INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team AxIS

User-Centered Design, Improvement and Analysis of Information Systems

Sophia Antipolis - Rocquencourt

THEME COG

Table of contents

1. Team

Team Leader

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

Team Vice-Leader

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

Administrative Assistants

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

Research Scientists

Thierry Despeyroux [Research Scientist (CR1), Inria Rocquencourt]
Florent Masséglia [Research Scientist (CR2), Inria Sophia Antipolis]
Fabrice Rossi [Research Scientist (CR1), on secondment from October 15, Inria Rocquencourt]
Bernard Senach [Research Scientist (CR1), since November, Inria Sophia Antipolis]
Anne-Marie Vercoustre [Research Scientist (DR2), 75 %, Inria Rocquencourt]

Research Scientists (partners)

Mireille Arnoux [Assistant Prof., Univ. Bretagne Occidentale, Inria Sophia Antipolis]
Marc Csernel [Assistant Prof., Univ. Paris IX Dauphine, Inria Rocquencourt]
Fabrice Rossi [Assistant Prof., Univ. Paris IX Dauphine, until October 15, Inria Rocquencourt]
Brieuc Conan-Guez [Assistant Prof., Univ. Metz, until January 31, Inria Rocquencourt]

Technical Staff

Mihai Jurca [Development engineer, EPIA project, until August 31, Inria Sophia Antipolis]
Aicha El Golli [Research engineer,EPIA project, until October 31, Inria Rocquencourt]
Doru Tanasa [Research engineer, EPIA project, since November 15, Inria Sophia Antipolis]

Ph. D. Students

Abdourahamane Baldé [Univ. of Paris IX Dauphine, Inria Rocquencourt]
Hicham Behja [France-Morocco Cooperation (STIC-GL network), Univ. Hassan II Ben M'Sik, Casablanca, Morocco, Inria Sophia Antipolis]
Sergiu Chelcea [Univ. Nice Sophia Antipolis (UNSA-STIC), Inria Sophia Antipolis]
Alzennyr Da Silva [Univ. Paris IX Dauphine, from October 1st, Inria Rocquencourt]
Alice Marascu [Univ. Nice Sophia Antipolis (UNSA-STIC), since October 1st, Inria Sophia Antipolis]
Doru Tanasa [Univ. Nice Sophia Antipolis (UNSA-STIC), until June 30, Inria Sophia Antipolis]

Visiting Scientists

Teresa Bernarda Ludermir [Prof., Federal Univ. of Pernambuco, Brazil, April, Inria Rocquencourt]
Francisco De Carvalho [Prof., Federal Univ. of Pernambuco, Brazil, April, September-November, Inria Rocquencourt]
Elvira Romano [PhD student, Univ. Federico II, Naples, November-December, Inria Rocquencourt]
Rosanna Verde [Prof., University of Napoli, Italy, March-May, Inria Rocquencourt]
Osmar Zaiane [Associate Prof., University of Alberta, Canada, June, Inria Sophia Antipolis]

Student Interns

Rémi Busseuil [ENS Cachan, June-July, Inria Sophia Antipolis]
Patrick Chastellan [Univ. Montpellier, June-September, Inria Sophia Antipolis & LIRMM]
Alzennyr da Silva [Federal Univ. of Pernambuco, Brazil, April-September, Inria Sophia Antipolis and Rocquencourt]
Marina Dufresne [Univ. Paris XIII Institut Galilée, March-July, Inria Rocquencourt]
Calin Garboni [Univ. of Timisoara, May-Nov, Inria Sophia Antipolis]
Saba Gul [MIT, since September, Inria Rocquencourt]
Selma Kebbache [Univ. Paris I Panthéon-Sorbonne, June-August, Inria Rocquencourt]

Nicomedes Lopes Calvacanti Junior [Federal Univ. of Pernambuco, Brazil, October-March, Inria Rocquencourt]
Alice Marascu [Univ. Nice Sophia Antipolis (UNSA), January-July, Inria Sophia Antipolis]
Mounir Fegas [Univ. Paris Sud XI LRI, April-September, Inria Rocquencourt]
Sofiane Sellah [Univ. Lyon II, March-June, Inria Sophia Antipolis]
Sattisvar Tandabany [Univ. Orsay - ENS Lyon, March-June, Inria Sophia Antipolis]

2. Overall Objectives

2.1. Objectives

Keywords: *KDD, Web mining, data mining, data stream mining, document mining, evaluation, information retrieval, information system, knowledge discovery, knowledge management, recommender system, semantic Web, semantic Web, semantics checking, usage mining, user-centered design.*

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

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

To achieve such objectives, an interdisciplinary approach is necessary and in fact, the AxIS team, which was created in July 2003, regroups people coming from different domains in computer sciences: Artificial Intelligence, Data Mining & Analysis, Software Engineering, all of them being involved in the world of XML documents and information systems.

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

- the structure and content point of view related to the design and the evaluation of static aspects of ISs (architecture, documents),
- the usage point of view related to dynamic aspects of ISs i.e. both the design of support tools (information retrieval support tools, recommender systems), the IS use and then the usage analysis (usage mining).
- the knowledge management point of view related to the capitalization of knowledge and experience in the evaluation process of IS: this concerns the expertise of combining the evaluation results according to different points of view and more generally the KDD[1] expertise applied on information systems data.

[1] KDD: Knowledge Discovery from Databases

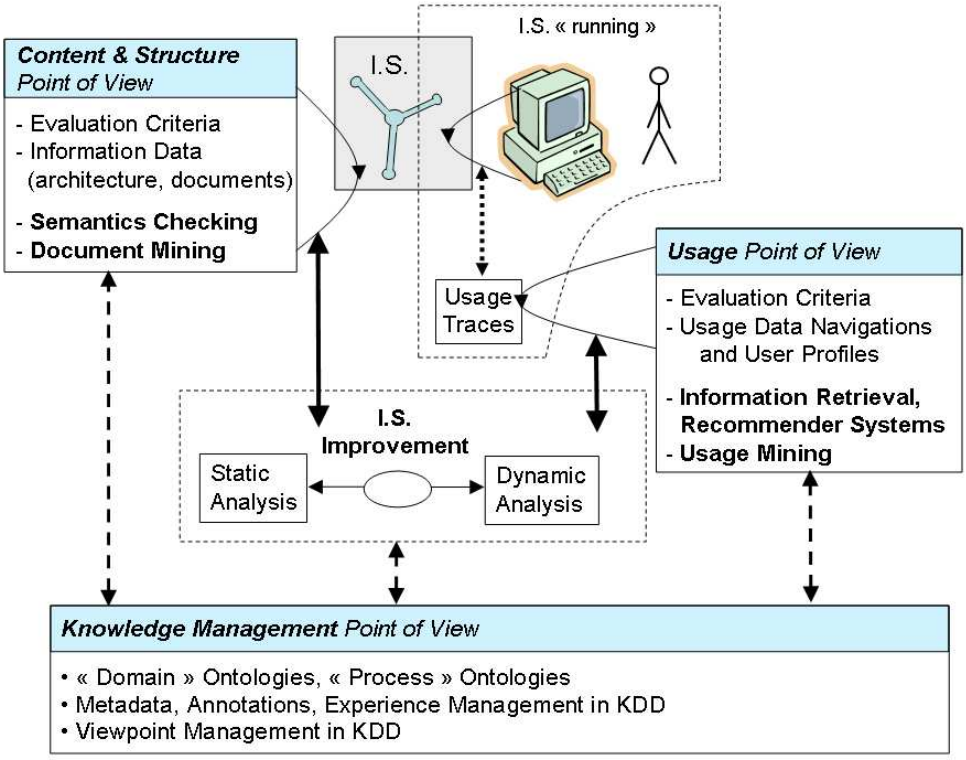

Figure 1. Global View of AxIS Research Topics

3. Scientific Foundations

3.1. Introduction

This section details the questions that we want to answer to:

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

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

3.2. Semantics and Design of Hypertext Information Systems

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

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

Most of the efforts deployed in the Web domain are related to languages for documents presentation (HTML, CSS, XSL) and structure (XML), to Web sites modeling and Web services (UML), but not to the formal semantics of Web sites to support their quality and evolution. The initiative led by the W3C consortium on Semantic Web (XML, RDF, RDF Schema) and ontologies aims at a different objective related to resource discovery. The term “semantics” has at least two significations:

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

To address the first definition of the word semantics, we use taggers, thesaurus, ontologies, to go deeper into the semantics of plain text.

But we are especially interested with the latter definition, trying to give a formal semantics to Web sites.

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

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

3.3. Information Systems Data Mining

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

3.3.1. Usage Mining

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

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

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

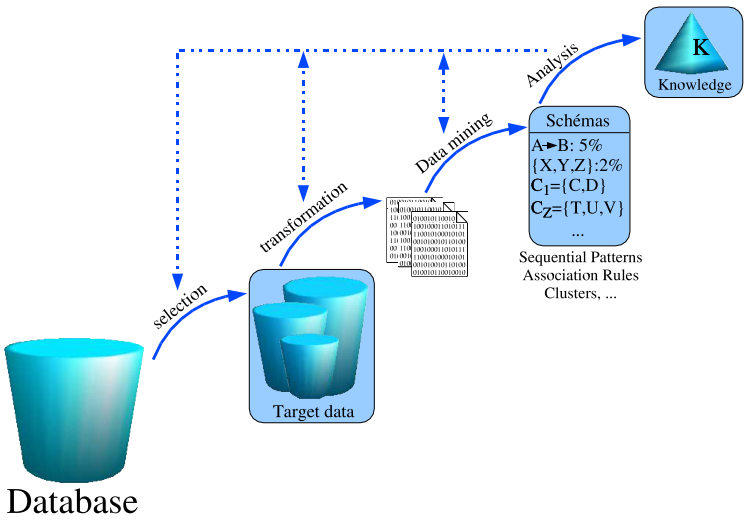

Figure 2. Steps of the KDD Process

Let us zoom on five following research topics involved in the first third steps:

3.3.1.1. Data selection and transformation

We insist on the importance of the pre-processing step in the KDD process composed of selection and transformation sub-steps.

The considered KDD methods applied on usage data will rely on the notion of user session, represented through a tabular model (items), an association rules model (itemsets) or a graph model. This notion of session enables us to act in the appropriate 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, AI., ...), a data transformation from input to output is needed and will be managed by the parsers. The input data will come from databases or from standard formatted file (XML) or a private format.

3.3.1.2. Data mining: extracting association rules

Our preprocessing tools (or generalization operators) given in the previous paragraph 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.

3.3.1.3. Data mining: discovering sequential patterns

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

- First of all by analyzing the causes of a sequential pattern extraction failure on large access logs. It is necessary to understand and incorporate the great variety of potential behaviours on a Web site. This variety is mainly due to the large size of the trees representing the Web sites and the very large number of combination of navigations on those sites.
- It is also necessary to incorporate all the available information related to the usage. Taking into account several information sources in a single sequential pattern extraction process is a challenge and can lead to numerous opportunities.
- Finally, sequential pattern mining methods will have to get adapted to a new and growing domain: data streams. In fact, in numerous practical cases, data cannot be stored more than a specified time (and even not at all). Data mining methods will have to provide solution in order to respect the specific constraints related to this domain (no multiple scan over the data, no blocking actions, etc.).

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

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

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

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

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

3.3.1.5. Data mining: reusing usage analysis experiences

This 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 an open research problem 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.

3.3.2. Content and Structure Document Mining

Keywords: *classification, clustering, document mining.*

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

The involved techniques from the KDD process are thus mainly clustering and classification. Our goal is to explore the possibilities of those techniques for document mining such as described below.

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

Recently much attention has been drawn towards using the structure of XML documents to improve information retrieval, classification and clustering, and more generally information mining. In the last four years, the INEX (Initiative for the Evaluation of XML retrieval) has focused on system performance in retrieving elements of documents rather than full documents and evaluated the benefits for end users. Other works are interested in clustering large collections of documents using representations of documents that involve both the structure and the content of documents, or the structure only ([68], [77], [63], [74]).

Approaches for combining structure and text range from adding a flat representation of the structure to the classical vector space model or combining different classifiers for different tags or media, to defining a more complex structured vector models [88], possibly involving attributes and links.

When using the structure only, the objective is generally to organize large and heterogeneous collections of documents into smaller collections (clusters) that can be stored and searched more effectively. Part of the objective is to identify substructures that characterize the documents in a cluster and to build a representative of the cluster [67], possibly a schema or a DTD.

Since XML documents are represented as trees, the problem of clustering XML documents is the same as clustering trees. One can identify two main approaches: 1) identify frequent common sub-patterns between trees and group together documents that share the same patterns; 2) define a similarity measure between trees that can be used with a standard clustering algorithm. A possible distance can be calculated by associating a cost function to the edit distance between two trees. However, it is well known that algorithms working on trees have complexity issues. Therefore some models replace the original trees by structural summaries or s-graphs that only retain the intrinsic structure of the tree: for example, reducing a list of elements to a single element, flattening recursive structures, etc.

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

3.4. Supporting Information Retrieval with adaptive recommender systems

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

We think that information retrieval support tools as recommender systems are very useful in very large information systems. The objective of a recommender system is to help system users to make their choices in a field where they have little information for sorting and evaluating the possible alternatives [83], [79], [71].

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

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

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.

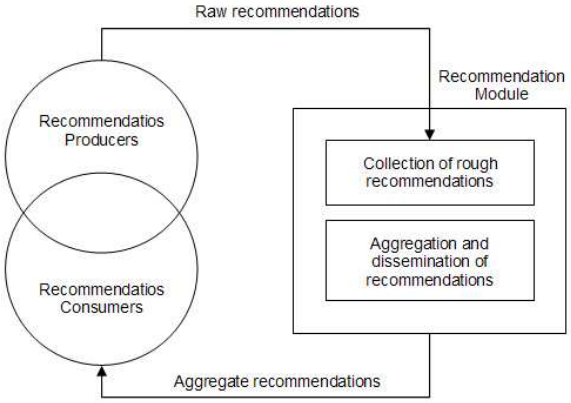

Figure 3. Architecture of a Recommender System

The user profile is a structure of data that describes user's centers of interest in the space of the objects which can be recommended. The user profile is a structure built in the first approach or specified by the user in the second approach.

The user profile is used either to filter available objects (content-based filtering), or to recommend to a user something that satisfied previous users with a similar profile (collaborative filtering) [79].

In the Axis project, we continue the development of a hybrid approach for recommendations based on the analysis of visited content and on collaborative filtering; The past behaviours of a user group are used to calculate the recommendations (collaborative filtering). Like this this approach is able to support some usage evolutions without a complete re-design.Also the usage analysis of such recommender systems may be very useful to support designers in an possible re-design or improvment of their IS.

Approaches based on data mining are mainly statistical approaches where the sequence of events in the history is not taken into account for the calculation of recommendations. There are some early examples in the field of navigation assistance on the Web: the FootPrints system [86] and the system of Yan et al. [87].

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

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

We study the class of recommender systems, based on the re-use of a user group's past experiences, using case based reasoning techniques (CBR).

Let us remind what is **Case-Based Reasoning (CBR)**. It is a problem solving paradigm based on the reuse by analogy of past experiences, called "cases". In order to be found, a case is generally indexed according to certain relevant and discriminating characteristics, called "indices"; these indices determine in which situation (or context) a case can be re-used.

Case-Based Reasoning [76] usually breaks up into four principal phases:

1. a "retrieve" phase for cases having similarities (i.e. similar indices) with the current problem,

2. a “re-use” phase where a solution to the current problem is built, based on cases identified in the previous phase,
3. a “revise” phase where the solution may be refined with an evaluation process,
4. a “retain” phase that updates the elements of the reasoning by taking into account the experiment which has been just carried out and which could thus be used for future reasoning.

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

We focus on two types of recommender systems:

- systems where the calculation of recommendations is based on the re-use of an users group’s experiences in searching for information in a 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 the design process.

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

4. Application Domains

4.1. Panorama overview

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

The project explores any applicative field on design, evaluation and improvement of a huge hypermedia information systems, for which end-users are of primary concern. We currently focus on web-based information systems (internet, intranet), or parts of such ISs, offering one of the following characteristics:

1. presence or wanted integration of services of assistance in the collaborative search of information and personalization (ranking, filtering, addition of links, etc.);
2. frequent evolution of the content (information, ontology), generating many maintenance problems, for example:

   – a web-based IS containing information about the activities of a group of people, for example an institute (Inria), a company, a scientific community, an European network on the internet or intranet, etc.

– a web-based IS indexing a wide range of productions (documents, products) resulting from the Web or a company, according to a thematic criteria, eg. 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 specification and checking of an information system,
- usage analysis of an information system (internet, intranet),
- document mining (XML documents, texts, Web pages)
- re-designing of an information system based on usage analysis,
- adaptive recommender systems for supporting information retrieval, Collaborative search of Information on the internet.

Ultimately, it should be noted that other fields (Life Science, health, transports, etc.) may be subject to the study since they provide an experimental framework for the validation of our research work in KDD, and in the reuse of experiences in story management: this type of approach may be relevant in applications that are badly solved in automatic of control type (e.g. nutrition of plants under greenhouses, control in robotics).

5. Software

5.1. Introduction

AxIS has developed several packages http://www-sop.inria.fr/axis/software.html:

- for classification and clustering : SODAS 2 Softawre (CF. section 5.2), Clustering Toolbox (cf. section 5.3),
- as well as frameworks for Case-Based Reasoning (cf. CBR*Tools section 5.4) and Recommender systems (cf. section 5.5).

5.2. SODAS 2 Software

Participants: Yves Lechevallier [correspondant], Marc Csernel.

The ASSO project designs methods, methodology and software tools for extracting knowledge from multidimensional complex data.

The SODAS 2 Software [78] is the result of the European project called “ASSO”(Analysis System of Symbolic Official data), that started in January 2001 for 36 monthsXS. It supports the analysis of multidimensional complex data (numerical and non numerical) coming from databases mainly in satistical offices and administration using Symbolic Data Analysis [57].

SODAS 2 is an improved version of the SODAS software developed in the previous SODAS project, following users’ requests. This new software is more operational and attractive. It proposes innovative methods and demonstrates that the underlying techniques meet the needs of statistical offices. It uses the SOM library [62].

SODAS allows for the analysis of summarised data, called Symbolic Data. This software is now in the registration process at APP. The latest executive version (version 2.50) of the SODAS 2 software, with its user manual (PDF format), can be downloaded at http://www.info.fundp.ac.be/asso/sodaslink.htm

5.3. Clustering Toolbox and Classification Software

Participants: Marc Csernel, Sergiu Chelcea, Francesco de Carvalho, Aicha El Golli, Brieuc Conan-Guez, Mihai Jurca, Yves Lechevallier [co-correspondant], Brigitte Trousse [co-correspondant].

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

- from AxIS Rocquencourt: 1) a partitionning clustering method on complex data tables called SCluster [78], 2) an adapted version of the SOM on the dissimilarity tables called DSOM [65](cf. section 6.3.1) and Div (in C++) [3];
- two partitionning clustering methods on the dissimilarity tables: 1) CDis (in C++) [78] issued from a collaboration between AxIS Rocquencourt team and Recife University, Brazil and 2) CCClust (in C++) issued from a collaboration between AxIS Rocquencourt team and Recife University, Brazil;
- 2-3 AHC (in Java)[4] (cf. section 6.3.4) from AxIS Sophia Antipolis.

We developed a Web interface for this clustering toolbox for the following methods: SCluster, Div, Cdis, CCClust. 2-3 AHC is available as a Java applet which runs the HIERARCHIES VISUALISATION TOOLBOX. The aim of this online interface is in a short term to allow other team members (and in the near future Internet users) to use these classification methods to process their own data via the Web. The Web interface is developed in C++, run on our Apache internal Web server.

For classification of functional data, we developed a functional Multi-Layer Perceptron Method called FNET (cf. section 6.3.2).

5.4. CBR*Tools

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

CBR*Tools is an object-oriented framework [70], [66] for Case-Based Reasoning which is specified with the UMT notation (Rational Rose) and written in Java. It offers a set of abstract classes to models the main concepts necessary to develop applications integrating case-based reasoning techniques: case, case base, index, measurements of similarity, reasoning control. It also offers a set of concrete classes which implements many traditional methods (closest neighbors indexing, Kd-tree indexing [85], prototypes indexing [69], neuronal approach based indexing, standards similarities measurements). CBR*Tools currently contains more than 240 classes divided in two main categories: the core package for basic functionality and the time package for the specific management of the behavioral situations. The programming of a new application is done by specialization of existing classes, objects aggregation or by using the parameters of the existing classes.

CBR*Tools aims application fields where the re-use of cases indexed by behavioral situations is required. The CBR*Tools framework was evaluated via the design and the implementation of five applications (Broadway-Web, educaid, BeCKB, Broadway-Predict, CASA and RA2001). We showed that, for each application, the thorough expertise necessary to use CBR*Tools relates to only 20% to 40% of the hot spots thus validating the assistance brought by our platform on design as well as on the implementation, thanks to the re-use of its abstract architecture and its components (index, similarity).

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

CBR*Tools is planned to be available in 2006 for research, teaching and academic purpose under the INRIA license. The user manual can be downloaded at the URL: http://www-sop.inria.fr/axis/cbrtools/manual/.

5.5. Broadway*Tools

Participants: Mihai Jurca, Brigitte Trousse [correspondant].

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

Different recommender systems have been developed:

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

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

6. New Results

6.1. Introduction

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

This year we obtained original results as previous years in our three research topics: data transformation and knowledge management, data mining and web usage mining methods. Let us note first results in document mining (cf. the content and structure point of view of an IS) and in data stream mining (cf. the usage point of view of an IS).

In section 6.2, we describe new results on data transformation and knowledge representation. For the latter, we have studied the use of metadata (cf. the KM point of view), in particular in two ongoing PhD thesis related semantic web and KDD, conducted by H. Behja and A. Baldé. Metadata have been used or annotating global KDD processes in terms of viewpoints to support the management and the reuse of past KDD experiences (cf. section 6.2.6), 2) for supporting the interpretation of extracted clusters. Moreover this year we have proposed and studied new distances and dissimilarities in various applicative contexts: XML Sanskrit documents (section 6.2.2), tourist itineraries (section 6.2.3) and Web navigations and content (cf section 6.2.1).

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

Finally on information systems data mining, we started this year to work on visualization problems and we proposed different representations of the organization of a web site based on usage data (cf. section 6.4.1). We also obtained our first results on XML document mining and XML search: we studied content and/or structure mining for clustering or classifying XML documents (cf. sections 6.5.1, 6.5.2) as well as the improvement of the relevance in XML search (cf. section 6.5.3). More classically we pursued our researches on intersites web usage mining in the context of the ECML/PKDD 2005 discovery Challenge (cf. section 6.4.2) and in extracting dense periods of sequential patterns (cf. section 6.4.3).

6.2. Data Transformation and Knowledge Management in KDD

6.2.1. Dissimilarities for Web Usage Mining

Keywords: *Benchmark, Clustering, Dissimilarities, Validation, Web Usage Mining.*

Participants: Fabrice Rossi, Francisco De Carvalho, Yves Lechevallier, Alzennyr Da Silva.

Many Web Usage Mining methods rely on clustering algorithms in order to produce homogeneous classes of documents (when the content of the web site is analyzed) and/or of users (when browsing behaviors are analysed). Information extracted from web server logs are complex and noisy, but can be used to define usage based dissimilarities between users of the site or between pages of the site. There are however many possibilities to define this type of dissimilarity measure.

We have defined a benchmark site that allows to compare dissimilarity measures via the clustering results they produce. The benchmark consists in one year of log of the web site of the CIn, the laboratory of Francisco De Carvalho. This site is small (91 pages) and is very well organized. This allows to define a meaningful semantic structure and to build an expert partition of its content. Then, this expert partition can be compared to the results of a clustering algorithm applied to the dissimilarity matrix constructed with a specific measure.

The results that will be published in the EGC 2006 conference show that the Jaccard index and the "term frequency inverse document frequency" approach obtain quite good results, whereas the cosine measure performs badly. It seems also that better results could be obtained by taking into account the structure of the site together with usage data.

6.2.2. Distances for Clustering Homogeneous XML Documents

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

Participants: Marc Csernel, Sergiu Chelcea, Yves Lechevallier, Sattisvar Tandabany, Brigitte Trousse.

In the context of a research project with India and some others French partners, about a hundred ancient manuscripts written in Sanskrit, all arisen from the same text (the Benares Glose), should be compared in order to make a critical edition and provide some classification between the manuscripts (cf. section 6.2.7).

During his internship, S. Tandabany developed some tools that were required for clustering homogeneous XML Sanskrit Documents. First, a modified longest common substring algorithm is proposed to deal with Sanskrit characters. Then, as in Sanskrit inversions of characters are not always meaningful, a detection of possible inversions is applied. Finally, the Agglomerative 2-3 Hierarchical classification (cf. section 6.3.4) is used as the classification algorithm. To do this, we proposed a new distance between texts taking into account some Sanskrit specificities and allowing the addition of meta-data (worn state and shape of the manuscripts as objects, annotations about words forgotten, etc.). The text is splitted into paragraphs, "sub-distances" are computed between each corresponding paragraph, taking into account adds, deletions, transformations and inversions. Then, some constraints needed to obtain a distance (triangular inequality) are removed to get a dissimilarity instead of a distance for the 2-3AHC. The impact of these modifications on the classification was analysed. Finally, our results are highlighted with some experiments and examples [55].

6.2.3. Distances for Clustering Downtown Tourist Itineraries

Keywords: *2-3 AHC*, *Clustering*, *Distance*, *Tourist itineraries*.

Participants: Rémi Busseuil, Sergiu Chelcea, Brigitte Trousse.

In the context of the MobiVIP project (cf. section 7.1.2), during Rémi Busseuil's internship we studied the possibility of clustering tourists itineraries in the town center (Antibes in this case). Thus, a software for tourist itineraries generation and clustering was developed (in Visual .NET), taking into account not only the geographical characteristics of an itinerary, but also the symbolical ones: street type, buildings type, etc. This new use of semantical data, opens new directions for the road itineraries recommendations, by addressing new issues like the purpose of the itinerary or the nature of the crossed areas.

Clustering itineraries has many advantages besides the possibility of choosing the most suitable one: it is also an analysis and comparison tool. This can have multiple applications: route or destination prediction, traffic anticipation, etc. As clustering algorithm, we used the Agglomerative 2-3 Hierarchical Classification (2-3 AHC) algorithm. The 2-3 AHC has the advantage of being easily visualized compared to a classical clustering method.

In order to compare different itineraries, we basically divided each itinerary into fragments and then we computed a distance/dissimilarity value using the Longest Common Subsequence algorithm (LCS) and a spread function developed in [55]. Different ways of defining the dissimilarity and of comparing the fragment were tested.

An example of clustering 40 itineraries issued from 4 different profession types is presented in Figure 4 bellow.

6.2.4. Semantics Tools for XML documents

Participant: Thierry Despeyroux.

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

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

- Web sites may be spread along a great number of files.

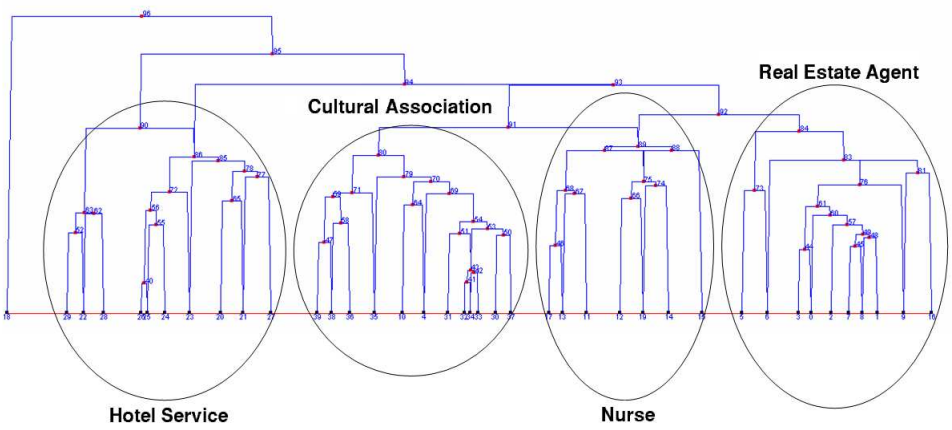

Figure 4. 2-3 AHC classification on 40 itineraries

- Information is scattered, with many forward references.
- We may need to use external resources to define the static semantics (thesaurus, ontologies, taggers, image analysis program, etc.).

We are developing a specification language to express global constraints in Web sites or in a collection of XML documents in an operational way.

An initial version of this language as been described in [6], together with its application to a real sized collection of documents: the Inria scientific activity report for the years 2001 and 2001.

The language and its implementation has been developped and improved in 2005, in particular for efficiency. At the same time our XML core parser has been extended to allow parsing of HXTML documents.

The same language has been used to extract information in XML documents. This has been the case to choose and extract words from different part of XML documents. These words was first passed to a tagger, then used to cluster the different documents. A first experiment has been done in 2004 and has been presented to the EG2005 conference [33], [32] (cf. section 6.5.1).

In a more long term experiment, we have initiated a regular monitoring of the Inria activity reports to see how the number of bad URLs in these reports evoluates. This monitoring, started in december 2004 and then performed every two weeks, takes into account the activity report for 2002, 2003 and 2004.

6.2.5. Metadata Extraction for Supporting the Interpretation of Clusters

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

Participants: Abdourahamane Baldé, Yves Lechevallier, Brigitte Trousse.

This work was conducted in the context of the PhD of A. Baldé.

A huge volume of data is produced by many applications. Data mining techniques are part of knowledge discovery methods whose aim is to discover knowledge in large databases without predetermined information about the application field which is well-known as KDD. But data mining is a complex process for an end-user and the main difficulties consist in the interpretation of the results. Metadata can help the interpretation process by providing additional information. Our objective is to facilitate the interpretation process and to point out that metadata can play a major role for this purpose. In spite of the visual representation of the results, the user should acquire a significant experience to be able to interpret the clusters. Data mining tools generally offer visualization modules which are not adapted to analysis. The original contributions of our work made

in collaboration with Marie-Aude Aufaure (Supelec) concern new approaches to representing clustering's metadata and interpreting clustering's results by using metadata.

First, we propose a metadata model that could be automatically exploited [17]. We also propose a tool in order to help the end-user to interpret the clusters obtained. This tool is based upon the architecture described in Figure 5.

This architecture is composed by three layers: metadata model, metadata manager which manages metadata extraction and storage and manipulations performed on these metadata and user query layer using XQuery. In order to implement these queries, we use the Saxon processor. Saxon is a set of tools dedicated to XML documents processing: it has established a reputation for fast performance, the highest level of conformance to the W3C specifications. This method can be applied to a wide variety of data mining methods.

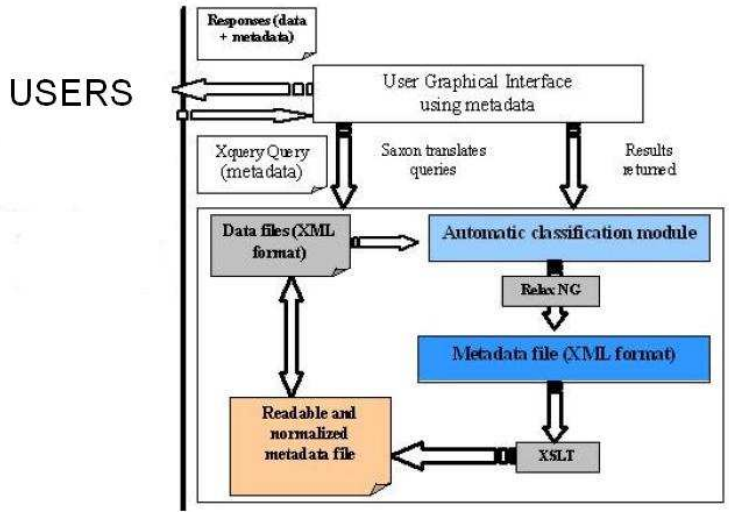

Figure 5. metadata production architecture

6.2.6. Viewpoint Managment for Annotating a KDD Process

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

Participants: Hicham Behja, Brigitte Trousse.

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

Our goal is to make more explicit the notion of "viewpoint" from analysts during their activity and to propose a new approach to integrate the viewpoint notion in a multi-view Knowledge Discovery from Databases (KDD) analysis. We define a viewpoint in KDD as an analyst's perception of a KDD process, perception referring to its own knowledge [12]. Our purpose is to facilitate both reusability and adaptability of a KDD process, and to reduce his complexity whilst maintaining the trace of the past analysis viewpoints. The KDD process will be considered as a view generation and transformation process annotated by metadata to store the semantics of a KDD process.

In 2004 we started with an analysis of the state of the art and identified three directions: 1) the use of the viewpoint notion in the Knowledge Engineering Community including object languages for knowledge representation, 2) modelling KDD process adopting a Semantic Web based approach and 3) the use of annotations of KDD processes. Then we designed and implemented an object oriented platform for KDD

processes including the viewpoint notion (via design patterns and UML using Rational Rose). The current platform is based on the Weka library.

In 2005 we proposed and implemented the knowledge conceptual model [25] integrating the viewpoint concept (cf. Figure 6). It is composed of four models structured in two types of knowledge:

First, for the domain knowledge, the domain model that describes the analyzed domain knowledge in terms of objects, attributes, data, etc. and the analyst domain knowledge that will relate to the tasks carried out by the analyst; choice of methods, variables, etc. We propose a formal representation of the domain model as a datawarehouse that allows the business information to be viewed from many viewpoints. For our example of the HTTP log in Web Usage Mining (WUM), the used database design is the star schema.

Second, for the strategic knowledge, we find:

- the task and method model which describes the KDD analyst domain knowledge. Here, the domain objects are methods, algorithms, parameters, etc. This model is a semi-formal generic ontology. For its construction we are mainly inspired from the DAMON system ontology for the data mining step, but we address all three KDD steps (preprocessing, data mining and postprocessing). This ontology is developed in Protégé-2000 system.
- the viewpoint model which describes the viewpoint specification in terms of preferences related to the decision-making process in KDD (choice of the attributes, methods and systems, etc.). This viewpoint model, described by a RDF scheme, manipulates both the analyzed domain and the analyst domain:

  – The viewpoint analyzed domain specifies the significant attributes for the expert from the analysed domain. This vision allows the analyst on the one hand to restrict the analyzed domain and on the other hand to guide the goal of the retrieval by defining a diagram on the raw data.

  – The viewpoint analyst domain allows to define a symbolic execution by choosing the methods and the algorithms for each KDD process step.

  – The viewpoint organizational model describes the organization of the viewpoints in terms of relations among them (in progress).

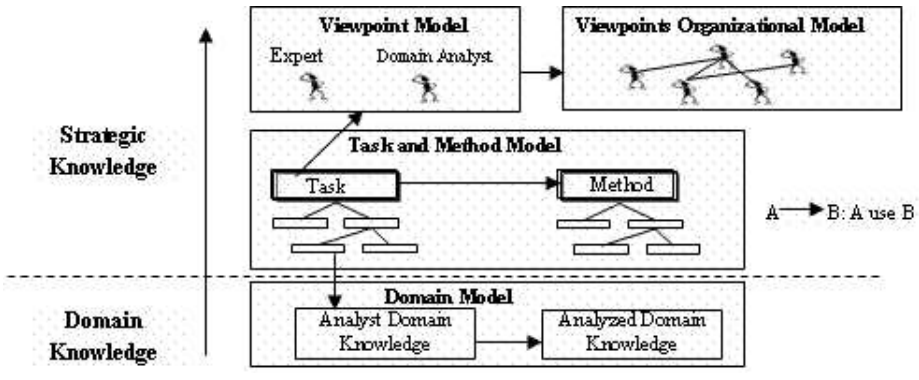

Figure 6. Conceptual model

This work is accepted for publication in january 2006 in a special issue on “Méthodes Avancées de Développement des SI” of the french journal ISI (D. Rieu and G. Giraudin editors).

6.2.7. Production and Display of a Critical Edition of Sanskrit documents

Keywords: *Critical Edition, Sanskrit, Text comparison, Unicode, XML, electronic display, transliteration.*

Participants: Marc Csernel, Marina Dufresne, Yves Lechevallier, Selma Khebache.

A critical edition is the edition of a well known text taking into account all possible versions of this text. Critical editions are particularly needed for texts issued from manuscripts where the variations can be very significant from one manuscript to the other.

Production of critical edition of Sanskrit Text. This is particularly important for the Indian subcontinent where at least one third of the manuscripts existing through the whole world are supposed to be found, the main part of them being written in Sanskrit. It was not an Indian tradition to deal with critical edition, so very few of them exist at present time. The idea is to provide a computer assisted construction of critical edition. Such tools exist for occidental languages, but, due to some Sanskrit specificities they could not suit our purpose:

- Sanskrit is written according to a transliterated alphabet.
- Separation between words are mostly absent in a manuscript and their presence is not meaningful. The text is generally formed by a sequence of thousands of characters;
- The writing of two words is different if there is a blank (or any separation) between them, or if they are following each other directly. This notion is called a sandhi.

In order to avoid the complexity problem induced by the thousands of characters sentences, and to be able to provide to the philologist the exact words where a difference occurs, we need either a lexicon, or a text where all the words appear separately. We will use the second solution and we call such a text a Padapatha according to a certain form of recitation used by the Sanskritist. Due to the sandhi, such a text is not readily comparable with the manuscript text. We must provide a pre-processing based on LEX to construct all the sandhi related to the padapatha in order to be able to provide a suitable comparison. For the comparison we use the Longest Common Subsequence (LCS) algorithm based on dynamic programming. The algorithm output will be used as input of the following project steps :

- Electronic display of critical edition of Sanskrit text
- Cluster (cf. section 6.2.2) and Phylogenetic trees

Electronic display of critical edition of Sanskrit Text. Traditionally critical editions are represented as particularly boring (for unfamiliar) books, where the text itself is very small, and where the notes are numerous and enormous. It's a philologist dream to see only the points they care about. An electronic form of critical edition could be a proper answer.

But there are a lot of problems related to the Sanskrit we have to care about. Thanks to Unicode it is now possible to get some standard about the display of sandhi characters. But because of the ligature formation, two Sanskrit characters separated by a blank do not look similar as if they were put one after the other.

Marina Dufresne during her internship (cf. section 9.2.3) developed a software tool what allows an interactive display of critical edition of Sankrit text starting from an XML text. This tool is not perfect but it has been greatly appreciated by the Sanskrit community. This work has bee done in collaboration with François Patte.

6.3. Data Mining Methods

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

6.3.1. Self Organizing Maps on dissimilarity matrices

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

Participants: Aicha El Golli, Yves Lechevallier, Fabrice Rossi, Nicomedes Lopes Calvacanti Junior.

The standard Self Organizing Map (SOM) is restricted to vector data from R^n. In our previous work [64], [65], we proposed an adapted version of the SOM, called the DSOM (for Dissimilarity SOM) that can be applied to any data for which a dissimilarity can be defined.

In 2005, we improved DSOM by defining a new algorithm associated to an improved implementation. This implementation significantly reduces the execution time of the method without changing the results [29]. The new algorithm is based on a factorization technique applied to the computation of the criterion optimized by the method (the sum of weighted dissimilarities between a prototype candidate and the data to cluster). It is associated to an early stopping scheme and to some memorization techniques that leverage the iterative nature of the method.

We have also applied the DSOM to usage-based web content clustering and visualization [42], see section 6.4.1 for details.

6.3.2. Functional Data Analysis

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

Participant: Fabrice Rossi.

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

In 2005, we have extended our approach based on neural methods for functional data to the case of Support Vector Machines (SVMs) applied to function classification:

- in [44], we have introduced functional kernels based on derivation operators and on B-spline smoothing. An application to spectrometric curves classification showed an improvement of these kernels over standard non functional kernels;
- in [46], we have studied the theoretical properties of a projection based functional kernel, in which functions are projected to a truncated Hilbert basis in a pre-processing step. The coordinates on this basis are handled by a standard SVM. We showed that this method, associated to a split sample procedure for the choice of the truncation level, is consistent (i.e., it can reach the Bayes error rate asymptotically). We also illustrated the method on several real world data set (speech recognition problems).

In 2005, our earlier work on functional multi-layer perceptrons has been published in international journals [21], [22], [23].

6.3.3. Partitioning Method: Adaptive Distances on Interval Data

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

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

The main contribution [24] is the proposal of a new partitional dynamic clustering method for interval data based on the use of an adaptive Hausdorff distance at each iteration. The idea of dynamical clustering with adaptive distances is to associate a distance to each cluster, which is defined according to its intra-class structure. The advantage of this approach is that the clustering algorithm recognizes different shapes and sizes of clusters. Here the adaptive distance is a weighted sum of Hausdorff distances. Explicit formulas for the optimum class prototype, as well as for the weights of the adaptive distances, are found. When used for dynamic clustering of interval data, these prototypes and weights ensure that the clustering criterion decreases at each iteration.

Let Ω be a set of n objects indexed by i and described by p interval variables indexed by j. An *interval variable* X [57] is a correspondence defined from Ω in $\Re$ such that for each $i \in \Omega$, $X(i) = [a, b] \in \Im$, where $\Im$ is the set of closed intervals defined from $\Re$, i.e., $\Im = \{[a, b] : a, b \in \Re,\ a \leq b\}$. Each object i is represented as a vector of intervals $\mathbf{x}_i = (x_i^1, \cdots, x_i^p)$, where $x_i^j = [a_i^j, b_i^j] \in \Im$.

An interval data table $\{x_i^j\}_{n \times p}$ which is used by our clustering method is made up of n rows that represent n objects to be clustered and p columns that represent p interval variables. Each cell of this table contains an interval $x_i^j = [a_i^j, b_i^j] \in \Im$. In our approach [24] a prototype $\mathbf{y}_k$ of cluster $C_k \in P$ is also represented as a vector of intervals $\mathbf{y}_k = (y_k^1, \cdots, y_k^p)$, where $y_k^j = [\alpha_k^j, \beta_k^j] \in \Im$.

It is now a matter of choosing an adaptive distance between vectors of intervals and properly defining the representation step of the dynamic algorithm with adaptive distances given in the previous section. In other words, we will give an explicit formula for the prototype $\mathbf{y}_k$ and for the vector of weights λ_k that minimizes both the adequacy criterion $\sum_{j=1}^{p} \lambda_k^j \sum_{i \in C_k} d(x_i^j, y_k^j)$.

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

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

Participants: Sergiu Chelcea, Mihai Jurca, Brigitte Trousse.

This work was conducted in the context of the PhD of S. Chelcea.

We have continued [28] this year our study of the Agglomerative 2-3 Hierarchical Clustering [60], [59] as a part of Chelcea Sergiu's PhD thesis. A study of different aggregation indexes and cluster indexing measures combined with the 2-3 AHC algorithm execution has revealed a particular case of clusters merging, which can influence the resulting induced dissimilarity matrix. This case that we denoted *blind merging*, is present when two clusters are merged whilst one of them is not maximal. Based on our previous theoretical study [58], the next (intermediate) merging, will merge together two clusters possibly at a high indexing degree. This can be avoided by minimizing the final cluster's indexing degree when choosing the two merging clusters.

A slightly modified version of a 2-3 AHC algorithm was proposed and implemented in order to avoid such situations. The interest of this new 2-3 algorithm variant is its resulting induced dissimilarity matrix which is "better" or equal to the classic ultrametric.

We experimentally validated this new 2-3 AHC algorithm variant on different artificial datasets and we also integrated it into our HIERARCHIES VISUALIZATION TOOLBOX[2].

This new 2-3 AHC algorithm variant was also applied and validated on other types of datasets: on Web Usage Data [28], on Sanskrit XML documents [55] (see also Section 6.2.2) and on tourists itineraries [47].

6.3.5. Sequential Pattern Extraction in Data Streams

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

Participants: Alice Marascu, Florent Masséglia.

This work was conducted in the context of the master of A. Marascu.

[2] http://axis.inria.fr/

In recent years, emerging applications introduced new constraints for data mining methods. These constraints are particularly linked to new kinds of data that can be considered as complex data. One typical kind of such data is known as *data streams*. In a data stream processing, memory usage is restricted, new elements are generated continuously and have to be considered as fast as possible, no blocking operator can be performed and the data can be examined only once.

At this time and to the best of our knowledge, no method has been proposed for mining sequential patterns in data streams. We argue that the main reason is the combinatory phenomenon related to sequential pattern mining. Actually, if itemset mining relies on a finite set of possible results (the set of combinations between items recorded in the data) this is not the case for sequential patterns where the set of results is infinite. In fact, due to the temporal aspect of sequential patterns, an item can be repeated without limitation leading to an infinite number of potential frequent sequences.

The SMDS (Sequence Mining in Data Streams) method, proposed in [37], [36], is designed for extracting sequential patterns from a data stream. More precisely, our goal is to extract significant patterns that will be representative of Web usage streaming data. To this end, SMDS performs as follows:

1. cutting down the data stream into batches of fixed size. The following operations are then performed for each batch;
2. clustering the sequences of the batch;
3. for each cluster c, providing the alignment of the sequences embedded in c. The aligned sequence will be considered as a summary of c;
4. filtering the aligned sequence in order to keep 1) frequent items only and 2) aligned sequences obtained on clusters having size greater than 2 only;
5. maintaining a prefix tree structure that will keep the history of frequency for each extracted sequence (the operations on this structure may be $insertion$, $update$ or $deletion$).

All those steps have to be performed as fast as possible in order to meet the constraints of a data stream environment. Approximation has been recognized as a key feature for this kind of applications, explaining our choice for an alignment method for extracting the summaries of clusters. The SMDS method is illustrated in figure 7. SMDS has been tested over both real and synthetic datasets. Experiments could show the efficiency of our approach and the relevance of the extracted patterns on the Web site of Inria Sophia Antipolis.

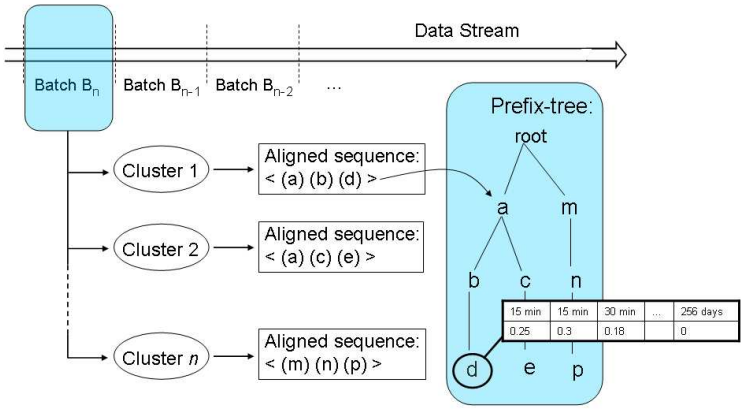

Figure 7. The SMDS method

6.4. Web Usage Mining Methods

6.4.1. Visualization

Keywords: *data visualization, dissimilarities, graph visualization, non linear projection, self organizing map, web usage mining.*

Participants: Fabrice Rossi, Yves Lechevallier, Aicha El Golli.

The analysis of the content of a web site based on usage data is an important task as it allows to obtain insight on the organization of the site and of its adequacy to user needs. The (dis)agreement between the prior structure of the site (in terms of hyperlinks) and the actual trajectories of the users is of particular interest. In many situations, users have to follow some complex paths in the site in order to reach the pages they are looking for, mainly because they are interested in topics that appeared unrelated to the creators of the site and thus remained unlinked. On the contrary, some hyperlinks are not used frequently, for instance because they link documents that are accessed by different user groups.

In 2005 we have studied two general tools for visualizing the content of a site based on usage data:

- in [43] we have used the logical and hierarchical organization of the web site to simplify the representation of user trajectories. Simplified trajectories are used to calculate dissimilarities between URL groups defined thanks to the site hierarchy (groups are also called topics in section 6.4.2). The groups, which reflect the prior semantic structure of the site, are represented thanks to the minimum spanning tree induced by the dissimilarity matrix. This allows to explore the relationship between prior categories and user browsing patterns. The method was applied to the INRIA web site and gave satisfactory results;
- in [42] we have applied the same general methodology for calculating dissimilarities between URL groups but we used an adapted version of the Self Organizing Map (SOM) to visualization clusters obtained via the dissimilarity matrix (see section 6.3.1 for details on this version of the SOM).

6.4.2. InterSites Web Usage Mining: preprocessing methodology and crossed clustering

Keywords: *Complex data, Crossed-clustering, Preprocessing, Web Usage Mining.*

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

In the context of the ECML/PKDD 2005 Discovery Challenge, we improved [26], [27] our preprocessing methodology for intersites Web Usage Mining [16]. A clickstream dataset was proposed in the Discovery Challenge this year for the first time. The dataset consisted in requests for page views on seven different e-commerce Web sites from the Czech Republic. A request contained a PHP SessionID automatically generated for each new user visit on each server (unique IDs).

Based on Tanasa's preprocessing methodology [16], we defined a new methodology to preprocess the provided datasets and to store it in a data warehouse. Since a user changing shops can have (during a single visit) multiple SessionIDs, one on each shop, we regrouped these PHP SessionIDs into intersite users visits. More precisely we regrouped SessionIDs belonging to a single user (same IP) into a *Group of SessionIDs*, corresponding to the user's actual (intersite) visit. This was done by comparing the Referrer with the previously accessed URLs (in a reasonable time window), each time the user moves to another shop. We thus reduced by 23.88% the number of user visits.

To analyze the traffic load in the seven shop sites, we grouped the requests in terms of *Time Periods* (slices of date and hour). We cross-clustered these time periods against the visited products using a generalized dynamic algorithm [72].

The result consisted in the confusion table containing classes of periods and products (see Figure 8).

Such analyses allow us to identify best hours for marketing strategies, like fast promotions, online advices and publish banners, etc. Others analyses could be planned in the future, exploiting for example the link between the consumer activities and the time periods by shop or focusing on multi-shop user visits, etc.

	Product_1	Product_2	Product_3	Product_4	Product_5	Total
Period_1	2847	5084	3284	2265	2471	15951
Period_2	11305	31492	12951	1895	9610	67253
Period_3	33107	55652	36699	5345	20370	151173
Period_4	22682	46322	30200	5165	27659	132028
Period_5	9576	20477	19721	2339	7551	59664
Period_6	1783	3515	2549	392	*11240*	**19479**
Period_7	15019	14297	8608	1397	6014	45335
Total	96319	176839	114012	18798	**84915**	490883

Figure 8. Confusion table for 7 classes of periods and 5 classes of products

In fact we apply a previous work published in 2003: indeed we proposed in [72] a crossed clustering algorithm in order to partition a set of objects in a predefined number of classes and to determine, in the same time, a structure (taxonomy) on the categories of the object descriptors. This procedure is a simultaneous clustering algorithm on contingency tables. The convergence of the algorithm is guaranteed at the best partitions of the objects in r classes and of the categories of the descriptors in c groups, respectively. This algorithm extended the dynamical algorithms hereafter proposed in the context of the Web Usage Mining. In particular, we had already performed it on the Web Logs Data, coming from the HTTP log files of INRIA web server [73].

6.4.3. Extracting Dense Periods of Sequential Patterns

Keywords: *period, sequential patterns, web usage mining.*

Participant: Florent Masséglia.

This work has been done in collaboration with the LGI2P and the LIRMM (see 8.2.5) and has been published in [38]. Existing Web Usage Mining techniques are currently based on an arbitrary division of the data (*e.g.* “one log per month”) or guided by presumed results (*e.g* “what is the customers behaviour for the period of Christmas purchases?”). Those approaches have two main drawbacks. First, they depend on this arbitrary organization of the data. Second, they cannot automatically extract “seasons peaks” among the stored data.

The work presented in this section performs a specific data mining process (and particularly to extract frequent behaviours) in order to automatically discover the densest periods. Our method extracts, among the whole set of possible combinations, the frequent sequential patterns related to the extracted periods. A period will be considered as dense if it contains at least one frequent sequential pattern for the set of users connected to the Web site in that period.

Our method is based on:

1. a new representation of the Web log file designed to retrieve the “login” and “logout” information associated to each user.
2. a rewriting of the log in order to build periods based on the information of step 1. A period will begin at the arrival of a new user or end at the departure of a “connected” user.
3. a heuristic designed for extracting approximate frequent sequences from each period built at step 2.

The third step is based on PERIO, the heuristic we have developed for that purpose and it is widely inspired from genetic algorithms.

6.5. XML Document Mining and XML Search

6.5.1. Structure and Content Mining

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

Participants: Thierry Despeyroux, Mounir Fegas, Saba Gul, Yves Lechevallier, Anne-Marie Vercoustre.

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

Last year we studied the impact of selecting different parts (sub-structures) of XML documents for specific clustering tasks. Our approach integrated techniques for extracting representative words from documents elements with unsupervised classification of documents. We illustrated and evaluated this approach with the collection of XML activity reports written by Inria research teams for year 2003. The objective was to cluster projects into larger groups (Themes), based on the keywords or different chapters of these activity reports. We then compared the results of clustering using different feature selections, with the official theme structure used by Inria between 1985 and 2003, and with the new one proposed officially in 2004. The results (published this year) show that the quality of clustering strongly depends on the selected document features [33], [32].

This year we developed a new representation model for clustering XML documents. The standard vector model for classification or clustering of documents represents documents by weighted vectors of words contained in the documents. This model takes into account only the textual content of documents. With XML documents, we want a representation that takes into account either the structure of the documents or both the structure and the content. Since XML documents can be seen as trees, we represent documents by the set of their (node) paths of length L, $n <= \mathrm{L} <= m$, n and m being two given values. Paths can be contrained to be root-beginning paths, or leaf-ending paths. For dealing with both the structure and the content, we define *text paths* that extend the node paths with the word contained in the subtree of their final node. Then by regarding paths as words, we can cluster documents by applying standards clustering methods based on the vector model. There is one difficulty, though, since the vector model is based on the independance between the dimensions of the vectors. In our case, when two paths are embedded in each other they are obviously not independant. To deal with this problem of dependency, we partition the paths by their length and treat each set of paths as a different modality in the clustering algorithm.

We evaluate our approach using four standard metrics, namely the F-measure, the Corrected-Rand, the entropy and the purity. That for a given clustering task, we compare the resulting clusters with a priori known classes. We made several experiments using the INEX IEEE collections and INRIA activity reports [48]. The results that will be published in the EGC 2006 conference show that our approach works both for structure-based clustering and Structure-and-content clustering. However, using leaf-ending paths may result in damaging the clustering time, as the number of paths increases dramatically. We need to find good ways to reduce the number of paths, especially for text paths.

We also started to apply this approach to the collections proposed by the INEX XML Document Mining tracks [56].

6.5.2. Sequential Pattern Mining for Structure-based XML Document Classification

Keywords: *XML document, classification, sequential pattern, structure mining.*

Participants: Calin Garboni, Florent Masséglia, Brigitte Trousse.

The goal of this work is to provide a classification (“classification supervisée” in french) over a collection of XML documents. For this purpose we consider that we are provided with a set of clusters coming from a previous clustering on an past collection. More formally: let us consider S_1 a first collection of XML documents and $C = \{c_1, c_2, ...c_n\}$ the set of clusters defined for the documents of S_1. Let us now consider S_2 a new collection of XML documents. Our goal is to provide a classification on S_2 by taking into account the distribution of documents in C.

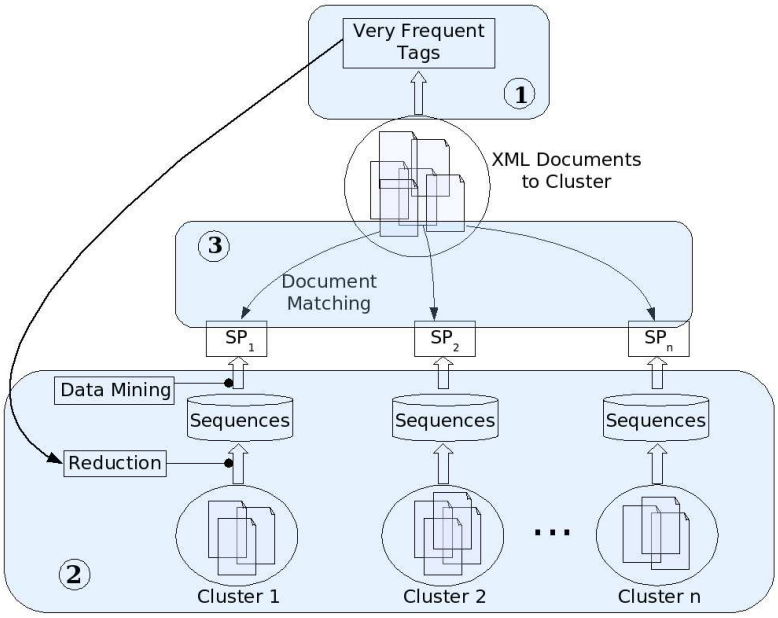

Figure 9. Overview of our structure-based classification method

To this end, our method will perform as illustrated in figure 9. It is based on the following three steps:

1. Pre-processing: first of all, we extract the frequent tags embedded in the collection. This step corresponds to step "1" in figure 9. The main idea is to remove irrelevant tags for clustering operations. A tag which is very frequent in the whole collection may be considered as irrelevant since it will not help in separating a document from another (the tag is not discriminative).
2. Characterising existing clusters: then we perform a data mining step on each cluster from the previous collection (namely "C" in the foreword of this section). This step corresponds to step "2" in figure 9. For each cluster, the goal is to transform each XML document into a sequence. Furthermore, during the mapping operation, the frequent tags extracted from step 1 are removed. Then on each set of sequences corresponding to the original clusters, we perform a data mining step intended to extract the sequential patterns. For each cluster C_i we are thus provided with SP_i the set of frequent sequences that characterizes C_i.
3. XML Document matching: finally the key step of our method relies on a matching between each document of the collection and the sequences extracted from the second step. This last step corresponds to step 3 in figure 9.

The matching techniques developed in this work are described in [49].

6.5.3. Relevance in XML search

Keywords: *User Relevance, XML Search.*

Participant: Anne-Marie Vercoustre.

When searching information from structured documents collections such as XML collections, it is expected that the use of the structure will help in two ways:

- specifying more precise queries

- identifying specific and relevant parts of documents instead of full documents.

In the context of INEX (an International Initiative for the Evaluation of XML search), we are interested in evaluating what granularity of elements the users find relevant. We first analysed the relevance assessments to identify the types of highly relevant elements and identified three retrieval scenarios: *Original*, *general* and *specific* and compare the performance of three different systems - a native XML database, a full text retrieval engine and a hybrid system -, for those different scenarios [39]. We then developed a novel retrieval heuristics that dynamically determines the preferable units of retrieval called *Coherent retrieval Element* [20].

In [40], we analyse and compare the assessors'judgement on the relevance of returned document components with the users' behaviour when interacting with components of XML documents. By analysing the level of agreement between the assessor and the users, we show that the highest level of agreement is on highly relevant and on non-relevant document components, suggesting that only the end points of the INEX 10-point relevance scale are perceived in the same way by both the assessor and the users.

7. Contracts and Grants with Industry

7.1. Industrial Contracts

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

Participants: Aicha El Golli, Mihai Jurca, Yves Lechevallier, Bernard Senach, Doru Tanasa, Brigitte Trousse [resp].

Inria Contract Reference: S04 AO485 00 SOPML00 1

The EPIA project "Evolution of an Adaptive Information Portal" got labeled by RNTL 2002, and started on September 2003 until march 2006. Partners are Dalkia, Mediapps and Inria. This year, as Mediapps was bought by Ever (http://www.ever-team.com/es/GetRecords?Template=ET/ET_HomePage expert in integrated ECM[3] software), a change of project management was done in June 2005. The objectives of this project are the following:

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

We hold two project meetings in Paris with Ever (june and september) in order to re-orient the project according to Everteam wishes, taking into account the future integration of Mediapps.net and Net.portal in the eversuite software.

[3] ECM: Enterprise Content Management

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

Participants: Sergiu Chelcea, Mihai Jurca, Bernard Senach, Brigitte Trousse [resp.].

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

MobiVIP, Individual Public Vehicles for Mobility in town centres, is a research project of Predit 3 (Integration of the Communication and Information systems Group). It involves five research laboratories and seven small business companies (SME), in order to experiment, show and evaluate the impact of the NTIC on a new service for mobility in town centres. This service is made up of small urban vehicles completing existing public transport. The MobiVIP project will develop key technological bricks for the integrated deployment of mobility services in urban environment.

The strengths of the project are: 1) the integration between assisted and automatic control, telecommunications, transport modeling, evaluation of service and 2) the demonstrations on 5 complementary experimental sites and 3) the evaluation of possible technology transfer.

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

In december 2004, we finalized in collaboration with B. Senach (Ergomatics Consultants) the delivrable 5.1 [82] which we coordonate with Georges Gallais (Visa Action, Inria Sophia Antipolis). This deliverable aimed at defining a common generic evaluation scenario and proposed a framework to facilitate the identification of the main evaluation dimensions for each planned test or experimentation. This year we had two main tasks: the preparation of the deliverable 5.2 [54] and the NancyCab event.

- The delivrable 5.2 addresses the definition of the main criteria of service quality from the user point of view: this evaluation takes into account the improvment of information access, information content, mobility and man-machine interaction.
- We work on the preparation of the intermediate project evaluation at the NancyCab 2005 (17-18 juin, Stanislas Place, Nancy): participation at a preliminary meeting (18 may), preparation of a demo of our recommender systems published in 2004 [61], [84], a poster and also a presentation on “Scénario générique d’évaluation” for the evaluation workshop. Indeed this intermediate evaluation review has given different articles in various medias (Le Figaro journal 19 june, TV5.org le 20 juin etc.).

More we continued our previous work [61], [84] in the travel information retrieval research field related to mobility in the transport domain. This year we welcomed one internship on this project:

- R. Busseuil who proposed in [47] a new distance for clustering user itineraries (cf. section 6.2.3) and developed an interface for our researches related to urban iteneraries clustering based of Benomad software (cf. Figure 10).

7.1.3. Industrial Contacts

Some contacts during this year:

- SAP, Sophia Antipolis related to data mining and data streams (security and environnement problems). Contact: B. Trousse and F. Masséglia
- Mondeca and Antidot (leadership), in the context of a proposal to RNRT call for proposals. Our proposal called EIFFEL related to semantic web and e-Tourism was accepted. Academic partners are: LIRMM and University of Paris X (Nanterre). Contact: Y. Lechevallier.

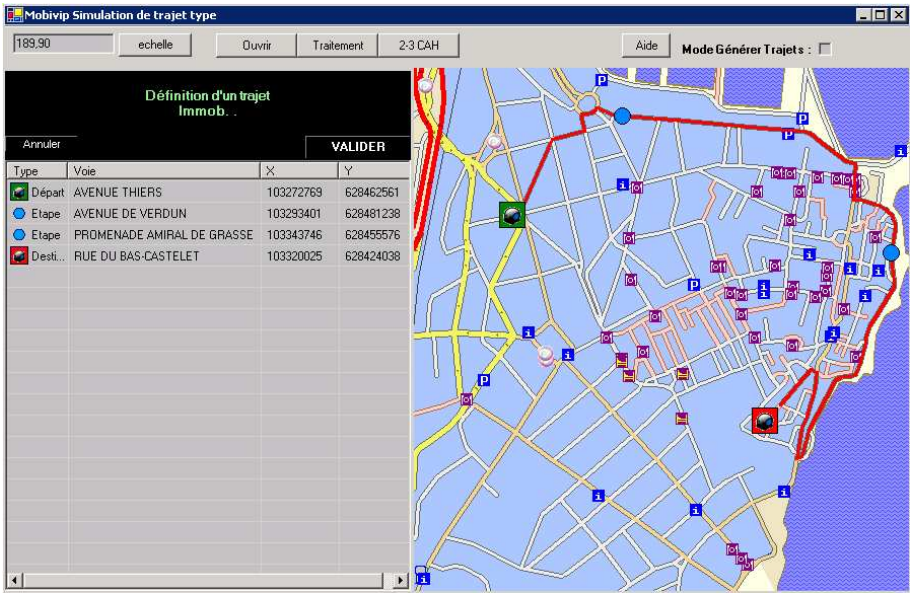

Figure 10. Interface of the urban transport support tool

8. Other Grants and Activities

8.1. Regional Initiatives

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

8.1.1. Color Action: "e-Mimetic"

Our partners are: LePont laboratory (E. Boutin) of the University South Toulon and the IHMH team of LIRMM (M. Nanard, J. Nanard, J-Y. Delors), related to defining and evaluating new Web pages ranking criteria based on page presentation. The web server of this action is e-mimetic.

During this action, we have two internships: Sofiane Sellah [53] located at Inria Sophia Antipolis on the use of generalized URLs for extracting sequential patterns according to different point of views (content or user access) and Patrick Chastellan, located at LIRMM (Montpellier) on the extraction of page presentation criteria.

8.1.2. "Pôle de compétitivité SCS "Solutions Communicantes Sécurisées"

AxIS (B. Senach and B. Trousse) is participating in the preparation of the ROSCOE project related to intelligent transport systems with different partners such: Hitachi Europe (leader), Vu Log, Nexo, Inria (Mascotte), CNRT Télius, etc.

8.1.3. Other initiatives

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

- Laboratoire des Usages, CNRT Télius, http://www.telius.org Sophia Antipolis. B. Trousse is a member of the scientific committee and a substitute member of the management committee. B. SEnach participated to the meeting related to: “Restitution de l’enquéte - Recherche SHS et Société de l’Information en région Provence-Alpes-Cote d’azur”, Marseille (november).
- Supelec: research collaboration with Marie-Aude Aufaure in the context of Baldé’s PhD thesis [17].

8.2. National Initiatives

AxIS is involved in several national working groups.

8.2.1. CNRS RTP 12: “information et connaissance: découvrir et résumer”

In the context of the pluri-disciplinary thematic http://rtp12.loria.fr, we participated to the CNRS Specific Action (AS 120) “Disco Challenge” animated by J.F. Boulicaut and B. Crémilleux. We made a presentation to the Discovery challenge proposed at PKDD/ECML 05 [26].

8.2.2. CNRS Action Concertée Incitative : “Histoire des savoirs”

This initiative (ACI RNR TTT Grammaire et math matique dans le monde indien 17/01/03 - 17/01/06) associates several French research teams from various research fields, such as computer science, data analysis, and Sanskrit literature. The main goal of this action is to provide help for the construction of critical edition of Indian manuscripts in Sanskrit, and to provide pertinent information about the manuscripts classification (construction of cladistic trees). The expected tools will not be restricted to Sanskrit language. This action is completed according some different aspects by the European AAT project which allows us to collect more Sankrit manuscripts and to care about some interactive aspect that we where not able to take into account with the ACI dotation. The action will end in december 2006.

8.2.3. EGC National Group on Mining Complex Data

URL: http://eric.univ-lyon2.fr/projets.php

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

- F. Masséglia with P. Gancarski (LSIIT, Strasbourg) co-organised and co-chaired the second workshop “Fouille de données complexes dans un processus d’extraction de connaissances” (January 18, 2005) [15]. Y. Lechevallier and B. Trousse were members of the program committee.
- F. Masséglia, B. Trousse participated to the meeting of the two national working groups of the national group on mining complex data in relation with the working group 3.4 “Data Mining” of the GDR I3 (May, Paris). F. Masséglia with O. Boussaïd co-animated one of these two working groups: “organisation and structuration of complex data”.
- H. Behja, F. Masséglia and B. Trousse participated in the main meeting of the working group held in Lyon (ERIC) on September 9, 2005. H. Behja made a presentation of his PhD thesis “Vers une approche Web sémantique pour le processus ECD”.

8.2.4. GDR-I3

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

- Working Group 3.4 (GT) on Data Mining animated by P. Poncelet and J.M. Petit. H. Behja, F. Masséglia and B. Trousse participated to the Paris meeting (May) in collaboration with the working group FDC. F. Masséglia and A. Baldé participated to the second meeting at Lyon (November 21). A. Baldé made a presentation of his PhD thesis “Utilisation de métadonnées pour l’aide à l’interprétation de classes et de partitions”.
- GRACQ (*Groupe de Recherche en Acquisition des Connaissances*) (GRACQ): B. Trousse.
- Working Group 3.1 “Sécurité des Systèmes d’Information” animated by D. Boulanger and A. Gabillon: F. Masséglia and B. Trousse.

8.2.5. Other Collaborations

- LIRMM (Montpellier) and Ecole des mines d'Alès (LGI2P): F. Masséglia with M. Tesseire (LIRMM) and P. Poncelet (LGI2P) proposed 1) a survey related to sequential pattern mining method and issues [19] 2) a method dedicated to the management of time constraints in the generalized sequential pattern extraction process [75] and 3) a distributed algorithm for mining users behaviours on a P2P network (a paper about this work has been accepted for the 6th French conference on knowledge extraction and management (EGC'06)).
- ENST: Y. Lechevallier collaborated with Georges Hébrail (ENST).
- Two ARC proposals: 1) ARC SéSur: "Sécurité et Surveillance dans les data streams" (resp: F. Masséglia) with M.O Cordier (DREAM, IRISA) P. Poncelet (LGI2P, Alès) and M. Teisseire (LIRMM, Montpellier); 2) ARC Valex: " Vérification et exploitation de collections de documents scientifiques semi-structurés" (resp: A.-M. Vercoustre) with Annie Morin (TEXMEX, IRISA Rennes), A. Napoli (Orpailleur, INRIA, Nancy), Nathalie Aussenac (IRIT, Toulouse)
- GRIMM-SMASH team (Université Toulouse Le Mirail): F. Rossi works with Nathalie Villa on Support Vector Machines and functional data (cf section 6.3.2 and [46], [44]).
- LITA EA3097 (Université de Metz): F. Rossi and A. El Golli work with Brieuc Conan-Guez on the Self Organizing Map for dissimilarity matrices (see section 6.3.1 and [29]). F. Rossi works with Brieuc Conan-Guez on functional data analysis (cf section 6.3.2 and [21], [22], [23]).

8.3. European Initiatives

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

Participants: Marc Csernel, Sergiu Chelcea, Marina Dufresne, Yves Lechevallier, Sattisvar Tandabany, Brigitte Trousse.

This year we started our project called "AAT" in the context of the EuropAid (DG1) projects and more precisely of the Asia Information Technology (I.T. Asia). We collaborated mainly with François Patte (UFR Maths-Informatique, UNiv Parie 5 René Descartes) and Pascale Haag (EHESS, Centre d'études de l'Inde et de l'Asie du Sud, Paris).

8.3.1.1. The objective of the AAT

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

So there is a chain of copies starting with the original text and continued by an immense tree of hundreds of copies that has grown more or less like a genealogical tree. They are never identical to each other, sometimes extremely different. Parts of the original are missing, fragments are not readable anymore, some have been miscopied, and some others have been voluntarily transformed. This is particularly true for the large Indian subcontinent where at least one third of the manuscript existing through the whole world are supposed to exist, mostly unpreserved , unreferenced , and being at mercy of any accidental event. Even during the 20th century manuscripts were copied by hand by armies of scholars.

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

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

8.3.1.2. Contributions to program

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

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

- Providing the software tools necessary to help the creation of critical edition of the Sanskrit texts. As a secondary result, a distance between the texts should be established based on the presence/absence of the different words in each manuscript (cf. sections 6.2.2, 6.2.7).
- Using the distance established by the first software sets, trying to establish which are the different cluster set of manuscript (for example via a 2-3 HAC clustering), try to establish more or less a phylogeny of the different manuscripts

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

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

Three internships were carried out on this project: S. Tandabany (cf. section 6.2.2), M. Dufresne and S. Kebbache (cf. section 6.2.7).

8.3.2. ERCIM

B. Trousse presented AxIS researches at the Kickoff meeting of the ERCIM Working group on “Data and Information Mining” organised by Christoph Schommer on january 14th at the Campus Kirchberg (University of Luxembourg).

8.3.3. Other Collaborations

- Portugal: B. Trousse collaborated with F. Amilcar Cardoso (amilcar@dei.uc.pt) in the context of the COST Action 282 (2001-2005): “Knowledge Exploration in Science and Technology”. URL:http://www.mpa-garching.mpg.de/~opmolsrv/COST282/.
- Italy, University of Napoli II (Profs C. Lauro and R. Verde) : Y. Lechevallier, A. El Golli, D. Tanasa and B. Trousse
- Italy, KDD Lab. Instituto ISTI, (Fosca Gianotti): some potential collaborations were identified after the ERCIM kickoff meeting.
- Belgium, Facultés Universitaires Notre-Dame de la Paix à Namur (Profs A. Hardy, M. Noirhomme and J.-P. Rasson) [78]: Y. Lechevallier.
- Belgium, Université Catholique de Louvain, DICE Laboratory (Prof. Michel Verleysen, Prof. Vincent Wertz, Dr. Amaury Lendasse, Damien François): F. Rossi was invited professor for one month in 2005 [81], [80].

8.4. International Initiatives

8.4.1. Australia

A-M. Vercoustre collaborates with RMIT, Computer Science department, Melbourne, Australia, on analysing and improving XML search from the point of view of the user [39], [40], [20]. This work is mostly done in the context of the Initiative for the Evaluation of XML Retrieval (INEX-2004), DELOS network of Excellence.

8.4.2. Brazil

We continue our collaboration on clustering and web usage mining with F.A.T. de Carvalho from Federal University of Pernambuco (Recife) and his team. We welcomed F.A.T. de Carvalho, during 3 months, September to November. During this year some analysis of Recife Log files are carried out by students from Recife University, Napoly University and by AxIS members. Common papers have been proposed at EGC and IFCS. Moreover, we have finished the submission of an article for Computational Statistics Journal (Springer). A poster was presented during the meeting (October 13 to 14) on information technologies of Brazil. We have proposed a joint research projet in the framework of the cooperation between INRIA and FACEPE. We also welcomed Thesera Ludemir, during 15 days which has worked with Fabrice Rossi on neural methods.

Two internships were done on this project:

- A. Da Silva, Clustering in Web Usage Mining with Recife Log files and workbench of PKDD (cf. section 6.4.2) [26], [27].
- N. Lopes Calvacanti Junior has started an internship in October 2005 (planned ending date: march 2006). He is working on the implementation of the Self Organizing Map on dissimilarity matrices (see section 6.3.1).

8.4.3. Canada

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

Osmar Zaiane, Professor at the university of Alberta (Canada), visited us during five days at Inria Sophia Antipolis. He participated in Doru Tanasa's Ph.D. committee on 3rd June.

8.4.4. India

Marc Csernel collaborated with the Bhandrakar Institute (India) and the Mahendra Sanskrit University (Nepal) via the CNRS action "History of Knowledge" (cf. section 8.2.2) and also via the consortium members of EuropeAid projet of Asia-Information Technology and Communications (6.2.2, 8.3.1).

8.4.5. Morocco

AxIS is involved in a France-Morocco thematic network in software engineering. In this context, B. Trousse co-supervises with Abdelaziz Marzark (University of Casablanca) a Ph.D. student: H. Behja (ENSAM, Meknès, Morocco). H. Behja visited us for his thesis work during the summer period and also for our annual AxIS workshop (october 24-26). Mr. Marzark visited us in November (10-13).

8.4.6. Romania

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

8.4.7. Tunisia

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

9. Dissemination

9.1. Promotion of the Scientific Community

9.1.1. Journals

AxIS members belongs to editorial boards of two international journals and three national journals:

- the Co-Design Journal (Editor: S. Scrivener, Coventry University, UK - Publisher: Swets & Zeitlinger): B.Trousse
- the Journal of Symbolic Data Analysis (JSDA) (Editor: E. Diday, electronic journal http://www.jsda.unina2.it): Y. Lechevallier, F. Rossi and B. Trousse.
- the RIA journal (<< Revue d'Intelligence Artificielle >>) (Hermès publisher; editor in chief: M. Pomerol): B. Trousse.
- the I3 electronic journal of the GDR-I3 (editor-in-chief: C. Garbay et H. Prade) http://www.Revue-I3.org/: B. Trousse.
- La revue MODULAD (electronic journal, http://www.modulad.fr/): Y. Lechevallier is one of the main editors and F. Rossi is a member of the editorial board

F. Masséglia and B. Trousse were invited editors (with O. Boussaid and P. Gancarski) of a Special issue of the RNTI (Revue des Nouvelles Technologies de l'Information) on "Complex Data Mining" [14]. Y. Lechevallier was a member of the editorial board and A-M. Vercoustre was an additional reviewer.

AxIS members were reviewers for sixteen international and national journals and for one international book

- the IEEE Journal of Transactions on Data and Knowledge Engineering (TDKE): F. Masséglia, B. Trousse (http://www.computer.org/tkde/)
- the International Journal 'Behaviour & Information Technology' (BIT: Taylor & Francis Publisher)
- the Information Systems (IS) Journal: F. Masséglia. (http://ees.elsevier.com/is/)
- the Data Mining and Knowledge Discovery (DMKD) Journal (twice): F. Masséglia (https://www.editorialmanager.com/dami/)
- the Journal of Systems and Software (JSS): F. Masséglia. (http://ees.elsevier.com/jss/)
- the Data and Knowledge Engineering Journal (DKE): F. Masséglia. (http://www.sciencedirect.com/science/journal/0169023
- La revue Modulad: F. Rossi (http://www.modulad.fr/)
- Computational Statistics: F. Rossi (http://comst.wiwi.hu-berlin.de/)
- Scandinavian Journal of Statistics: F. Rossi (http://www.blackwellpublishing.com/journal.asp?ref=0303-6898)
- Computational Statistics and Data Analysis: F. Rossi (http://www.elsevier.com/locate/csda)
- Neural Processing Letters: F. Rossi (http://www.springerlink.com/openurl.asp?genre=journal&issn=1370-4621)
- Revue de Statistique Appliquée: F. Rossi (http://www.sfds.asso.fr/publicat/rsa.htm)
- Neurocomputing (twice): F. Rossi (http://www.elsevier.com/locate/issn/09252312)
- Computational Geosciences: F. Rossi (http://www.springeronline.com/journal/10596/about)
- Control and Intelligent Systems Journal: F. Rossi (http://www.actapress.com/Content_of_Journal.aspx?journalID=58)
- AI Communications: F. Rossi (http://aicom.web.cse.unsw.edu.au/)
- Book on "Processing and Managing Complex Data for Decision Support": F. Masséglia, B. Trousse. (http://chirouble.univ-lyon2.fr/bderic/publications/index.php)

9.1.2. Program Committees

Several AxIS members were involved at national or international conferences/worhshops as member of Program Committee or as additional reviewer. Let us note that we organized this year two workshops this year (FDC at EGC05 and MDM/KDD'05).

9.1.2.1. National Conferences/Workshops

- Atelier FDC Fouille de Données Complexes (at EGC'05): F. Masséglia (co-chair), B. Trousse, F. Rossi (additional reviewer) (http://www-sop.inria.fr/axis/fdc-egc05/FDC05_CFP.htm)
- EGC'05 (Paris, January) Extraction et Gestion des Connaissances: Y. Lechevallier, F. Rossi (additional reviewer) and B. Trousse. (http://www.math-info.univ-paris5.fr/egc2005/index.php)
- Atelier sur la modélisation utilisateurs et personnalisation de l'interaction homme-machine (at EGC'2005): B. Trousse (http://www-connex.lip6.fr/~artieres/EGC/Atelier_MU_EGC.html)
- Plateforme AFIA 2005, atelier RàPC Raisonnement à Partir de Cas (Nice, May): B. Trousse (http://www-sop.inria.fr/acacia/afia2005/rapc/)
- CORIA 2005 (Grenoble, March) Conférence en Recherche d'Informations et Applications: B. Trousse (http://www-clips.imag.fr/mrim/coria05/main_coria.html)
- UBIMOB 2005 Mobilité et Ubiquité (Grenoble, France, June): B. Trousse (http://www-lsr.imag.fr/UbiMob05/index.html)
- SSTIC 2005 (Rennes, June) Symposium sur la Sécurité des Technologies de l'Information et des Communications: F. Rossi (additional reviewer) (http://www.sstic.org/SSTIC05/info.do)
- BDA 2005 (Saint-Malo, October) Conférence Bases de Données Avancées: F. Masséglia (additional reviewer).

9.1.2.2. International Conferences/Workshops

- MDM/KDD'05 the sixth International Workshop on Multimedia Data Mining [13] held in conjunction with KDD'05, Chicago, USA): F. Masséglia was co–chair with Fatma Bouali (Univ. Lille) and Latifur Khan (Univ. Texas at Dallas). (http://www-sop.inria.fr/axis/mdm-kdd05/MDM05.htm
- GfKI 2005 (Magdeburg, Germany, March) 29th annual conference of the German Classification Societe - From Date and Information Analysis to Knowledge Engineering: F. Rossi (additional reviewer) (http://omen.cs.uni-magdeburg.de/itikmd/gfkl2005)
- WWW 2005 (Valencia, Spain, March) 1st Int'l Workshop on Automated Specification and Verification of Web Sites: T. Despeyroux (http://www.dsic.upv.es/workshops/wwv05/)
- ESANN 2005 (Bruges, Belgium, April) 13th European Symposium on Artificial Neural Network: F. Rossi (http://www.dice.ucl.ac.be/esann/)
- CSCW 2005 (Coventry, UK, May) 9th International Conference on CSCW in Design: B. Trousse (http://2005.cscwid.org/)
- CIR 2005 (Paris, July)International Workshop on Context-Based Information Retrieval: A-M. Vercoustre (http://wwwsi.supelec.fr/bld/CIR-2005/)
- IJCAI 2005 (Edinburg, Scotland, Aug.) 19th International Joint Conference on Artificial Intelligence: B. Trousse (http://ijcai05.csd.abdn.ac.uk/)
- ACM SIGIR 2005 (Salvador, Brazil, August)Conference on Research and Development in Information Retrieval, posters: A-M. Veroustre (http://www.dcc.ufmg.br/eventos/sigir2005/)
- 6th International Workshop on Multimedia Data Mining, in conjunction with KDD-2005 (Chicago, USA, August): F. Masséglia, B. Trousse (http://www-sop.inria.fr/axis/mdm-test/content.htm)

- ICCBR 2005 (Chicago, USA, August) 6th International Conference on Case-Based Reasoning: B. Trousse (http://www.iccbr.org:8080/iccbr05/index.jsp)
- ICANN 2005 (Warsaw, Poland, September) International Conference on Artificial Neural Networks: F. Rossi (additionnal reviewer) (http://www.ibspan.waw.pl/ICANN-2005/)
- Intellicom 2005 (Montreal, Canada, October) IFIP International Conference on Intelligence in Communication Systems: A-M. Vercoustre (http://www.congresbcu.com/intellcomm2005/)
- ACM DocEng 2005 (Bristol, UK, November) Symposium on Document Engineering: A-M. Vercoustre (http://www.documentengineering.org/)
- ADCS 2005 (Sidney, Australia, December) 10th Australasian Document Computing Symposium: A-M. Vercoustre (http://goanna.cs.rmit.edu.au/~aht/adcs2005/)
- ICPReMI'05 (Kolkata, India, December) Technical session on Symbolic Data Analysis: M. Csernel (http://www.isical.ac.in/~premi05/)
- MSTD 2005 (Porto, Portugal) the first international workshop on Mining Spatio-Temporal Data (held in conjunction with PKDD'05): F. Masséglia. (http://www.di.uniba.it/~malerba/activities/mstd/)
- IDEAS 2005 (Montreal, Canada) the 9th International Database Engineering & Application Symposium: F. Masséglia (additional reviewer). (http://ideas.concordia.ca/ideas2005/)

9.1.3. Invited Seminars

- ENIT Tunisia, Séminaire RAIDI: Y. Lechevallier ("Classification automatique dans le Web Usage Mining"), january.
- Project Team CORTEX (LORIA), March 2005: F. Rossi ("Une implémentation efficace du SOM sur tableau de dissimilarités")
- Common day between the following scientific associations SFDS, EGC, SFC, INFORSID and AFIA, March 21, organised by the French Society of Statistics (SFdS): Y. Lechevallier ("Le tableau de données, une structure unique, des réalités multiples").
- Machine Learning Group (DICE Laboratory), Université Catholique de Louvain (Belgium), June 2005: F. Rossi ("Classification in Functional Spaces with Support Vector Machines")
- National Group on "Mining Complexe data", H. Behja, september 9, Uiversity of Lyon II. "Vers une approche web sémantique du processus d'ECD".
- Seminar "Logiciels libres en Data Mining", October 13, organised by the French Society of Statistics (SFdS), InfoStat group , Data Mining et Logiciels: Y. Lechevallier ("WEKA, un logiciel libre en Data Mining").
- E.N.S. "Ecole Normale Supérieure"(Ulm, Paris): M. Csernel presented a conference entitled "Grammaire et mathématiques dans le monde indien: histoire des savoirs, histoire des textes et nouvelles technologies au service de la philologie" in december.

9.1.4. Organization of Conferences or Workshops

Besides the organization of the workshops FDC/EGC'05) and MDM/KDD'05, we are involved in others organization tasks:

- Member of the organizing committee of the first IEEE international workshop on "Mining Complex Data" (held in conjunction with ICDM, Houston, November): F. Masséglia.
- Organisation of our annual AxIS workshop at Inria Rocquencourt (24-26 october): S. Aubin, S. Honnorat, Y. Lechevallier and B. Trousse. Monthly team meetings were organised by videoconferecne between AxIS Sophia Antipolis and AxIS Rocquencourt.

9.1.5. AxIS Web Server

AxIS maintains an external and an internal Web site allowing the access to lots of information, including software developed in the team, our publications, relevant events (conferences, workshops) and information related to the conferences and seminar we organise. URL:http://www-sop.inria.fr/axis/.
S. Chelcea delevoped our publication management tool called "BibAdmin". BibAdmin is a collection of PHP/MySQL scripts for bibliographic (Bibtex) management over the Web. Publications are stored in a MySQL database and can be added/edited/modified via a Web interface. It is specially designed for research teams to easily manage their publications or references and to make their results more visible. Users can build different private/public bibliographies which can be then used to compile LaTeX documents. BibAdmin is made available under the GNU GPL license on INRIA's GForge server at: http://gforge.inria.fr/projects/bibadmin/

9.1.6. Activities of General Interest

- T. Despeyroux is president of AGOS (Inria Works Council), a permanent member of the "commission technique paritaire (CTP)" and a member of the Inria Board of Directors (Conseil d'Administration) as a scientific staff representative.
- T. Despeyroux is participating in the project for redesigning the intranet Web site of Inria-Rocquencourt.
- B. Senach is involved in the Inria Sophia Antipolis support committee of the world-wide competitivity pole "Solutions Communicantes Sécurisées"
- B. Trousse is a member of the scientific committee and also a substitute member of the decision committee of the "Laboratoire des Usages des NTIC" of Sophia Antipolis.
- B. Trousse is a member of the RSTI scientific committee related to the << ISI, L'OBJET, RIA, TSI >> journals (Hermès publisher).
- A-M Vercoustre is involved (25%) in the Department for Scientific Information and Communication (DISC), working on Inria policy and tools for scientific publications, in particular the development of an Open Archive in cooperation with CNRS.

9.2. Formation

9.2.1. University Teaching

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

- "DEA Informatique" (resp. Mr Kounalis) at UNSA Sophia Antipolis: Optional tutorial on "Web usage Mining" (F. Masséglia, B. Trousse).
- "Licence professionnelle franco-italienne: Statistiques et Traitement Informatique de Données (STID)" (resp. J. Lemaire) at UNSA, Menton: Supervision of a student project (60h by students, 10 students, 30h supervised) on *Mining HTTP Logs From Inria's Web Sites*: S. Chelcea, B. Trousse.
- International University of Monaco, three courses (D. Tanasa): Programming Techniques (90h), Business Analysis and Systems Design (45h), Management of Information Systems (25h).
- "DEA Modélisation et traitement des données et des connaissances" (resp: S. Pinson) of the University Paris IX-Dauphine (4h): Tutorial on *"Analyse des connaissances numériques et Symboliques"*: Y. Lechevallier.
- "DESS Mathématiques appliquées et sciences économiques (MASE)" of the University Paris IX-Dauphine: Tutorial (18h) on *"Méthodes neuronales en classification"*: Y. Lechevallier.
- "ENSAE": Tutorial on *"Data Mining"* (12h): Y. Lechevallier.
- "ENIT /Tunis": Tutorial on *"Data Mining et méthodes neuronales"* (12h): Y. Lechevallier.

9.2.2. Ph.D. Thesis

Ph.D. defence in 2005:

1. **D. Tanasa**, (start: end of 2001), “Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support” [16]), University of Nice-Sophia Antipolis (director: B. Trousse), june 3rd.

Ph.D. in progress:

1. **S. Chelcea**, (start: end of 2002), “Agglomerative 2-3 Hierarchical Clustering: theoretical and applicative study”, Université de Nice-Sophia Antipolis (directors: J. Lemaire and B. Trousse with the support of P. Bertrand on 2-3 AHC).
2. **H. Behja**, (start: end of 2002), “Gestion de points de vues multiples dans l’analyse d’un observatoire sur le Web”, University of Casablanca, (directors: A. Marzark and B. Trousse). This thesis is done in the context of the STIC Software engineering network of France-Morocco cooperation (2002-2005).
3. **A. Baldé**, (start: end of 2003), ”Extraction de méta-données à partir de prototypes issus d’une classification” (Metadata Extraction from classification prototypes), University of Paris IX Dauphine, (directors: E. Diday and Y. Lechevallier) with the participation of B. Trousse and M.-A. Aufaure (Supelec).
4. **A Da Silva**, (start: October 2005), "Modélisation de données agrégées ou complexes par l’approche symbolique, application au Web Usage Mining", University of Paris IX Dauphine (directors: Edwin Diday and Yves Lechevallier).
5. **A. Marascu**, (start: October 2005), “Extraction de Motifs Séquentiels dans les Data Streams”, Université de Nice-Sophia Antipolis (director: Yves Lechevallier) with the participation of B. Trousse and F. Masseglia).

F. Rossi is a member of the thesis committee of **N. Delannay** (start: October 2003) on “Méthodes neuronales pour les données structurées”, Université Catholique de Louvain, Belgium (director: Michel Verleysen).

AxIS researchers were members of Ph.D. committees in 2005:

- Doru Tanasa, “Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support”, defended on 3rd. June 2005: B. Trousse, F. Masséglia
- Cherif Mballo, “ Ordre, codage et extension du critère de Kolomogorov-Smirnov pour la segmentation de données symboliques”, December 12, Y. Lechevallier
- Hani Hamdan, “Developpement de méthodes de classification pour le contrôle d’émission accoustique d’appareils sous pression”, November 22, Y. Lechevallier
- Nathalie Villa “Éléments d’apprentissage en statistique fonctionnelle - Classification et régression fonctionnelles par réseaux de neurones et Support Vector Machine”, defended on October 21th, 2005: F. Rossi
- Jonathan Mamou, “XSEarch, un moteur de recherche pour XML combinant structure et contenu”, Univ. Paris 11, defended on 30th September 2005, Orsay: A.-M. Vercoustre

9.2.3. Internships

We welcomed eleven students in AXIS this year:

1. **R. Busseuil** (supervisors S. Chelcea and B. Trousse), ENS Cachan and Inria Sophia-Antipolis, “Classification des itinéraires pour l’aide á la navigation assistée par GPS” [47].
2. **M. Fegas** (supervisor A.-M. Vercoustre), University of Orsay Paris-11 and Inria Rocquencourt, “Classification de documents XML” [48].
3. **M. Dufresne** (supervisors Marc Csernel, Yves lechevallier, Francois Patte), Univ. Paris XIII Institut galilée, Inria Rocquencourt, Interactive presentation of Critical edition of Sankrit texts.
4. **S. Kebbache** (supervisors Marc Csernel, Yves lechevallier), Univ. Paris I Panthéon-Sorbonne, Inria Rocquencourt, Comparison of Sanskrit texts, aligment procedures.
5. **A. Marascu** (supervisor F. Masséglia), University of Nice and Inria Sophia-Antipolis, “Extraction de motifs séquentiels dans les data streams” [52].
6. **S. Sellah** (supervisors F. Masséglia and B. Trousse), University of Lyon 2 and Inria Sophia Antipolis, [53].
7. **S. Tandabany** (supervisors S. Chelcea and B. Trousse), University of Orsay Paris-11, ENS Lyon and Inria Sophia Antipolis, “Elaborating a Distance for Clusterig Homogeneous Sanskrit Documents” [55].

Four of which were in the context of the Inria international internship program:

1. **A. Da Silva** (supervisors Y. Lechevallier and B. Trousse), Recife University of Permenbouc (Brazil), Inria Rocquencourt, Inria International Internship Program.
2. **C. Garboni**, West University of Timisoara (Romania) and Inria Sophia Antipolis, Inria International Internship Program [49].
3. **S. Gul** (supervisor A.-M. Vercoustre), MIT and Inria Rocquencourt, XML Document mining (in progress), Inria Internship Program.
4. **N. Lopes Calvacanti Junior** (supervisor F. Rossi), Federal Univ. of Pernambuco, Brazil and Inria Rocquencourt, Implementation of a fast Dissimilarity Self-Organizing Map (in progress), Inria Internship Program.

9.3. Participation to Workshops, Conferences, Seminars, Invitations

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

- EDA’05 (Entrepôts de Donnes et Analyse en ligne), Lyon, June 10: F. Masséglia.
- Creating the information Commons for e-Science: Towards Institutional Policies and Guidelines for Action, UNESCO Headquarters, Paris, France, September 2005.

F. Rossi was invited professor for one month at the Université Catholique de Louvain (Belgium).
M. Csernel was invited in August in the framework ot the AAT project in two different EFEO branches in India: the Poonah branch and the Pondichery branch.

10. Bibliography

Major publications by the team in recent years

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

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

[3] M. CHAVENT. *A monothetic clustering method*, in "Pattern Recognition Letters", vol. 19, n° 11, September 1998, p. 989-996.

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

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

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

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

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

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

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

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

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

Books and Monographs

[13] F. BOUALI, L. KAHN, F. MASSEGLIA (editors). *Proceedings of MDM'05, the sixth international workshop on « Multimedia Data Mining »*, (held in conjunction with KDD'05), 2005, http://www-sop.inria.fr/axis/mdm-kdd05/.

[14] O. BOUSSAID, P. GANÇARSKI, F. MASSEGLIA, B. TROUSSE (editors). *Numéro spécial « Fouille de données complexes »*, Revue des Nouvelles Technologies de l'Information (RNTI), vol. 7, Cépaduès, 2005.

[15] P. GANÇARSKI, F. MASSEGLIA (editors). *Actes de FDC'05, le second atelier sur la « Fouille de données complexes dans un processus d'extraction de connaissances »*, (Atelier de la conférence EGC'05), 2005, http://www-sop.inria.fr/axis/fdc-egc05/FDC05_CFP.htm.

Doctoral dissertations and Habilitation theses

[16] D. TANASA. *Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support*, Ph. D. Thesis, University of Nice Sophia Antipolis, 3 June 2005, http://www-sop.inria.fr/axis/personnel/Doru.Tanasa/these_TANASA.pdf.

Articles in refereed journals and book chapters

[17] A. BALDÉ, M.-A. AUFAURE. *How can metadata contribute to add semantic information to clusters?*, in "Journal of Symbolic Data Analysis", vol. 3, n° 1, 2005, p. 32-44, http://www-sop.inria.fr/axis/Publications/uploads/pdf/Article.pdf.

[18] M. CSERNEL, P. BERTRAND. *Comparaison de manuscrits sanskrits, édition critique et classification*, in "La Revue Modulad", n° 23, 2005, p. 1-20.

[19] F. MASSEGLIA, M. TEISSEIRE, P. PONCELET. *Sequential Pattern Mining: A Survey on Issues and Approaches*, Information Science Publishing, 2005.

[20] J. PEHCEVSKI, J. THOM, A.-M. VERCOUSTRE. *Hybrid XML Retrieval: Combining Information Retrieval and a Native XML Database*, in "Journal of Information Retrieval, Special Issue on INEX", Postprint version. The editor version can be accessed through the DOI., vol. 8, n° 4, 2005, http://hal.inria.fr/inria-00000183.

[21] F. ROSSI, B. CONAN-GUEZ. *Estimation consistante des paramètres d'un modèle non linéaire pour des données fonctionnelles discrétisées aléatoirement*, in "Comptes rendus de l'Académie des Sciences - Série I", vol. 340, n° 2, January 2005, p. 167–170, http://hal.inria.fr/inria-00000223.

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

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

[24] F. D. A. T. DE CARVALHO, R. M. C. R. DE SOUZA, M. CHAVENT, Y. LECHEVALLIER. *Adaptive Hausdorff distances and dynamic clustering of symbolic interval data*, in "Pattern Recognition Letters", article in press,

2005.

Publications in Conferences and Workshops

[25] H. BEHJA, B. TROUSSE, A. MARZAK. *Prise en compte des Points de Vue pour l'annotation d'un processus d'Extraction de Connaissances à partir de Données*, in "Actes des 5ème journées Extraction et Gestion des Connaissances (EGC 2005), Revue des Nouvelles Technologies de l'Information (RNTI-E-3)", S. PINSON, N. VINCENT (editors). , vol. 1, Cépaduès-Editions, ISBN 2.85428.677.4, January 2005, p. 245-256.

[26] S. CHELCEA, A. DA SILVA, Y. LECHEVALLIER, D. TANASA, B. TROUSSE. *Benefits of InterSite Pre-Processing and Clustering Methods in E-Commerce Domain*, in "Proceedings of the ECML/PKDD2005 Discovery Challenge, A Collaborative Effort in Knowledge Discovery from Databases, Porto, Portugal", P. BERKA, B. CRÉMILLEUX (editors). , 3-7 October 2005, p. 15-21, http://hal.inria.fr/inria-00000880.

[27] S. CHELCEA, A. DA SILVA, Y. LECHEVALLIER, D. TANASA, B. TROUSSE. *Pre-Processing and Clustering Complex Data in E-Commerce Domain*, in "Proceedings of the First International Workshop on Mining Complex Data 2005 (IEEE MCD'2005), held in conjunction with the Fifth IEEE International Conference on Data Mining (ICDM'05), Houston, Texas", 27 November 2005, http://hal.inria.fr/inria-00000881.

[28] S. CHELCEA, B. TROUSSE. *Classification 2-3 hiérarchique de données du Web*, in "Actes des 5ème journées Extraction et Gestion des Connaissances (EGC 2005), Revue des Nouvelles Technologies de l'Information (RNTI-E-3)", S. PINSON, N. VINCENT (editors). , vol. 1, Cépaudès-Editions, ISBN 2.85428.677.4, January 2005, 219, http://hal.inria.fr/inria-00000864.

[29] B. CONAN-GUEZ, F. ROSSI, A. EL GOLLI. *A Fast Algorithm for the Self-Organizing Map on Dissimilarity Data*, in "Proceedings of the 5th Workshop on Self-Organizing Maps (WSOM 05), Paris (France)", September 2005, p. 561–568.

[30] M. CSERNEL, P. BERTRAND. *Comparaison de textes sanskrits en vue d'une édition critique*, in "Compte rendu des 12-emes Rencontres de la Société Francophone de classification, Montreal", V. MAKARENKOV, G. CUCUMEL, J.-F. LAPOINTE (editors). , SFC, 29 May - 1 June 2005, p. 108-111.

[31] A. DA SILVA, F. D. A. T. DE CARVALHO, Y. LECHEVALLIER, B. TROUSSE. *Uma Abordagem para Descoberta do Perfil da Clientela em Web Sites Institucionais*, in "Proceedings of the first ECML/PKDD Workshop on Data Mining for Business (DMBiz'05), held in conjunction with the 16th European Conference on Machine Learning (ECML'05) and the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05), Porto, Portugal", Poster, 3 October 2005.

[32] T. DESPEYROUX, Y. LECHEVALLIER, B. TROUSSE, A.-M. VERCOUSTRE. *Experiments in Clustering Homogeneous XML Documents to Validate an Existing Typology*, in "Proceedings of the 5th International Conference on Knowledge Management (I-Know), Vienne, Autriche", (postprint), n° 1, Journal of Universal Computer Science, July 2005, http://hal.inria.fr/inria-00000002.

[33] T. DESPEYROUX, Y. LECHEVALLIER, B. TROUSSE, A.-M. VERCOUSTRE. *Expériences de classification de documents XML homogènes*, in "Actes des 5ème journées Extraction et Gestion des Connaissances (EGC 2005), Revue des Nouvelles Technologies de l'Information (RNTI-E-3), Paris, France", N. VINCENT, S. PINSON (editors). , vol. 1, Cépaduès-Editions, January 2005, p. 183-188,

http://hal.inria.fr/docs/00/03/49/10/PDF/raclass.pdf.

[34] A. EL GOLLI. *Criterion-based divisive clustering: application to symbolic objects generation*, in "Proceedings of XIth International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005), Brest, France", 17-20 May 2005, p. 709–713, http://www-sop.inria.fr/axis/Publications/uploads/pdf/ASMDAsectionSymbolic.pdf.

[35] Y. LECHEVALLIER, R. VERDE. *Clustering Methods in Symbolic Data Analysis*, in "Proceedings of the 55th session of the International of Statictics Society (ISI 2005), Sydney", 2005.

[36] A. MARASCU, F. MASSEGLIA. *Mining Data Streams for Frequent Sequences Extraction*, in "Proceedings of the First International Workshop on Mining Complex Data 2005 (IEEE MCD'2005),held in conjunction with the Fifth IEEE International Conference on Data Mining (ICDM'05), Houston, USA", 27 November 2005, http://www-sop.inria.fr/axis/Publications/uploads/pdf/MCD_05.pdf.

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

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

[39] J. PEHCEVSKI, J. THOM, S. TAHAGHOGHI, A.-M. VERCOUSTRE. *Hybrid XML Retrieval Revisited*, in "Advances in XML Information Retrieval, Proceedings of the Third International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2004), Schloss Dagstuhl, Germany", N. FUHR, M. LALMAS, S. MALIK, Z. SZLÁVIK (editors). , LNCS, Postprint, vol. 3493, Springer, 2005, http://hal.inria.fr/inria-00000003.

[40] J. PEHCEVSKI, J. THOM, A.-M. VERCOUSTRE. *Users and Assessors in the Context of INEX: Are Relevance Dimensions Relevant?*, in "Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, Glasgow, Scotland", A. TROTMAN, M. LALMAS, N. FUHR (editors). , 30 July 2005, http://hal.inria.fr/inria-00000182.

[41] J.-P. RASSON, Y. LECHEVALLIER. *Parsimoious representation of symbolic objects : the case of multidimensional intervals*, in "Proceedings of the 55th session of the International of Statictics Society (ISI)", ISI, 2005.

[42] F. ROSSI, A. EL GOLLI, Y. LECHEVALLIER. *Usage Guided Clustering of Web Pages with the Median Self Organizing Map*, in "Proceedings of the 13th European Symposium On Artificial Neural Networks (ESANN 2005), Bruges, Belgium", April 2005, p. 351–356.

[43] F. ROSSI, Y. LECHEVALLIER, A. EL GOLLI. *Visualisation de la perception d'un site web par ses utilisateurs*, in "Actes des 5ème journées Extraction et Gestion des Connaissances (EGC 2005), Revue des Nouvelles Technologies de l'Information (RNTI-E-3), Paris, France", S. PINSON, N. VINCENT (editors). , vol. II, Cépaduès-Editions, January 2005, p. 563–574.

[44] F. ROSSI, N. VILLA. *Classification in Hilbert Spaces with Support Vector Machines*, in "Proceedings of XIth International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005), Brest, France", May 2005.

[45] R. VERDE, Y. LECHEVALLIER. *Classification croisée d'un tableau de données de type intervalle*, in "Actes des 12èmes Rencontres de la Société Francophone de Classification (SFC 2005), Montréal", 2005.

[46] N. VILLA, F. ROSSI. *Support Vector Machine For Functional Data Classification*, in "Proceedings of the 13th European Symposium On Artificial Neural Networks (ESANN 2005), Bruges, Belgium", April 2005, p. 467–472.

Miscellaneous

[47] R. BUSSEUIL. *Classification des itinéraires pour l'aide à la navigation assistée par GPS*, 2005, http://www-sop.inria.fr/axis/Publications/uploads/pdf/rapportREMY.pdf.

[48] M. FEGAS. *Classification de documents XML, Application au corpus d'INEX et aux rapports d'activité INRIA*, 2005, http://www-sop.inria.fr/axis/Publications/uploads/pdf/Rapport-Mounir.pdf.

[49] C. GARBONI. *Sequential Pattern Mining for Structure-based XML Document Classification*, 2005.

[50] Y. LECHEVALLIER. *Le tableau de données, une structure unique, des réalités multiples*, 21 March 2005, RDC'2005 ENST Paris.

[51] Y. LECHEVALLIER. *WEKA, un logiciel libre d'apprentissage et de Data Mining*, 13 October 2005, Après-midi d'INFOSTAT, Paris.

[52] A. MARASCU. *Extraction de motifs séquentiels dans les data streams*, 2005, http://www-sop.inria.fr/axis/Publications/show.php?author=Alice_Marascu.

[53] S. SELLAH. *Web Usage Mining : Extraction de motifs séquentiels selon plusieurs points de vue*, 2005.

[54] B. SENACH, B. TROUSSE. *Spécifications des critères d'évaluation et des données d'usage à tracer*, INRIA Sophia Antipolis, Delivrable D5.2. du Contrat PREDIT MOBIVIP, December 2005.

[55] S. TANDABANY. *Elaborating a Distance for Clustering Homogeneous Sanskrit Documents*, 2005, http://www-sop.inria.fr/axis/Publications/uploads/pdf/ReportSatti.pdf.

[56] A.-M. VERCOUSTRE, M. FEGAS, S. GUL, Y. LECHEVALLIER. *A Flexible Structured-based Representation for XML Document MiningPre-Proceedings of the Fourth International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2005), Schloss Dagstuhl, Germany*, November 2005, http://hal.inria.fr/inria-00000839.

Bibliography in notes

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

[58] S. CHELCEA, P. BERTRAND, B. TROUSSE. *Theoretical study of a new 2-3 Hierarchical Clustering Algorithm*, in "Proceedings of the 4th International Workshop on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2002, Timisoara, Romania", 8-12 October 2002.

[59] S. CHELCEA, P. BERTRAND, B. TROUSSE. *A New Agglomerative 2-3 Hierarchical Clustering Algorithm*, in "Innovations in Classification, Data Science, and Information Systems. Proc. 27th Annual GfKl Conference, University of Cottbus, March 12 - 14, 2003, Heidelberg-Berlin, Germany", D. BAIER, K.-D. WERNECKE (editors). , Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, ISBN 3-540-23221-4, 2004, p. 3-10, http://www.springeronline.com/sgw/cda/frontpage/0,11855,5-10129-22-36021200-0,00.html.

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

[61] S. CHELCEA, G. GALLAIS, B. TROUSSE. *Recommandations personnalisées pour la recherche d'information facilitant les déplacements*, in "Premières Journées Francophones : Mobilité et Ubiquité 2004, ESSI, Nice, Sophia-Antipolis, France", Cepadues - ISBN : 2-85428-653-7 / ACM Digital Library - ISBN : 1-58113-915-2, 1-3 June 2004, p. 143 - 150, http://portal.acm.org/citation.cfm?id=1050873.1050905.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[81] F. ROSSI, A. LENDASSE, D. FRANÇOIS, V. WERTZ, M. VERLEYSEN. *Mutual information for the selection of relevant variables in spectrometric nonlinear modelling*, in "Chemometrics and Intelligent Laboratory Systems", In press, 2005.

[82] B. SENACH, B. TROUSSE. *Définition du scénario générique guidant l'évalaution du service VIP*, INRIA Sophia Antipolis, Delivrable D5.2. du Projet PREDIT MOBIVIP, December 2004.

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

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

[85] S. WESS, K.-D. ALTHOFF, G. DERWAND. *Using K-d Trees to Improve the Retrieval Step in Case-Based Reasoning*, in "Lecture Notes in Artificial Intelligence, Topics in Case-Based Reasoning", S. WESS, K. ALTHOFF, M. M. RICHTER (editors). , Springer-Verlag, 1994, p. 167-181.

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

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

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