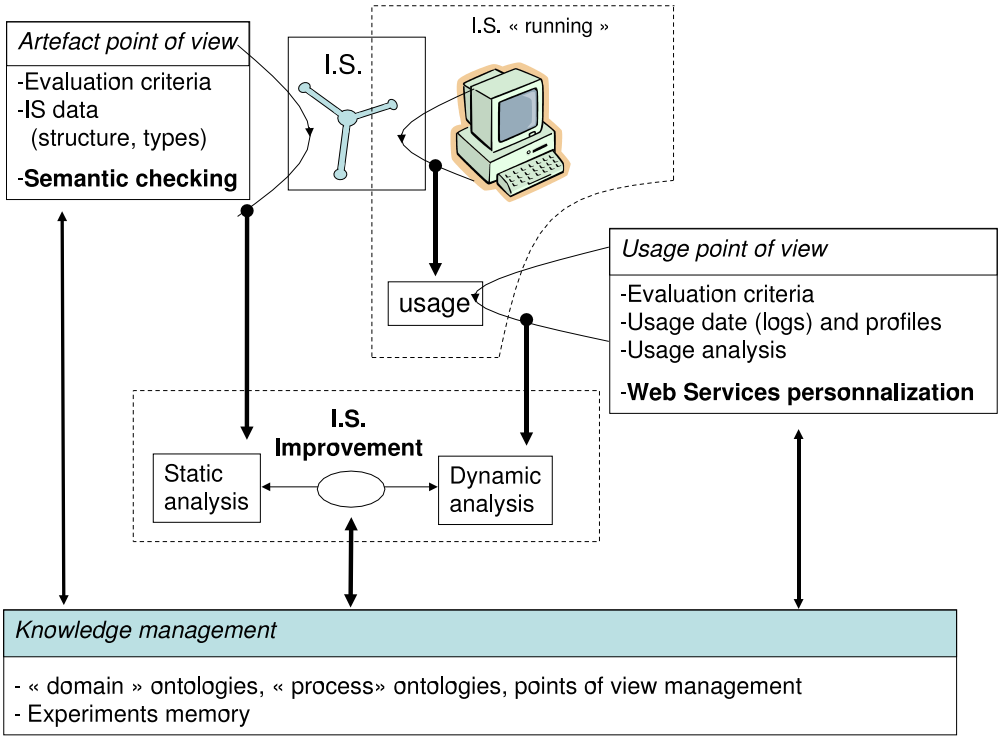

Figure 1. Global View of Research Topics

3. Scientific Foundations

3.1. Semantics and Design of Hypertext Information Systems

Key words: *Web semantics*, *static aspects of information systems*, *semantic checking*, *type checking*, *semi-structured documents*, *specification*, *maintenance*, *evolution*, *Web sites*, *formal semantics*, *software engineering*.

Designing and maintaining an hypertext information system, such as a Web site, is 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, using a multi-disciplinary approach (Software Engineering and Artificial Intelligence)

There is a strong relation between structured documents (such as Web sites) and a program; on the other hand the Web is a good candidate to experiment some of the ideas which have been developed in the software engineering world.

Until now most of the efforts developed within the Web domain were related to the presentation (HTML, CSS, XSL) or to the documents structure (XML) but not to the formal semantics of Web sites. We must anyway mention studies leaded by the W3C consortium on Web Semantic (XML, RDF, RDF-schema) and ontologies.

Our approach is different, as we want to explore the analogy between Web sites and programs to provide a formal semantics for Web sites.

The term « semantics » has at least two significations :

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

We will use this last definition when trying to give a formal semantics to Web sites.

We distinguish the static aspects of a site which are a set of global constraints (not only syntactic, but also semantic and context dependent) and must be verified, from the dynamic aspects. Dynamic aspects formalize the navigation inside a Web side (cf. the execution of a program). its use.

3.2. Usage Mining: Applying KDD to Usage Data

Key words: *usage mining*, *web usage mining*, *data warehouse*, *data mining*, *sequential patterns*, *user behaviour*.

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:

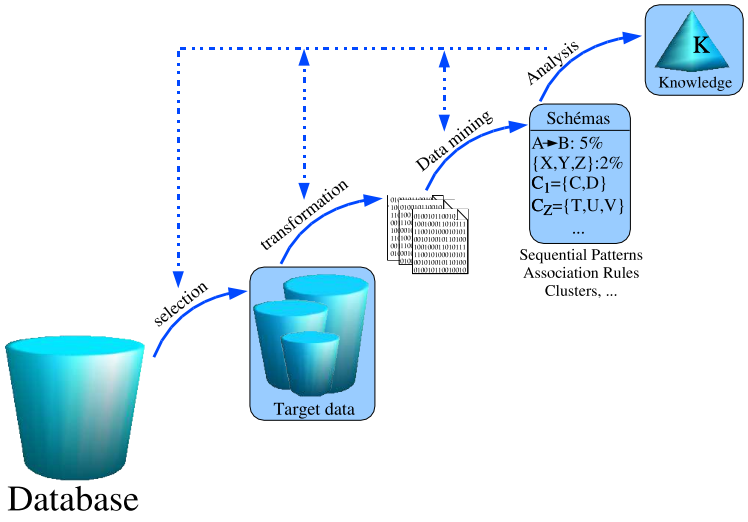

Figure 2. Steps of the KDD Process

3.2.1. a) 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 formated file (XML) or a private format.

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

3.2.2. b) 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 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. c) Discovering sequential patterns

Knowing the user allows to discover sequential patterns (which are inter itemsets association rules sorted by timestamps).The results obtained by the classical extraction algorithms are not accurate enough for a

detailed analysis of the users behaviour over the time. The log analysis can give more than the current research work does. We plan to work on the quality of the results in different ways: by taking into account the results timestamps, by clustering the results (to target a precise part of the users) or by considering a textual information (to be filtered before analyzing) or by taking into account the structure of the site while extracting the patterns.

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

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

Clustering methods reduce the volume of data in data warehouses, preserving the possibility to perform needed analysis. 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 they are decompressed. Our work differs in the sense that our compression technique has a semantic basis.

Clustering, in fact, leads to a classification, i.e. the identification of homogeneous and distinct subgroups in data [11] [15], where the definition of 'homogeneous' and 'distinct' 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

In spite of the great wealth of clustering algorithms, 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 item is in a unique cluster and ending with the trivial clustering in which all items are in the same cluster. A divisive method starts with all items in a single cluster and performs splitting until a stopping criterion is met (usually, until we obtain a partition of singleton clusters). Partitioning methods aim at obtaining a single partition of the input data into a fixed number of clusters. These methods identify the partition that optimizes (usually locally) an adequacy criterion. To improve the cluster's quality, the algorithm is run multiple times with different starting points, and the best configuration obtained from all the runs is used as the output clustering.

3.2.5. e) 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 (Mihai)

Key words: *recommender system*, *personalization*, *collaborative filtering*, *user support Web*, *hypermedia*, *KDD*, *CBR*.

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 [59][58][55].

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.

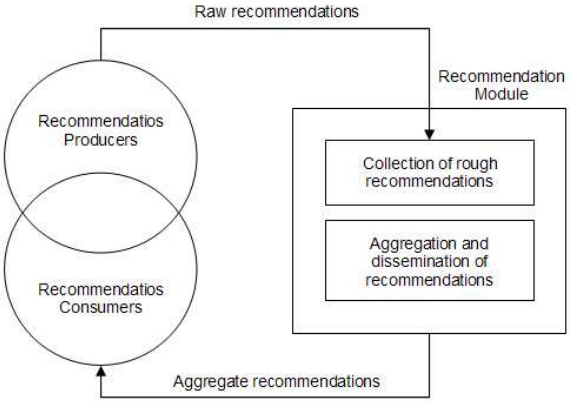

Figure 3. Architecture of a Recommender System

A major challenge in the field of recommender systems design is the following:

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

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) [58]

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 other 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 examples in the field of navigation assistance on the Web: the FootPrints system and the system of Yan et al.

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 study all three problems above by exploring the possibility of using CBR techniques and more generally the KDD technique.

We study the class of recommender systems, based on the re-use of a user group past experiences, using case based reasoning techniques (CBR). In this class of systems, we focus on the following two types of 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

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

Glossary

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 again be re-used.

Case-Based Reasoning [57] usually breaks up into four principal phases [44][54]:

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 very frequently related to: the definition and the representation of a case, the organization of the database containing the cases, the various used indexing methods and the definition of « good » similarities measurements for the case search, the link research-adaptation link of case (the best case being the most easily adaptable case), the definition of an adaptation strategy starting from the found case(s), the training of new indices, etc.

We continue the evaluation of our results in CBR, in particular of our indexing model by behavioural situation, of our object-oriented framework CBR*TOOLS nd toolbox BROADWAY*TOOLS. Moreover, more particularly 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

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

The project explores any applicative field on the design, the evaluation and the improvement of a big size hypermedia information system, for which the taking in account of the end-user is of primary importance. In the short run, we will focus on web sites (internet, intranet), or parts of web sites, having 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, generating many maintenance problems, eg.:
   - A site 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 site 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,
- Adaptative recommender systems for supporting information retrieval, Collaborative search of Information on the internet.

Ultimately, it should be noted that other fields (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

The AxIS software are mostly designed using the Rational Rose environment and developed with the Java programming language. These software are described at the following URL http://www-sop.inria.fr/axis/software.html

5.2. CLF - “Computer Language Factory”

Key words: *language specification*, *semantics*, *Centaur*, *Prolog*.

Participant: Thierry Despeyroux [correspondant].

The Computer Language Factory (CLF) is a set of tools and formalisms created to ease the creation of computer languages. It has been developed in the ex Croap project at Inria.

Parts of CLF were directly adapted to Prolog. The result is an extension to the traditional Prolog DCG (Definite Clause Grammars). Compared to DCG our parser generator has several advantages : it eases the writing by allowing left recursion, it optimizes the compilation of grammar rules to Prolog taking advantage of clause indexing, the structure that is returned by the parser is rich enough to allow back references in error messages (for example line numbers) when it is used as object of a compiler.

This extension is used to easily produce a Prolog parser for the XML language, and also to produce a parser for our definition language presented in section 6.4.1.

5.3. Clustering ToolBox

Key words: *clustering*, *clusters*, *visualization*, *Kohonen maps*.

Participants: Yves Lechevallier [correspondant], Marc Csernel, Mihai Jurca, Brigitte Trousse.

The clustering toolbox, written in C++, allows to group all tools and classification methods developed within the team and uses the library developed by M. Csernel. Such a library proposes in such a way a common data interface to every algorithm. This toolbox provides to developers an easy way of integration for every classification method, and, an easy way to test and compare other methods.

In 2003, we followed the development of an Web interface of this clustering toolbox for our short term own needs.

The aim of this Clustering Tools Online interface is to give the possibility to our team members and to other possible Internet users to use these classification methods to processing their data. The Web interface is developed under C++, run on our Apache internal Web server and supports execution of following classification methods:

- SCluster (AxIS Rocquencourt)
- Div (AxIS Rocquencourt)
- CDis (collaboration between by AxIS Rocquencourt team and Recife University, Brazil)
- CCClust (collaboration between by AxIS Rocquencourt team and Recife University, Brazil)

All these methods are written in C++ and use the same input and output format but take different parameters. The format of the input/output is the format recognized by the SOM parser (used by SCluster, Div, CDis and CCClust methods).

Classification Tools Online has proved to be a tool that makes easy the teamwork.

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

Key words: *case-based reasoning*, *object-oriented framework*, *UML*, *design pattern*, *re-use*.

Participants: Semi Gaieb, Mihai Jurca, Sébastien Siémard, Brigitte Trousse [correspondante].

CBR*TOOLS is an object framework developed in the team since 97 with the purpose to facilitate the development of applications requiring case-based reasoning technologies.

CBR*TOOLS [7] [51] is an object framework (or « object-oriented framework » [52][48]) in Case Base Reasoning, which offers a set of abstract classes which modelize the main concepts necessary to develop an application integrating case base reasoning technics: 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 [60], prototypes indexing [49], neuronal approach based indexing, standards similarities measurements). CBR*TOOLS currently contains more than 240 classes divided in two main categories: the package core for basic functionality and the package time for the specific management of

the behavioural 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.

Particularly, CBR*TOOLS aims application fields where the re-use of cases indexed by behavioural situations is required.

CBR*tools was installed at France Telecom (R&D) at Lannion in 1998 and 2000 within BROADWAY-WEB and EDUCAID (FT-CTI) framework and was used within XRCE-INRIA contract (98). A documentation on the Web is accessible to the following address. http://www-sop.inria.fr/axis/cbrtools/manual/.

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).

During 2003, CBR*Tools has been migrated to the CVS server of Inria Sophia Antipolis (cvs-sop) managed by the SEMIR service in order to prepare future distributions. More several scripts make easier and automatically the process of dvelopment-build-distribution, the versioning and the software installation. This work was mainly done by S. Simard.

5.5. Broadway*Tools - Generator of Adaptative Recommender Systems

Key words: *customized agent*, *re-use*, *adaptive service*, *personalization*, *recommender system*, *information retrieval*.

Participants: Semi Gaieb, Sébastien Siemard, Brigitte Trousse [correspondante].

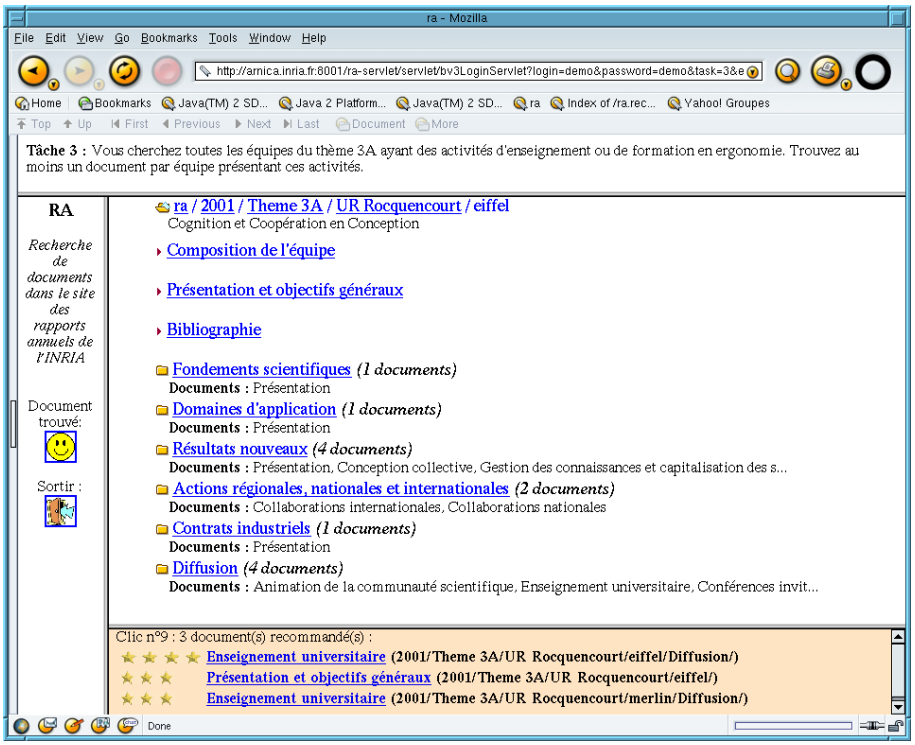

Figure 4. Example of the e-behaviour system developed with BROADWAY*TOOLS

BROADWAY*TOOLS is a toolbox used to facilitate the creation of adaptive recommendation Web-systems for the assistance in the search of information on the Web or in a Internet/intranet site. This toolbox offers different servers of which a server of calculation of recommendations based on our behavioural indexing model for the observation of the user sessions and on the re-use of user groups old sessions. A recommender system created with BROADWAY*TOOLS observes navigations of various users and gather the evaluations and annotations of these users in order to draw up a list of relevant recommendations (Web documents, keywords, etc).

This toolbox was used for our navigation assistance systems, specially for the e-behaviour system (cf. 6.4.5). In 2003, we migrated it to the CVS server of Inria Sophia Antipolis (cvs-sop).

5.6. Broadway-Web - Personalized Supporting Web Browsing

Key words: *collaborative filtering*, *browsing support*, *user behaviour*, *recommender system*, *Web*, *user profiles*.

Participants: Sergiu Chelcea, Semi Gaieb, Mihai Jurca, Sébastien Simard, Brigitte Trousse [correspondante].

BROADWAY-WEB (called in the past BROADWAY (V1) - « BROwsing ADvisor reusing pathWAYs ») [50] is a browsing advisor reusing past navigation from a group of users.BROADWAY-WEB observes browsing of various users and gathers the evaluations and the annotations of these users to draw up a list of relevant documents.

BROADWAY-WEB (http://www-sop.inria.fr/axis/broadway/), the result of Mr. Jaczynski's thesis (1998), is a HTTP server used like proxy: it is a middleware between the browser and the Web and it intercepts all the requests for documents by HTTP protocol. BROADWAY-WEB is then able to observe various user navigations while recording specific information: addresses of the visited documents, a set of keywords resulting from the analysis of HTML pages, and the evaluations of the documents.

During the navigation, BROADWAY-WEB can display a set of documents which it advises according to the navigation current state, and allows users to evaluate or annotate the visited documents by using a toolbar which is dynamically inserted in the displayed HTML pages.

BROADWAY-WEB integrates the HTTP Jigsaw server of the W3C developed in Java and uses our CBR*TOOLS platform to implement the case base reasoning system, allowing the re-use of past navigations.

BROADWAY-WEB was installed at France Telecom (R&D) in Lannion in 98 and was presented during the 8th and 9th INRIA-Industry meetings (November 1998 and December 2001) and during the conference TTNL'02 in Sophia Antipolis (November 2002).

In 2003, BROADWAY-WEB has been migrated to the cvs-sop server of Inria Sophia Antipolis.

6. New Results

6.1. Data Transformation and Knowledge Representation

6.1.1. Extraction and Construction of Aggregated Data

Participants: Aicha El Golli, Yves Lechevallier.

Key words: *generalization*, *aggregated data*.

In the field of data analysis, a generalization method allows us to aggregate the data issued from a database in such a way that the underlying concepts clearly appear. Such a function can be considered as well as a user oriented tool, or a primary step to further analysis conducted upon these concepts. For a given set of individuals G of Ω, the whole population set, the generalization function is expected to create a good G representation through a multidimensional vector, summarizing all individual description (using real, binary, categorical ... values).

An already well known method consists in the association with each vector of a scalar value, summarizing the individuals dispersion. A second approach, proposed in the symbolic data analysis framework, allows to summarize an observation set through a symbolic description (interval, distribution...).

In 2003, we have developed a new generalization method [24]. This method builds a set of homogeneous classes, each of them being a symbolic object. The goal we try to reach is the extraction from a set of relational databases, of an assertion set providing a good description of the concept we try to analyze. This is done using the generalization operator provided by V. Stephan in her thesis.

If some concept descriptions provided according to this method are too heterogeneous, we provide a divisive classification method, in order to improve the quality of the description. The new description is provided as a disjunction of assertions.

This new approach has been introduced in DB2SO, a program used in the ASSO software. ASSO is a data analysis software built up with the help of the EEC. It enables to build symbolic objects from relational databases. In the first version of the ASSO software, DB2SO was able to group all observations belonging to the same concept, and then by the use of a generalization operator obtain a symbolic object call « assertion ». Some of these assertion have got « atypical description » trough their extension.

A new version provides improvement, by enabling us to represent each group, not only with a single assertion, but also, if necessary, with a disjunction of assertions. In such a case we obtain a more homogeneous description, with a better representation quality. The decomposition allowing such a representation is based on a divisive classification algorithm.

6.1.2. N. S. F.: Normal Symbolic Form

Participants: Marc Csernel, [F. A. T. De Carvalho], Yves Lechevallier.

The Normal Symbolic Form research topic conducted by Marc Csernel is a decomposition of symbolic description, slightly related to Codd's third normal form, which allows in nearly every occasion a polynomial computation time if rules are constrains the description. Without the NSF, in rules presence, the computation time would be exponential according to the number of rules.

The work about N.S.F mainly focused on application. We propose in [20] how to build clusters of symbolic objects in presence of rules constraining the description space, in a polynomial time. On the other hand the N.S.F is planned to be used in the ASSO project by the University of Porto team in order to compute within an acceptable amount of time the description potential of symbolic objects .

6.1.3. Meta-Data

Participants: Marie-Aude Aufaure, Abdourahamane Balde, Luc Baubois, Marc Csernel, Yves Lechevallier.

Another important research topic this year was Meta Data, and more specifically upon Meta Data related to symbolic data such as the one we treat in the ASSO project. This work just lies between research and application. The Meta Data concerning symbolic data may be in some special cases not only important (as for any data), but absolutely necessary if one wants to compare them according to specific criteria. This is the case if symbolic objects have been built using any aggregative process such as DB2SO. Then the meta data provided the number of initial objects which have been aggregated during the process, and tells it us whether comparisons between objects are meaningful.

First we started by defining a certain amount of metadata related to symbolic data, some of them in collaboration with the U.O.A. (University of Athens) team, namely Maria Vardaki and Babis Papageorgiou. Then we did proceed to the implementation of these Meta Data within the Asso Library, with a special attention dedicated to the parser modification. The XML representation of the Meta Data was achieved, including an XSL file providing an easily Human readable representation.

With the collaboration of Luc Baubois [38] the Symbolic Object Generator DB2SO was modified in order to generate both Data and Meta Data, according to the definition previously defined. Now in the Asso Project all data can be accompanied with MetaData. The trainee period was technically oriented, and we do not, until now, have gotten any report. But the program works quite correctly.

During his trainee period A. Balde [37] made a study about the Meta Data and XML. He pointed two kinds of Meta Data: 1) one specific related to the Dublin core : it provides to the different search engines a standardized way of getting the information related to the author, publication date, etc. and 2) more general, the RDF

schema as defined by the W3C enables to describe the relation between the different items composing a Web site.

6.1.4. Kohonen Maps Visualization

Participants: Gérome Bernon, Luc Baubois, Olivier Famin, Yves Lechevallier.

The « classical » representation associated with Kohonen Map is a set of U cell connected in a lattice. The lattice structure is free, but usually a grid is used as visualization structure.

During their trainee period, Gérome Bernon, Olivier Famin et Sigrid Herstain [40] (resp: Y. Lechevallier) proposed three new form for the cell visualization. A first version of a program have been written in JAVA. These descriptions have been constructed using a weight vector associated to each cell of the map.

6.2. Data Mining Methods

Participants: Patrice Bertrand, [Marie Chavent], Marc Csernel, [F. A. T. De Carvalho], Aicha El Golli, Yves Lechevallier, [Rosanna Verde].

Key words: *clustering*, *aggregated data*, *distances*, *intervals*, *Hausdorff distance*, *topological cards*, *symbolic data*, *stability*, *validation*, *symbolic data*.

This year we proposed in [21][19][25][11] clustering algorithms which partition a set of SO's into a predefined number of classes. The classes are suitably interpreted and represented by generalized class prototypes (again in the form of symbolic objects). The clustering criterion, based on a context-dependent proximity function or on a resemblance measure, is optimized by a symbolic batch algorithm (dynamic or k-means clustering) or by sequential exchange algorithms. Concerning partitioning methods, used an approach based on a transfer algorithm to partition a set of symbolic objects into clusters described by vectors of weight distributions. have presented a fuzzy k-means algorithm to cluster symbolic data described by different types of symbolic variables. Last year we have introduced a dynamic cluster algorithm for symbolic data considering context dependent proximity functions where the clusters prototypes are vectors of weight distributions and we presented an iterative relocation algorithm to partition a set of symbolic objects into classes in order to minimize the sum of the description potentials of the classes. A dynamic cluster algorithm to interval data where the prototype is defined by the optimization of a criterion was proposed this year. However, none of the methods to cluster symbolic data already presented uses adaptive distances [20].

The proposed Dynamic Clustering Method determines iteratively a series of partitions which improve at each step the underlying clustering criterion. The algorithm is based on:

- prototypes to represent the classes
- context-dependent proximity functions, to assign the elements (SO's) to the classes at each step.

The clustering criteria to be optimized is based on the sum of proximities between individuals and the prototype of the assigning clusters.

6.2.1. Partitioning Method : Clustering of Interval Data and of Set-Valued Variables

Participants: [Marie Chavent], [F. A. T. De Carvalho], Yves Lechevallier, [Rosanna Verde].

6.2.1.1. Clustering of Interval Data

In [21] we presented partitioning cluster methods for interval data using a dynamic cluster algorithm based on a Hausdorff distance. Two methods are considered: one with an adaptive distance and the other without. These methods provide a partition and a prototype to each cluster by optimizing an adequacy criterion that measures the fitting between the clusters and their representatives. The first method uses a non-adaptive measure based on a Hausdorff distance which is able to compare a couple of intervals. In the second method, an adaptive version of the Hausdorff distance is presented. This adaptive version at each iteration changes for each cluster according to its intra-class structure. The advantage of these adaptive distances is that the clustering algorithm is able to recognizes clusters of different shapes and sizes.

In Symbolic Data Analysis, a symbolic description of an object is a data vector whose columns are symbolic variables. These symbolic descriptions are organized in a Symbolic Data Table, which is more complex than the usual one because each cell may contain more than a single value. Moreover, weights can be associated to intervals and categories, structures such as taxonomies of values can be take into account and logical rules between symbolic descriptors can be considered.

6.2.1.2. Clustering of Set-valued Variables

In [21][25][11] we proposed a dynamical clustering algorithm in order to partition a set of *symbolic objects*, described by set-valued variables, in a predefined number *k* of classes.

In the proposed algorithm, each class is represented as a *modal symbolic object* characterized by the distributions associated to their descriptors. Suitable dissimilarity functions (*context dependent*) are considered in order to allocate, at each step, the objects to the classes. The procedure is iterated until the convergence to a stationary value of the optimized criterion is reached. According to nature of the complex data, the first phase of the proposed algorithm consists in choosing a suitable type of representation of the classes of objects. Similarly than in dynamical clustering method on individual data, we refer to suitable prototypes G_i of the clusters in order to represent them. In this paper we propose to performing prototypes by synthesizing the whole information of the SO's belonging to the different classes. Each prototype associated to a class can be modeling as a *modal SO* (Diday (1998)). The description of a modal SO is given by frequency (or probability) distributions or a weighting for the p variables.

The other main aspect of such algorithm concerns the choice of the proximity function in order to assign SO's to the classes. In particular we consider a suitable function $\partial(\cdot)$ in order to compare SO's with prototypes, so called *context dependent proximity functions* (De Carvalho et al. (1998)). The allocation function $\partial(\cdot)$ can be assimilated to an extension mapping (Bock and Diday (2000)) of prototypes expressed as modal SO's.

6.2.2. Partioning Method: Clustering Large Amount of Data with a Kohonen SOM-based approach

Participants: [Antonio Ciampi], Brieuc Conan-Guez, Aicha El Golli, Yves Lechevallier.

Our approach is based on two key ideas:

- 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 affect the mixing coefficients, and not the parameters of the distributions.

The Kohonen Self Organizing Map (**SOM**) introduced by Professor Kohonen [53] is an unsupervised neural network method which has both clustering and visualization properties. It can be considered as an algorithm that maps a high dimensional data space, $\mathbf{R}^p$, to lattice space which usually has a lower dimension, generally 2 and is called a Map. This projection enables a partition of the inputs into « similar » clusters while preserving their topology. Its most similar predecessors are the k-means algorithm and the dynamic clustering method (Diay 1989), which operate as a **SOM** without topology preservation and so without easy visualization.

In data analysis, new forms of complex data have to be considered, most notably structured data (data with an internal structure such as intervals data, distributions, functional data, etc) and semi-structured data (trees,

XML documents, SQL queries, etc.). of a dissimilarity for each type of data is necessary to apply the method and therefore to process complex data. In this context, classical data analysis based on calculating the center of gravity can not be used because inputs are not $\mathbf{R}^p$ vectors. In order to solve this problem, several methods can be considered according to the type of data (for example recoding techniques for symbolic data or projection operators for functional data (Ramsey & Silverman 1997). However, those methods are not fully general and an adaptation of every data analysis algorithm to the resulting data is necessary.

We propose in our work, [23], an adaptation of the **SOM** to dissimilarity data as an alternative solution. Indeed, Kohonen's **SOM** is based on the notion of center of gravity and unfortunately, this concept is not applicable to many kind of complex data, especially semi-structured data. Our goal is to modify the SOM algorithm to allow its implementation on dissimilarity measures rather than on raw data. With this alternative only the definition of a dissimilarity for each type of data is necessary to apply the method and therefore to process complex data.

6.2.3. Partitioning method: Symbolic Clustering Interpretation and Visualization

Participants: [Marie Chavent], [F. A. T. De Carvalho], Yves Lechevallier, [Rosanna Verde].

In the framework of Symbolic Data Analysis [17][14], the algorithms to cluster a set E of symbolic data (Verde, De Carvalho, Lechevallier, 2000) are based on different types of assignment functions and different kinds of prototypes which represent the classes. The classical interpretative aids are usually based on inertia criterion. In our approach we propose to generalize this criterion to symbolic data, where the barycenter is replaced by the prototype of the clusters. The several indexes measure the improvement of a partition in k clusters with respect to the global cluster E. This comparison allows to evaluate the contribution and the discrimination of the symbolic objects and variables to the partition. Further suitable measures are considered to value the homogeneity of the clusters of the partition. In (Celeux *et al*. 1989) the indexes proposed to describe a partition are based on the decomposition of the total inertia into within and between inertia. These indexes provide a valuable help for the interpretation of the clusters obtained by partition methods like « *Nuées Dynamiques* » method. In our approach, the barycenter of the clusters is replaced for a prototype and the *total inertia* is generalized by the homogeneity of the partition in one class and the *within inertia* by the homogeneity criterion of the description of each cluster; while the *between inertia* is defined as the difference between the homogeneity measure of the partition in one cluster and the partition in k clusters.

The **quality measures** [11] of a partition and of their clusters, hereafter proposed, can be interpreted as the gain between the null hypothesis « No structure = Partition into one cluster » and the solution carried out a classification algorithm into K clusters optimizing the fitting criterion between a partition P and the corresponding vector of prototypes of D. The proposed quality measures provide an interpretation of a partition of multi-valued data.

6.2.4. Partitioning Method: Cluster Stability

Participants: [G. Bel Mufti], Patrice Bertrand.

During the last years, we have proposed in collaboration with G. Bel Mufti, a method for measuring cluster stability when a few objects are removed from the partitioned data set (see Bel Mufti and Bertrand (2001)). Three stability indices have been defined in this approach. Two of them estimate the inherent stability respectively in the isolation and in the cohesion of the studied cluster. By combining these two stability measures, the third index estimates the overall stability of the cluster. Using Monte Carlo tests, levels of significance are computed in order to assess how likely the values taken by the stability indices are, under a null model that specifies the absence of cluster stability.

During this year, we have improved the definitions of these indices in order to normalize their values in the particular cases of independence and perfect stability. We have proved the existence of a linear dependency between the isolation and the cohesion indices that are computed on a same sample. We have also extended our validation approach to a partition.

6.2.5. Functional Data Analysis

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

Publications: [26][31][30][32]
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, B. Conan-Guez proposed the extension of MLPs (Multi-Layer Perceptrons) to functional inputs : the Functional Multi-Layer Perceptron (FMLP). We showed two important properties : this model is a universal approximator, and the parameter estimation is consistent when we know only a finite number of functions known on a finite number of evaluation points.

In 2003, we showed that the assumptions on the observed functions can be weakened compare to those used in previous results (we get rid of the compacity assumption). In the same time, we compared the proposed model to well-known functional tools on two different databases (computations were carried out on Cristal-Cluster):

- the first one is a spectrometric application from food industry. The goal was to estimate the fat proportion in meat samples. We first used these spectra as FMLP inputs. Results are in this case satisfactory. In the second time, we used the second derivative of spectra as inputs. Results are even better.
- We applied FMLPs to a phoneme recognition problem. The goal is to discriminate 5 different phonemes. Obtained results are again satisfactory.

Now, we work on a pre-processing stage based on the Partial Least Squares Regression technique. This approach allows to obtain a set of components which describes accurately and parsimoniously the studied functions. The result of this decomposition step is then submitted to an FMLP. We try to apply this approach to the two databases described above.

6.2.6. Extensions of Ascendant Hierarchical Clustering (AHC): the 2-3 AHC

Key words: *hierarchy*, *clustering*, *AHC*, *single link*, *complexity*.

Participants: Patrice Bertrand, Sergiu Chelcea, Mihai Jurca, Brigitte Trousse.

The main motivation of Chelcea's thesis (directed by Lemaire and Trousse with the support of Bertrand) concerns the use of clustering techniques for user profiling based on Web user sessions and for indexing cases [45] inside a Case-Based Reasoning framework (CBR*TOOLS) for Web-based recommender systems [7]. It is in this context that we have proposed [22] a new Agglomerative 2-3 Hierarchical Classification algorithm which improves the one proposed in 2002 [46] by reducing its complexity from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2 \log n)$. This complexity reduction is based on the integration of the refinement step into the merging step and also due to the simplification of the data structure manipulation.

Compared to the previous algorithm, our new algorithm has a principle similar to the one of the classical AHC algorithm based on the introduction of a intermediate merging step. This provides us, at the end of the mergings during the 2-3 AHC algorithm execution, a partitioning of the initial set by the maximal clusters, which is as well the case of the AHC algorithm.

We have designed an object-oriented model of the algorithm, which was implemented in Java, and we begin to test this algorithm as an indexing method within the CBR*TOOLS framework [7].

We carried out a series of tests on the Ruspini data set, on random generated data and in a CBR application for risk factor determination of a car, based on a data base[1] usually used in this context. The 2-3 AHC algorithm complexity has been experimentally validated on simulated data sets.

The comparative analysis on random simulated data sets between our 2-3 AHC algorithm and the classical AHC algorithm has lead to the following observations:

- Using the single link and compared to the classical AHC, we obtained on average 10.9% more clusters (with a standard deviation of 1.8) for the 2-3 AHC algorithm with the integrated refinement

[1] ftp://ftp.ics.uci.edu/pub/ml-repos/machine-learning-databases/autos

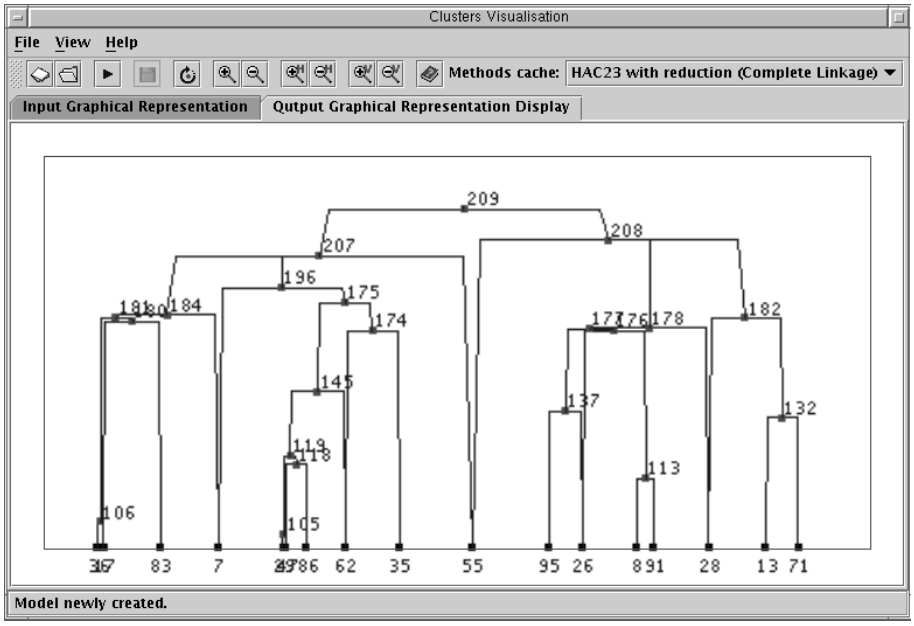

Figure 5. The clusters visualization application

step, and 45.6% more clusters (with a standard deviation of 0.8) for the 2-3 AHC algorithm without the refinement step. In the case of the complete link, these values are respectively 44.6% (1.2) against 47.5% (0.8) more clusters.

- Using the complete link and the *Stress* formula, we have compared the induced dissimilarity matrix with the initial one, and obtained a 23% gain.

In order to better visualize and compare the output of the different classification methods, an application for visualizing the hierarchical classification structures has been developed in Java (see Figure 5).

6.3. Viewpoint Management in KDD

Key words: *knowledge management*, *viewpoint*, *KDD*, *design pattern*, *object-oriented model*, *reusability*.

Participants: Hicham Behja, Abdelaziz Marzark, 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 andor methods and then offering more flexibility and adaptability.
In 2003, we started by a state of the art of the use of knowledge in the KDD process, both knowledge from the analysed domain and from the the analysts themselves. We focused on the existing works adopting a Semantic Web based approach, whether for the analysed domain (cf. MiningMart [56] or for the analyst domain (cf. Daemon [47] and Clementine [2]). Then we designed a first version of our object-oriented KDD model (design patterns, UML notation) which integrated the specification of analyst viewpoints inside such a process. We are now applying our model on analysing the use of a Web site, specially Inria Sophia Antipolis site (according to the « fiability » and « usability » viewpoints)and the use of a e-learning system (which is the case study of our network).

[2] http://www.spss.com/spssbi/clementine/

6.4. Web Mining and Web Applications

6.4.1. Semantics of Web Sites

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

Participants: Thierry Despeyroux, Brigitte Trousse.

The main goal of Semantic Web is to ease a computer based data mining, formalizing data that is most of the time 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.

In the past, we have evaluated the possibility to apply the Natural Semantics to the specification and the verification of the formal semantics of Web sites. This approach comes from the fact that we see an analogy between Web sites and programs. The use of Natural Semantics seems adequate but not well adapted to the world of Web sites : the rules are too detailed and difficult to write by non specialists.

Differences between Web sites and programs have also been pointed out :

- Web sites may be spread along a great number of files. This is also the case for programs, but these files are 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.

- The information is scattered, with a very frequent use of forward references. A forward reference describes an object (or a piece of information) that is used before it as been defined or declared. In programs, forward references exist but are most of the time 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.

After experiments conducted directly in Prolog, we have defined in 2003 a specification language to express global constraints in Web sites. A first version of this language has been implemented. Its parser is constructed as an extension of the XML parser done with CLF (see 5.6). The compiler is directly in Prolog and produces Prolog code.

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. Each file is the reflect of the activity of a research project/team. Even if a large part of the document is written in French, the structure and some parts of the document are formalized (parts concerning people and bibliography).

We tested whether each person cited as a participant was declared as a member of the team, and whether the dates in the bibliography of the year were correct. All URLs cited in the document were also tested. We also extracted hidden information from the document : how many papers were produced in cooperation with several teams.

Our system reported respectively 1372 and 1432 messages.

In the near future, we plan to extract indicators from the activity reports . These indicators are already produced manually and are used for the management purposes.

6.4.2. Ontology-Guided Image Retrieval

Key words: *ontology*, *image data*, *image mining*.

Participants: Marie-Aude Aufaure, Yves Lechevallier.

Image data are omnipresent for various applications. A considerable volume of multimedia data is produced and we need to develop tools to efficiently retrieve relevant information. Image mining [18][29][28] is a new and challenging research field and deals with making associations between images coming from large databases or internet and to present a summarized view. Image mining is a way to fill the semantic gap between visual properties and semantic content. We propose an architecture combining clustering and characterization rules which integrates visual and textual features. These features are separately processed and the characterization process is then performed on each previous calculated cluster. While clustering is used to reduce the search space, rules are performed to describe each cluster and to classify new images. A visual ontology is built according to the set of annotations. Ontology-based navigation is a user-friendly and powerful tool to retrieve relevant data. We just begin to experiment a part of this architecture on texture images, using Kohonen maps developed in the ASSO project. New experiments on more complex image databases, including semantic annotations and different visual features are considered. We will also study methods to extract relevant descriptive metadata and to build a visual ontology.

6.4.3. Pre-processing of Web Usage Data

Participants: Mireille Arnoux, Eric Guichard, Yves Lechevallier, Doru Tanasa, Brigitte Trousse.

The preprocessing step of the Web Usage Mining in a multi-sites context is one of the research topic of Tanasa's thesis. This step, known for its importance, is required before applying the data mining techniques for analyzing the data. We detail in [27] a general method of preprocessing the HTTP logs that we used within an experiment done in 2003 on the Web log files extracted from three of the Inria's Web servers during the first two weeks in January. One of the originality of our method consists in the multi-site feature that we considered. This feature is crucial for a better apprehension of the internaut's habits. We also propose, taking into account the data mining step, a relational schema for a Web usage datawarehouse [15].

In 2003, we developed Log2DB, a software tool for Web Usage Mining to store a processed and cleaned log file into a relational database. The aim is to easily distribute these logs to interested third parties. The format of the input log file for the Log2DB is a specific one. The input log file is the result of the HTTP log cleaning algorithms developed mainly in Perl within our project. Log2DB software tool can be used either in command line mode or through its graphical user interface.

The database format offers the advantage of reduced data redundancy and contains a set of computed statistics about sessions and navigations from the log file. Also is much easier to use it than the log file in text format.

6.4.4. Two New Hybrid Sequential Pattern Extraction Methods

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

Participants: Florent Masseglia, Doru Tanasa, Brigitte Trousse.

The goal of this work is to increase the relevance and the interestingness of patterns discovered by a Web Usage Mining process. Indeed, the sequential patterns extracted on web log files, unless they are found under constraints, often lack interest because of their obvious content. Our goal is to discover minority users behaviours having a coherence which we want to be aware of (like hacking activities on the Web site or a users activity limited to a specific part of the Web site). To solve this problem we proposed two new hybrid approaches based on neural clustering [16] and sequential pattern mining. Such approaches aim 1) at dividing the log sessions and create sublog corresponding to clusters created with the sessions of the original log and then 2) to extract sequential patterns from each obtained sublog.

6.4.4.1. "Divide and Discover" Method

By means of a clustering method on the extracted sequential patterns, we propose a recursive division of the problem. The developed clustering method is based on patterns summaries and neural networks. Our experiments show that we obtain the targeted patterns whereas their extraction by means of a classical process is impossible because of a very weak support (down to 0.006%).

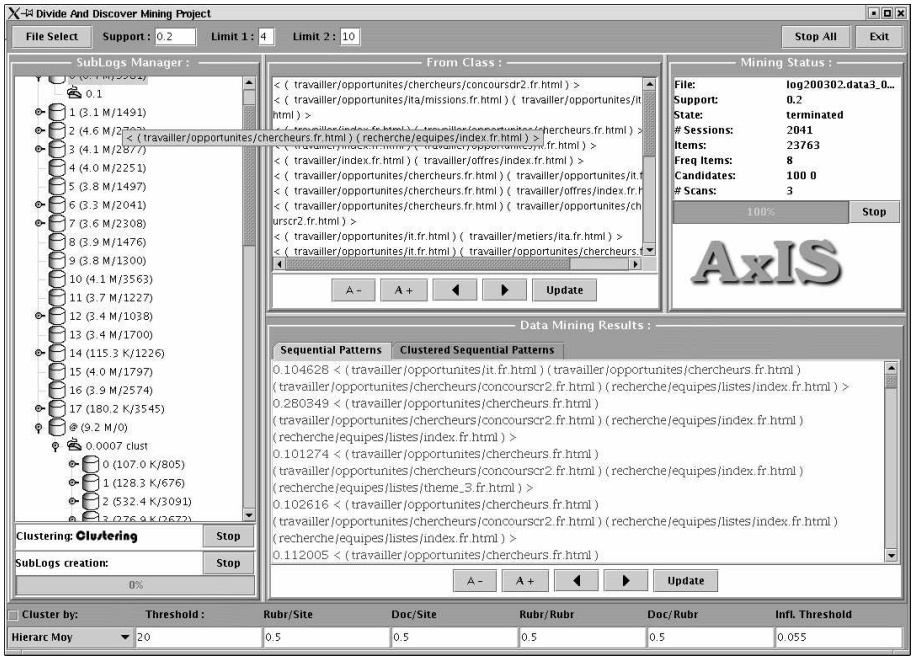

Figure 6. The Divide and Discover system

The very low global support of these patterns is mainly due to the great diversity of the behaviours on the analyzed logs and to the large number of URLs contained in that site. Let us now imagine that we specify a very low support for a classical algorithm. Two problems will then appear:

- The response time will be too long (in most cases, the result won't even be obtained due to the complexity of the process).
- The amount of frequent patterns generated by this process (in the case the process ends) would be very large.

The principle of our first method is thus the following:

1. Extracting sequential patterns on the original log.
2. Clustering these sequential patterns.
3. Dividing the log according to the clusters obtained above. Each sub-log contains sessions from the original log, corresponding to at least one behaviour of the cluster which enabled to create this sub-log. A special sub-log is then created to collect the sessions from the original sub-log which do not correspond to a cluster from the previous step.
4. For each sub-log, apply this whole process (recursively).

6.4.4.2. "Cluster and Discover" Method

In the context of Tanasa's thesis, we proposed another hybrid method called "Cluster and discover". The main objective was to overpass the main limit of our first method i.e. the great number of manual iterations to do and also to increase the quality of extracted patterns. Here the developed clustering method is based on sessions summaries and neural networks. Our second method proceeds as follows :

1. Cluster the sessions with a neural network based algorithm.

2. Group the atypical clusters into a single cluster and for each cluster create the associated sublog.
3. Extract sequential patterns from the sublogs.
4. Re-calculate the support of each sequential patterns as the real support of the pattern on the original log.

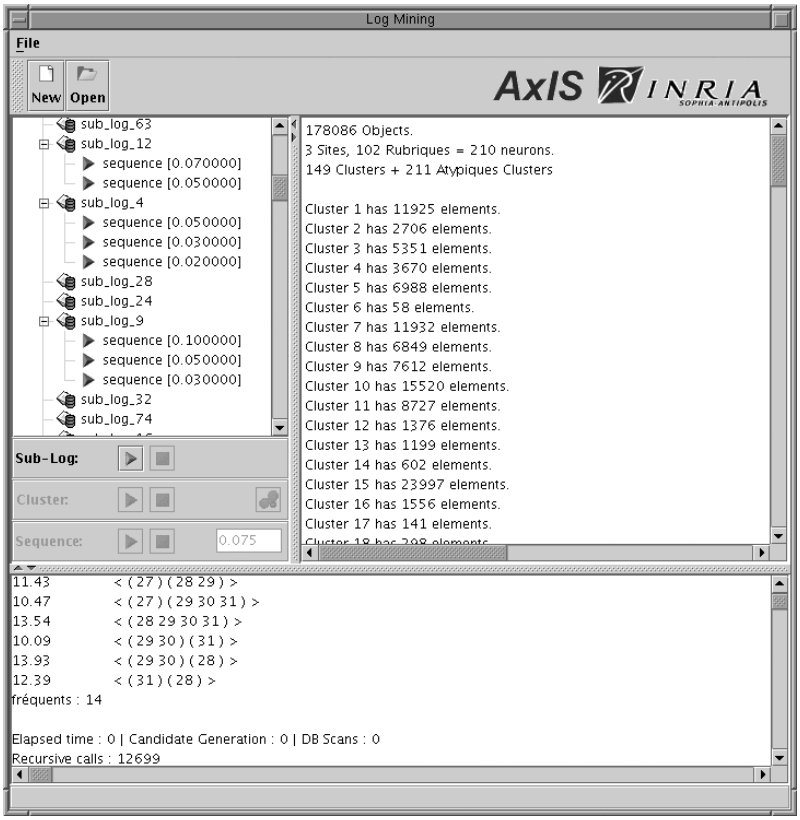

Figure 7. The Cluster and Discover system

6.4.4.3. Experimental Results

We present here some of the discovered behaviours:

- C1: with the common prefix **travailler/opportunites/**:
  <(it.fr.html) (ita/missions.fr.html) (ita/concoursit.fr.html) (ita/annales2001/index.fr.html)>. This behaviour is related to ET jobs opportunities.
- C2: <(travailler/opportunites/chercheurs.fr.html)
  (travailler/opportunites/chercheurs/concourscr2.fr.html)
  (recherche/equipes/index.fr.html) (recherche/equipes/listes/index.fr.html)>.
  This behaviour is related to the career opportunities offered by the Inria. The users read the job opportunities page, then the page describing the competitive selection and finally the pages describing the research teams.

- C3: <(scripts/root.exe) (c/winnt/system32/cmd.exe)
  (..%255c../..%255c../winnt/system32/cmd.exe)
  (..%255c../..%255c/..%c1%1c../..%c1%1c../..%c1%1c../winnt/system32/cmd.exe)
  (winnt/system32/cmd.exe) (winnt/system32/cmd.exe) (winnt/system32/cmd.exe)>.
  This behaviour is typical of a search for a security hole in the system. Usually, these attacks are programmed once and then shared and used by different individuals. This could explain the high relative support (S1) of this behaviour (more than 80%).

We have done a comparison by using the same sequence summaries for the clustering algorithm used by the two methods. For the D&D method we needed several iterations and parameter tuning for finding some of the behaviors listed above. C&D only needed two iterations for finding them, and a third one for computing the global support. Because of the very weak support, these presented behaviors could not be found using only a classical algorithm for frequent pattern mining (i.e. PSP [8]).

6.4.5. Eye Tracking Data Based Recommandation Computation : e-behaviour

Participants: Karine Deletre, Olivier De Maeyer, Sémi Gaieb, Mihai Jurca, Sébastien Simard, Brigitte Trousse.

Key words: *eye-tracking*, *adaptive interface*, *evaluation*, *visual user behaviour*, *navigation*, *web site*.

We process this year the third experiment planned in the e-Behaviour project which is a common project with the LPEQ laboratory of the University of Nice Sophia Antipolis http://www-sop.inria.fr/COLOR/2001/e-BEHAVIOUR.html. This project was partially funded by the COLOR program the first year (2001-2002).

Two students of the DESS Ergontic of UNSA with S. Simard within its training course of DESS Ergontic directed by T.Colombi and B.Trousse carried out an experiment. This experiment aims at validating the assistance brought by the built recommender systems (according to their level of experience on the Web) and at study the impact of utilization of this recommender on the users behaviour [7] (cf. Figure 4). Futhermore this experiment enabled us to validate the current version of CBR*TOOLSput on a CVS server.

In 2002, an experiment was carried out with two groups of 12 subjects, each group having to carry out 8 predefined tasks, a group without assistance and another with assistance based on non visual user behaviours. This year we carried out the same tasks with a third group (having the same charcateristics than the two others, for example level of expertise in internet etc.) and with an assistance based on « visual » and non visual user behaviour.

A statistical analysis highlighted the assistance brought by the recommenders and their impact on the visual behaviour of the users from the two "with assistance" groups. The experiment in 2003 with the second recommender (reasoning on visual and non visual user behaviours) showed good results compared with the group without assistance but more data analysis are necessary to compare precisely the benefits of the two recommenders.

The main results of this project are:

- Conception and development of an experimental system called « e-behaviour system ». It allows us to acquire visual and non visual user behaviour during browsing tasks inside a Web site. Such a web site is added by an adaptive recommender system for supporting browsing in a Web site.
- Applying our « e-behaviour system » for a part of the national Inria site related to the Inria Activity reports (2002).
- Three user groups have been constitued: one without support, one with a non-visual data based support and another one with a support based both on visual and non visual user behaviours.

6.4.6. Design and Development of an EyeTracking Data Oriented Mining Tool

Participants: Patrick Guirchowski, Mihai Jurca, Brigitte Trousse.

After our first experience in manipulating Eye-tracking data (cf. the e-behaviour project), we decided to make a state of the art related to the use of clustering technics applied on visual data from users interacting with an information system or a Web site. We noted that there is no work done. So we designed and developed a tool for clustering visual and non visual data as those obtained in the e-behaviour project [42]. Our objective is now to apply our main AxIS clustering methods on these data in order to extract some interesting patterns.

This tool allows analyzing of the user activity on the RA Web site. The RA Web site is a Web site that contains the Activity Report's of INRIA teams and provides to users, which visit this site, a Web pages recommendation mechanism like BROADWAY-WEB, and a trace of the user eyes movement. All these data are recorded in a relational database system.

This tool is composed on two main components:

- Server side component that build a copy of a page visited by an user (hosted on RA Web site)
- GUI component that implement the user interface with the application* *and communicates with the server side component

With this software tool we can see details about all the users sessions; details like user name, user task, visited pages, user eyes traces on every page etc.

Also we can analyze user sessions to find the similarities between two sessions or to retrieve other statistical data. These analyzes are made using classification algorithms developed by our team (e.g. PspID).

The user interface with RA Usage Analyzer is friendly and makes possible creation of more analyzes (MDI interface). All results of analyses are stored in the relational database and can be retrieved for visualization using RA Usage Analyzer.

6.4.7. Applying Web Usage Mining on Usage Data

Participants: Fabient Benoit, Brieuc Conan-Guez, Mihai Jurca, Yves Lechevallier, Florent Masseglia, Fabrice Rossi, Doru Tanasa, Brigitte Trousse.

In 2003, we started to apply different clustering algorithms on Web Usage Data [15] and differetn KDD algortihms eye-tracking data issued from our three experiments of the « e-behaviour » project :

- Sequential pattern extraction : the first goal of the Benoit's internship [39] (director: F. Masseglia) was to implement the SPAM algorithm for sequential patterns. The difficulty was in the way of implementing the algorithm since the efficiency relied on the data structures (based on bit vectors). Once the implementation realised, a second goal was to propose and realise improvements for this algorithm (SPAM+). Finally the student had to propose experiments of SPAM+ on eyetracking data.
- Clustering : we started to apply Kohonen 's neural maps on such data and also to study how to apply our results in « clustering of functional data » on such data.

This work will be going on next year in order to address the interpretation step.

6.4.8. Geography of the Internet

Participant: Eric Guichard.

This new subject refers to the geography of the internet and of cyberspace. It interests specialists of metrology of the internet, geographers, sociologists, physicists. David T. Horn, an American student of the DESS of ENSSIB (« Ecole nationale supérieure des sciences de l'information et des bibliothèques ») produced under Guichard's direction an exhaustive bibliography of this subject.

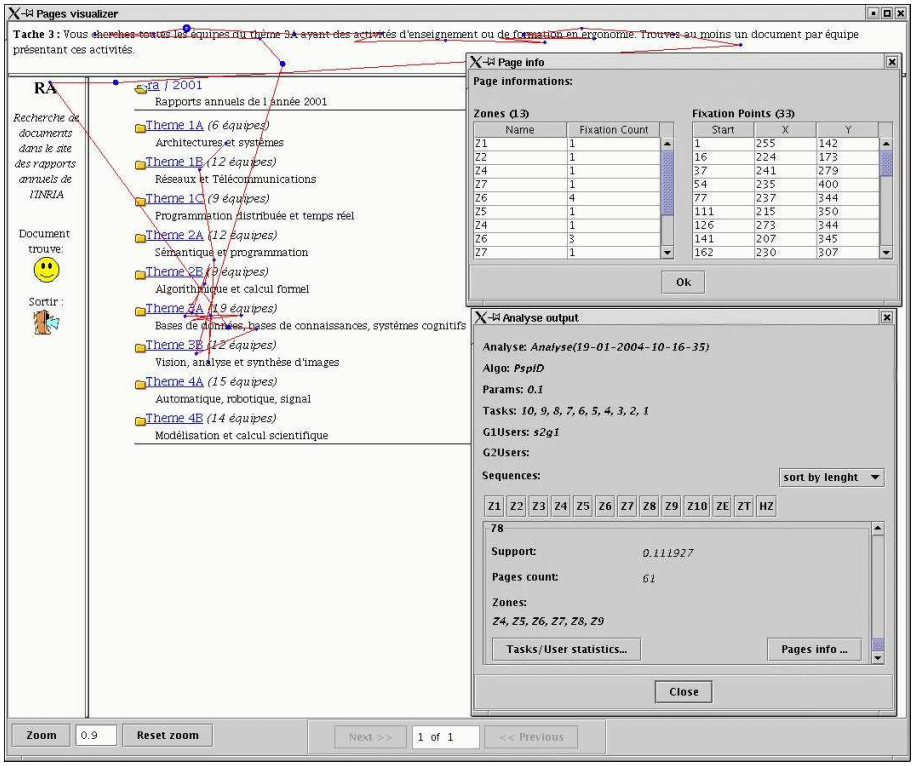

Figure 8. EyeTracking Data Oriented Wum Tool

7. Contracts and Grants with Industry

7.1. Industrial Contracts

7.1.1. EDF: Curve Clustering and Web Usage Mining

In 2002 we have signed a new contract with E.D.F's Research Direction for two years. The objective related to this new contract was to follow on the previous collaboration on curses classification, more precisely on main client consumption curves. We show that curves are well adapted for a classification through a Kohonen Map because the displayed representation is immediately understandable by any current user.

7.1.2. EPIA an RNTL project (2003-2005)

The EPIA project "Evolution of an Adaptive Information Portal" has been labeled by RNTL 2002. The partners are: Dalkia, Mediapps et Inria (AxIS Sophia Antipolis et Rocquencourt). Due to some delay, the starting date of such a contract is officially the 1 September 2003.

We have installed a license of Net.Portal on our experimental machine axis and delivered the first deliverable [33] in cooperation with mediapps related to the specification of the trace engine inside the mediapps.net system.

Our first meeting was in Paris 15-16 Oct.(M. Jurca, Y. Lechevallier and B. Trousse)

7.1.3. Industrial Contacts

- Toyota IT Center (July 22-23) : Axis presentation by B. Trousse at Inria Sophia Antipolis
- Toyota (October 27): AxIS presentation by B. Trousse to Mr. Masahiro Miyaji and Mr. Yuji Ninagawa and Mr. Mikio Danno
- Alcatel (December 9) : presentation by P. Loyer of the « Mediaspace » and « Space for Science » projects at Eurecom (Sophia Antipolis)
- IRD, « Institut de Recherche pour le Développement », ex-ORSTOM: participation of the supervision of theInvited post-doc research of Agathe Petit on the theme *Geography of the internet* (18 months): E. Guichard.
- Contacts with some industrial partners of the « Laboratoire des Usages des NTIC » of Sophia Antipolis (France Télécom Sophia Antipolis, Alcatel Cannes).

8. Other Grants and Activities

8.1. Regional Actions

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

- UNSA LPEQ (2001-2003): The e-Behaviour project is a cooperation between AxIS and LPEQ (UNSA) http://www-sop.inria.fr/COLOR/2001/e-BEHAVIOUR.html on the topic of analyzing visual and non visual behaviours of internautes on the Web [42]. LPEQ Visit November 12: B. Trousse and D. Tanasa
- INRIA VISA Action: Collaboration with G. Gallais et P. Rives of the Action VISA (Inria Sophia Antipolis), M. Riveill of the Rainbow team (I3S UNSA) on the topic "adaptation and evaluation of services in the context of transports". The MobiVIP proposal with 22 partners was made to AO PREDIT, such a proposal has been accepted and notified at the end of 2003.
- INRIA Direction (July 8 and 10): presentation (both at Sophia and Rocquencourt) of AxIS research work to M. Berthod, vice scientific chair of Inria

- Laboratoire des Usages, CNRT, Sophia Antipolis: the laboratory was created on the initiative of the Centre national de recherche technologique (CNRT) of Sophia-Antipolis. The laboratory is structured as a "GIS" ("Groupement d'Interet Scientifique") between the four founders: CNRS (STIC, SHS), GET, INRIA and UNSA. The laboratory aims at observing the current usages of "TIC", and to anticipate future usages « by a pluridisciplinary research gathering technologists, economists, sociologists, ergonomists, marketing specialists with rigorous methodologies around effective technological platforms and relevant and various users ».
  B. Trousse is a member of the scientific committee and also a substitute member of the management committee. Participation of AxIS members 1) via the e-behaviour project, 2) via the participation to various seminars organized by its director: (B. Trousse, F. Masséglia et D. Tanasa) and finally via administrative support provided by S. Honnorat.
- Laboratoire LISE-CEREMADE (UMR CNRS STIC Maths) of the University of Paris IX Dauphine : research activity with F. Rossi on the functional approach in the non supervised clustering methods and mainly the neural networks [31][30][32]

8.2. National Actions

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 (URL= http://rtp12.loria.fr/), the AxIS team (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, and more specifically to the seminars hold on Sept 08, 2003 and Nov. 03, 2003, Lyon

8.2.2. CNRS RTP 15: « économie, organisation & STIC »

In the context of the pluri-disciplinary thematic network 35
(URL=http://www.cnrs.fr/STIC/actions/rtp/PagesRTP/RTP35.html), the AxIS team participated to 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), more specifically to the seminar hold on November 20, 2003 Paris. Doru Tanasa maintained the web server of this action.

8.2.3. CNRS AS 120 « Web sémantique »

AxIS was involved in the CNRS Specific Action (AS 42) on *"Semantic Web"*
http://www.lalic.paris4.sorbonne.fr/stic/:.

8.2.4. GDR-I3

Members of the AxIS team participated in two working groups of GDR-PRC I3 National Research Group « Information - Interaction - Intelligence » of CNRS
http://www-timc.imag.fr/I3/

- Working Group (GT) on Data Mining animated by P. Poncelet and J.M. Petit (June 12, 2003) : F. Masseglia.
- Member of the GRACQ (*Groupe de Recherche en Acquisition des Connaissances*) (http://www.irit.fr/GRACQ): B. Trousse.

8.2.5. EGC « National Group on Mining Complex Data »

AxIS participated to the Working Group "Fouille de données complexes" created by D.A Zighed in June in the context of the EGC association.
F. Masséglia participated at the first meeting (sept 18) and is co-animator with O. Boussaid (ERIC, Lyon) of one of the two emerged topics related to the Data pre processing.

8.2.6. Inria Action : Syntax

We took part in Syntax, an INRIA national research and development action on electronic documents. AxIS contribution will consist in integrating clustering methods for producing classes of documents into the syntax platform, to study a formal approach to check the level of consistency and to analyze the usefulness of integrating an adaptative system for supporting information retrieval

8.2.7. CNRS « Action Concertée : Histoire des savoirs »

Participation: M. Csernel and Y. Lechevallier
This action associates several french research teams from various research field, such as computer science, data analysis, and Sankrit literature. The main goal of this action is to provide help for the construction of critical edition, in Sankrit, of Indian manuscripts, and to provide pertinent information about the manuscripts classification (construction of cladistic trees). The expected tools will not be dedicated to Sankrit language.

8.2.8. Others Collaborations

- MAD Laboratory of the university of Bordeaux (M. Chavent)
- University of Technology of Compiégne (Gérard Govaert)

8.3. European Actions

8.3.1. IST European Project: ASSO

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

8.3.1.1. Objectives

The objective of the ASSO project is to design methods, methodology and software tools for knowledge extraction from multidimensional complex data saved in large databases of Statistical Offices and Administrations. The project aims at: 1) extending standard data analysis to more complex data taking care of metadata; 2) providing better explanation of statistical results by new interpretation ; and finally 3) providing new tools for the creation, the manipulation and the dissemination of knowledge in order to get better and easier communication between community members.

8.3.2. 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.
- **Metadata Model Design** 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).
- **Data Format SOM library** Every algorithm within the project uses the 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.

- **DB2SO Improvement 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.
- **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** 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. It is an adaptation to symbolic data, of the cluster validation approach proposed for 'numerical' data by Bel Mufti and Bertrand (2001).

For 2003, see the new research contributions in Sections 6.4.1 and 6.4.2.

8.3.3. IST European Network : Ontoweb

AxIS participated to the european network called 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). http://www.ontoweb.fr/.

8.3.4. COST Action 282

COST Action 282 (2001-2005) : "Knowledge Exploration in Science and Technology". B. Trousse is one of the two french representatives in the Management Committee" (3rd meeting in October in Milano).

8.3.5. Others Collaborations

- University of Naples II (Italy): M. Csernel has been invited for two seminars on F.N.S and on the knowledge extraction (as symbolic objects) from a database (may 6-7)
- Facultés Universitaires Notre-Dame de la Paix à Namur (Belgique):

8.4. International Actions

8.4.1. Brazil

We followed our collaboration with F. A. T. de Carvalho [21][20][11] from Federal University of Pernambuco (Récife) and his team. We welcomed F.A.T. de Carvalho, in march-april and september;
In 2003 we started a new collaboration on WEb usage Mining : D. Tanasa preprocessed, with our Web Log preprocessing tool, Web logs from Recife University Web server for one of his students.

8.4.2. Canada

Collaboration avec A. Ciampi of Mc Gill Univerity
Contact avec S. Proulx of the University of Montréal: visit of E. Guichard (summer) [12].

8.4.3. Morocco

AxIS is involved in a France-Morocco thematic network in software engineering and B. Trousse co-supervises with Abdelaziz Marzark (UNiversity of Casablanca) a Ph-D student in this context, H. Behja (ENSAM, Meknès, Morocco). Third Joint Meeting of the France -Morocco Cooperation Program (« Atelier STIC ») between the three current thematic networks at Rabat dec 15-16 (B. Trousse and H. Behja). H. Behja presented his thesis work to the scientific comittee on december 16.

8.4.4. Tunisia

P. Bertrand follows his collaboration with Bel Mufti Ghazi of « Ecole Supérieure des Sciences Economiques et Commerciales » (Tunis)

9. Dissemination

9.1. Promotion of the Scientific Community

9.1.1. Journals

B. Trousse is a member of

- the edition board of the Co-Design journal (editor-in-chief S. Scrivener, Coventry University, UK);
- the scientific board of RSTI -« séries ISI, L'OBJET, RIA, TSI » (Hermès publisher) (meeting on april 10);
- the edition board of the RIA journal (« Revue d'Intelligence Artificielle ») (Hermès publisher) (editor in chief: M. Pomerol);
- the edition board of the I3 electronic journal of the GDR-I3 (editor-in-chief: C. Garbay et H. Prade) http://www.Revue-I3.org/.

Y. Lechevalier is a member of the edition board of "Journal of Symbolic Data Analysis (revue électronique http://www.jsda.unina2.it.

F. Masseglia was twice (june, november) a reviewer for the KAIS journal in 2003 (Knowledge and Information Systems, http://www.emba.uvm.edu/~xwu/kais/

9.1.2. Program Committees

Different AxIS members participated to Programm Committees at national or international levels:

9.1.2.1. National Conferences/Workshops

- EGC'03 (Lyon, January) Extraction et Gestion des Connaissances
- SFdS'03 (Lyon) Journées de la Société Française de Statistique : Y. Lechevallier
- "Entreposage et fouille de données " Workshop of SFDS (« Société Française de Statistique »), Lyon, june 2003: M-A Aufaure
- INFORSID 2003 (Nancy, june) : F. Masseglia
- H2PTM'03 Hypermedia, Hypertexts, Products, Tools and Methods (september 24-26) http://h2ptm.univ-paris8.fr/h2ptm03.htm : B. Trousse

9.1.2.2. International Conferences/Workshops

- ECCBR'03 European Conference of Case-Based Reasonng (UK): B. Trousse
- JFT03 Journées Francophones de la Toile, june 30-july 2 (Tours) : B. Trousse
- Colloque "Mesures de l'internet" [41]: E. Guichard (president), B. Trousse, Y. Lechevallier
- CoPSTIC'03, 1st Plenary Information Technology Conference, 11-13 December 2003, Rabat Morocco : B. Trousse
- Workshop on " Multimedia Discovery and Mining ", collocated with ECML/PKDD 2003): M-A Aufaure
- IFCS'04 (Chicago) Conference of the International Federation of Classification Societies: Y. Lechevallier

9.1.3. Invited Seminars

- F. Masséglia. Title « Fouille de données : algorithmes et applications pour l'extraction de motifs séquentiels »
  - University of Strasbourg I in the LSIIT laboratory (AFD team).
  - University of Lyon, ERIC laboratory, 1 december.
- B. Trousse
  - Forum « Systèmes & Logiciels pour les NTIC dans le Transport » , « Nouveaux Systèmes d'Information au Service des Usagers », CNAM, Paris, may 22: B. Trousse Title: « Calcul de recommandations personnalisées sur le Web: application aux sites Web sur les transports »
    http://www.inrets.fr/services/manif/ForumNTIC/index.htm
  - Lyon University, AS Disco Challenge, october 18. Title: Web Log Mining
  - ENST Paris, AS « Mesures des flux de l'internet », november 20. Title: "Mesurer les comportemens des utilisateurs des sites".
- : M-A Aufaure [36], University of Paris IX Dauphine, CEREMADE Laboratory
- Y. Lechevallier : Workshop at ILYADE (Lyon) december 2003 [43]
- Eric Guichard
  - « The digital divide today » [13], Sustainable Ties in the Information Society Conference, University of Tilburg, Netherlands (26–28 March).

9.1.4. Organization of conferences or workshops

- Organisation of the International Workshop *Mesures de l'internet* (Nice, Hotel Westminster, 12–14 May, http://www-sop.inria.fr/axis/cmi): E. Guichard (responsible), S. Honnorat and B. Trousse from AxIS, D. Sergeant and M-H Zeitoun (BMC) with the support of SEMIR and GENER services of INRIA Sophia Antipolis.
- Participation to the organisation of the Pluri-disicplinary Working day *Cartographie, standards ouverts et partage de données* (Paris, Telecom, November 20): E. Guichard
- Animation of the round table « Sociabilités en ligne » (days *éditions électroniques*, BPI — Public Library of Information —, Georges Pompidou museum), Paris, 14 Novembre : E. Guichard
- CNRS Representative in the scientific committee for the exhibition *Mission Biospace* (Cité de l'Espace, Toulouse).

9.1.5. AxIS Web Server

AxIS maintains his Web site allowing the access to a lot of information, mainly the software developed in the team, our publications, a list of relevant events (conferences, workshops) for the team and the information related to the conferences we organise.
http://www-sop.inria.fr/axis/.

9.1.6. Activities of General Interest

- T. Despeyroux is AGOS president, permanent member of the « commission technique paritaire (CTP) » and also of the « conseil d'administration » of INRIA as staff representative.

- F. Masséglia: Member of the working group on “Politique des sauvegardes de l’unité de Sophia Antipolis”.
- B. Trousse is a member of the scientific committee and also of the decision committee of the « Laboratoire des Usages des NTIC » de Sophia Antipolis.

9.2. Formation

9.2.1. University Teaching

AxIS team is involved in teaching in differents university diploma and is an associated team for « teh STIC Doctoral school » at the University of Nice-Sophie Antipolis.

Teaching activity of the sophiapolitan component of AxIS :

- « DEA Informatique » (resp Mr Kounalis) à l’UNSA Sophia Antipolis:
- « DESS “Ergonomie et NTIC » (resps T. Baccino et J. Araszkiewiez) à l’UNSA on Web Usage Analysis and Adaptive Services For Web Information retrieval : B. Trousse.
- « Licence professionnelle Franco-italienne STID: Statistiques et Informatique Décisionnelle » (resp. J. Lemaire) à l’UNSA, Menton:
  - supervision of a student project (100h by students, 20 students, 50h supervised) on *Mining Web HTTP Logs From Multi Inria Sites* : D. Tanasa, B. Trousse (resp.).
- IUT GTR (Génie des Télécommunications et Réseaux) à l’UNSA, Sophia Antipolis : TP (140h 2002-2003) de *Réseaux* : S. Gaieb.
- DEA of social sciences ENS-EHESS (Feb-May), Cycle *Lectures de l’internet*: E. Guichard (responsible).

Teaching activity of the parisian component of AxIS :

- « 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.
- « DEA MIASH », University of Paris 5, course on data mining for complex data (spatial and multimedia) [35], march 2003. M-A Aufaure
- « 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.
- « EPFL » Information & Communication Faculty, Lausanne. Course on Data Mining and Data Warehouse [34], 3rd-4th year of the engineer school, from march to june 2003. M-A Aufaure

9.2.2. Student Visits

september 18: AxIS presentation (B. Trousse) to 15-20 students of the Master in Information Systems Management (http://msi.ensg.ign.fr)

9.2.3. Participation to Summer Schools

- A Joint International Summer School was organized in Lisbon (July 2003)by IASC-IFCS. A topic on Data Mining with SODAS approach and SODAS software has be presented by Y. Lechevallier [19].
- A summer school was organized in Greece (October 2003) for ASSO Group (Sodas software) and a course in DB2SO was presented by A. El Golli.

9.2.4. PhD Thesis

PhD in progress:

1. **Aicha El Golli** (starting date: end 2001), « Cartes topologiques et modèles statistiques : application à la classification de données symboliques », University of Paris IX Dauphine (director : E. Diday).
2. **Doru Tanasa**, (starting date 2001), "Trace et analyse de l'usage pour l'aide à la reconception d'un site Web", University of Nice-Sophia Antipolis (director: B. Trousse).
3. **Sergiu Chelcea**, (starting date end 2002), "Classification de profils utilisateurs d'un site Web", Université de Nice-Sophia Antipolis (directors: J. Lemaire et B. Trousse with the support of P. Bertrand on 2-3 CAH).
4. **Hicham Behja**, (starting date end 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 the France-Morocco cooperation (2002-2005).

Y. Lechevallier was a member of the following Ph-D committees in 2003 :

- Isabelle Marie-Joseph « Méthodologie de diagnostic appliquée à la maintenance préventive d'unités de production d'électricité en sites isolés » in march t the Université of Guyanne,
- Rodolphe Priam « Méthodes de carte auto-organisatrice par mélange de lois contraintes. Application à l'exploration dans les tableaux de contingence textuels » in November at the Université of Rennes 1,
- Francois-Xavier Jollois "Contribution de la classification automatique à la Fouile de données" in December at the University of Metz,

9.2.5. Internships

1. **L. Baubois** [38] (University of Paris IX Dauphine, « UFR Informatique de Gestion », Miage, 2002-2003) « Extension de DB2SO » (resp: M. Csernel),
2. **A. Baldé** [37] (« DEA Informatique Systèmes Intelligents » , University of Paris IX-Dauphine) [37] « Extraction de métadonnées sur les prototypes issus de la classification d'objets symboliques » (resp: M-A Aufaure et Y Lechevallier);
3. **G. Bernon, O. Famin, S. Herstain** [40] (Télécoms 3, ISPG), « Représentation Graphique des Cartes de Kohonen » (resp: Y. Lechevallier);
4. **F. Benoit** [39] (University of Nice-Sophia Antipolis, ESSI 2002-2003 (3rd year). Subject: Implementation and improvement of the SPAM algorithm, application on eyetracking data. (resp: F. Masséglia)

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 (cf. section Bibliography). More we have participated to the following conferences or workshops:

- JFT03 « Journées Francophones de la Toile », Tours, Laval, June 29-July 3: B. Trousse
- Plateforme AFIA, July, Laval, France : B. Trousse
- Annual RNRT meeting, Lille, January 27-28 : E. Guichard
- « ICT: wrong theories, real questions », 53rd Pugwash Conference on Science and World Affairs, Halifax (Canada), 17–22 July.: E. Guichard
- « Scientific aspects of digital divide », Uppsala University (Sweden), November 6. E. Guichard
- « Ecritures cartographiques », Conference *Cartographie, standards ouverts et partage de données*, Paris, ENST, November. E. Guichard
- « Rencontres d'Autrans-Vercors, Internet: les réseaux de perosnnes January 8-11 » : E. Guichard http://www.autrans.net/2003/programme.html

10. Bibliography

Major publications by the team in recent years

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

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

[3] G. CARAUX, Y. LECHEVALLIER. *Règles de décision de Bayes et méthodes statistiques de discrimination.* in « Revue d'intellignece artificielle », number 2-3, volume 10, 1996, pages 219-284.

[4] M. CHAVENT. *A monothetic clustering method.* in « Pattern Recognition Letters », 1999, pages 989-996.

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

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

[7] 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.

[8] F. MASSEGLIA. *Algorithmes et Applications Pour l'Extraction de Motifs Séquentiels Dans le Domaine de la Fouille de Données : de l'Incrémental au Temps Réel.* Ph. D. Thesis, Universite de Versailles St-Quentin en Yvelines, France, January, 2002.

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

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

Articles in referred journals and book chapters

[11] M. CHAVENT, F. A. T. D. CARVALHO, Y. LECHEVALLIER, R. VERDE. *Trois nouvelles méthodes de classification automatique de données symboliques de type intervalle.* in « Rev. Statistique Appliquée », volume Vol 4, December, 2003, pages 5–29.

[12] E. GUICHARD. *Mesures de la fracture numérique.* S. PROULX, F. JAURÉUIBERRY, editors, in « Internet, nouvel espace citoyen ? », l'Harmattan, 2003, pages 37–47.

[13] . GUICHARD. *Does the 'Digital Divide' Exist?*. P. VAN SETERS, B. DE GAAY FORTMAN, A. DE RUIJTER, editors, in « Globalization and its new divides: malcontents, recipes, and reform », Dutch University Press, 2003.

[14] G. HÉBRAIL, Y. LECHEVALLIER. *Data mining et analyse des données.* in « Analyse des données », Hermes, June, 2003, pages 340-360.

Publications in Conferences and Workshops

[15] M. ARNOUX, Y. LECHEVALLIER, D. TANASA, B. TROUSSE, R. VERDE. *Automatic Clustering for the Web Usage Mining.* in « Proceedings of the 5th Intl. Workshop on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC03) », Editura Mirton, Timisoara, D. PETCU, D. ZAHARIE, V. NEGRU, T. JEBELEANU, editors, pages 54 – 66, 1-4 October, 2003.

[16] A. BENEDEK, B. TROUSSE. *Adaptation of Self-Organizing Maps for Case Indexing.* in « 27th Annual Conference of the Gesellschaft fur Klassifikation », Cottbus, Germany, 12-14 march, 2003.

[17] H. H. BOCK, Y. LECHEVALLIER. *The Analysis of Symbolic Data.* in « Statistical Computing 2003 », Statistical Computing 2003, Reisensburg, Germany, July, 2003.

[18] M. BOUET, M.-A. AUFAURE. *Towards better large image databases exploration and exploitation : image mining & visual ontology.* in « Fourth International Workshop on Multimedia Data Mining, in conjunction with the 9th ACM Conference on Knowledge Discovery & Data Mining », Washington, USA, August, 2003.

[19] P. BRITO, Y. LECHEVALLIER, C. MARCELO. *Data Mining with SODAS approach and SODAS software.* in « IASC-IFCS Joint International Summer School », JISS-2003, Lisbonne,Portugal, July, 2003.

[20] F. A. T. D. CARVALHO, M. CSERNEL, Y. LECHEVALLIER. *Partioning of Constrained Symbolic Data based on Dissimilarity Function.* in « 27th Annual Conference of the Gesellschaft für Klassifikation », GlKl, Cottbus, germany, March, 2003.

[21] M. CHAVENT, F. A. T. D. C. Y. LECHEVALLIER, R. VERDE. *Classification automatique d'objets décrits par un vecteur d'intervalles.* in « XXXVèmes Journées de Statistique », SFdS, pages 317-320, Lyon, France, June, 2003.

[22] S. CHELCEA, P. BERTRAND, B. TROUSSE. *Agglomerative 2-3 Hierarchical Clustering: theoretical improvements and tests.* in « 27th Annual Conference of the Gesellschaft fur Klassifikation », Cottbus, Germany, 12-14 March, 2003.

[23] A. E. GOLLI, B. CONAN-GUEZ. *Adaptation des cartes topologiques auto-organisatrices aux tableaux de dissimilarités.* in « Méthodes et Perspectives en Classification (10èmes Rencontre de la Société Francophone de Classification) », Presse Académiques Neuchâtel, Y. DODGE, G. MELFI, editors, pages 99 – 102, Case Postale 1420 CH-2001 Nechâtel Suisse, 10-12 September, 2003.

[24] A. E. GOLLI, Y. LECHEVALLIER. *Extraction de classes homogenes et creation d'objets symboliques.* in « XXXVemes Journees de Statistique », 2-6 june, 2003.

[25] Y. LECHEVALLIER, R. VERDE. *General dynamic clustering methods on symbolic data tables.* in « Classification and Data Analysis Group », CLADAG2003, pages 245-248, Bologne,Italie, September, 2003.

[26] F. ROSSI, B. CONAN-GUEZ. *Estimation consistante d'un modèle paramétrique fonctionnel en présence de discrétisation aléatoire.* in « Actes des trente-cinquièmes journées de la SFDS », June, 2003.

[27] D. TANASA, B. TROUSSE. *Le prétraitement des fichiers logs Web dans le "Web Usage Mining" multi-sites.* in « Journées Francophones de la Toile (JFT'2003) », Tours, June - July, 2003.

[28] S. YU, M.-A. AUFAURE, N. CULLOT, S. SPACCAPIETRA. *A collaborative framework for location-based services.* in « CAiSE 2003 », Klagenfurt, Velden, Austria, 16-20 June, 2003, short paper.

[29] S. YU, M.-A. AUFAURE, N. CULLOT, S. SPACCAPIETRA. *Location-based spatial modeling using ontology.* in « 6th AGILE Conference on Geographic Information Science », Lyon, France, 24-26 April, 2003.

Internal Reports

[30] F. ROSSI, B. CONAN-GUEZ. *Estimation consistante des paramètres d'un modèle semi-paramétrique pour des données fonctionnelles discrétisées aléatoirement.* Technical report, number 0334, LISE/CEREMADE, http://www.ceremade.dauphine.fr/, October, 2003, and also INRIA Research Report, number 2-7261-1270-6 , UR Rocquencourt, december.

[31] F. ROSSI, B. CONAN-GUEZ. *Functional Multi-Layer Perceptron: a Nonlinear Tool for Functional Data Analysis.* Technical report, number 0331, LISE/CEREMADE, http://www.ceremade.dauphine.fr/, September, 2003, and also INRIA Research Report, number 2-7261-1269-2 , UR Rocquencourt, december.

[32] F. ROSSI, B. CONAN-GUEZ. *Un modèle semi-paramétrique neuronal pour la régression et la discrimination sur données fonctionnelles.* Technical report, number 0338, LISE/CEREMADE, http://www.ceremade.dauphine.fr/, October, 2003, and also INRIA Research Report, 2-7261-1271-4, UR Rocquencourt, december.

Miscellaneous

[33] B. ABERT, B. TROUSSE, M. JURCA. *Moteur de traces de Mediapps.net: cahier des charges et d'évaluation (D1).* December, 2003.

[34] M.-A. AUFAURE. *Course on Data Mining and Data Warehouse.* 3rd-4th year of the engineer school, Information & Communication Faculty, EPFL, Lausanne, Switzerland, March - June, 2003.

[35] M.-A. AUFAURE. *Course on data mining for complex data (spatial and multimedia).* DEA MIASH, University of Paris 5, France, March, 2003.

[36] M.-A. AUFAURE. *Vers une recherche intelligente dans des grandes bases d'images.* Séminaire LISE-CEREMADE, Université Paris-Dauphine, 26 June, 2003, organised by Y. Lechevallier (Inria AxIS) and Mireille Summa (Ceremade).

[37] A. BALDE. *Extraction de métadonnées sur les prototypes issus de la classification d'objets symboliques.* 2003, Internship Report.

[38] L. BAUBOIS. *Extension de DB2SO.* 2003, Internship Report.

[39] F. BENOIT. *Méthodes de data Mining.* 2003, Internship Report.

[40] G. BERNON, O. FAMIN, S. HERSTAIN. *Représentation Graphique des Cartes de Kohonen.* 2003, Internship Report.

[41] . GUICHARD. *Mesures de l'internet (actes provisoires).* May, 2003, Inria Sophia Antipolis.

[42] P. GUIRCHOWSKI. *Analyse des comportments visuels pendant la navigation sur le Web, Catégorisations et Représentations Graphiques.* September, 2003, Internship Report, DESS Ergontic.

[43] Y. LECHEVALLIER. *L'Analyse des Données Symbolique : méthodes exploratoires pour l'Analyse de connaissances.* Séminaire du groupe ILYADE, Lyon, December, 2003.

Bibliography in notes

[44] A. AAMODT, E. PLAZA. *Case-Based Reasoning : Foundational Issues, Methodological Variations, and System Approaches.* in « The European Journal oo Artificial Intelligence », number 1, volume 7, 1994, pages 39-59.

[45] A. BENEDEK, B. TROUSSE. *Adaptation of Self-Organizing Maps for CBR case indexing.* in « Proceeding of the 4th International Workshopon Symbolic and Numeric Algorithms for Scientific Computing (toappear) », pages 31-45, Timisoara, Romania, October, 2002.

[46] P. BERTRAND. *Set Systems for Which Each Set Properly Intersects at Most One Other Set - Application to Pyramidal Clustering.* in « IFCS2002, Classification, Clustering, and Data Analysis », IFCS2002, K. JAJUGA, A. SOKOLOWSKI, editors, pages 38-39, Cracow, Poland, juillet, 2002.

[47] M. CANNATRO. *A data Mining Ontology for Grid Programming.* in « Proceedings of SemPGRID' 03 », pages 115-134, 2003.

[48] M. E. FAYAD, D. C. SCHMIDT. *Object-Oriented Application Frameworks.* in « Communication of the ACM », number 10, volume 40, 1997, pages 32-38.

[49] 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) », pages 313-320, Chantilly, 1994.

[50] M. JACZYNSKI, B. TROUSSE. *Broadway : a Case-based System for Cooperative Information Browsing on the World-Wide-Web.* in « Collaboration between Human and artificial Societies. Coordination and Agent-based distributed Computing », LNAI Series, J. A. PADGET, editor, pages 264-283, 1999.

[51] M. JACZYNSKI, B. TROUSSE. *Patrons de conception dans la modélisation d'une plate-forme po ur le*

raisonnement à partir de cas. in « Revue l'Objet », number 2, volume 5, 1999, Numéro Spécial sur les patterns orientés objets, D. Rieu et J-P. Giraudon (guest editors).

[52] R. JOHNSON, B. FOOTE. *Designing Reusable Classes.* in « Journal of Object-oriented programming », number 2, volume 1, 1988, pages 22–35.

[53] T. KOHONEN. *Self-organizing maps.* Springer, Berlin, 1995.

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

[55] 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 », number 3, volume 40, 1997, pages 77-87.

[56] K. MORIK, M. SCHOLZ. . *The MiningMart Approach to Knowledge Discovery in Databases.* in « Intelligent Technologies for Information Analysis », 2003.

[57] 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 », hermes, Paris, S. PESTY, P. SIEGEL, editors, pages 261-288, mars, 1997.

[58] P. RESNICK, H. R. VARIAN. *Recommender systems.* in « Communications of the ACM », number 3, volume 40, 1997, pages 56-58.

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

[60] 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 », Springer-Verlag, S. WESS, K. ALTHOFF, M. M. RICHTER, editors, pages 167-181, 1994.