Improving Effectiveness on Clickstream Data Mining

Cristina Wanzeller[1] and Orlando Belo[2]

[1] Departamento de Informática, Instituto Superior Politécnico de Viseu,
Escola Superior de Tecnologia de Viseu, Campus Politécnico de Repeses,
3505-510 Viseu, Portugal
cwanzeller@di.estv.ipv.pt
[2] Departamento de Informática, Escola de Engenharia,
Universidade do Minho, Campus de Gualtar,
4710-057 Braga, Portugal
obelo@di.uminho.pt

Abstract. Developing and applying data mining processes are often very complex tasks to users without deep knowledge in this domain, particularly when such tasks involve *clickstream* data processing. One important and known challenge arises in the selection of mining methods to apply on a specific data analysis problem, trying to get better and useful results for a particular goal. Our approach to address this challenge relies on the reuse of the acquired experience from similar problems, which had provided successful mining processes in the past. In order to accomplish such goal, we implemented a prototype mining plans selection system, based on the Case-Based Reasoning paradigm. In this paper we explain how this paradigm and the implemented system may be explored to assist decisions on the data mining or Web usage mining specific scope. Additionally, we also identify the underlying issues and the approaches that were followed.

1 Introduction

Web Usage Mining (WUM) concerns to the application of mining methods to data related to the interaction processes between visitors and Web sites. This data, as we know, is usually called as *clickstream* or usage data. The WUM aim is to discover relevant usage patterns, to understand and satisfy what a site visitor wants, as an insight to improve site user friendliness and effectiveness levels. This knowledge is quite useful to support decisions on several application areas, such as Web personalization, business intelligence, site restructuring and content alteration and system performance improvement [20]. The basic intention behind all of this consists of catching and keeping attention from visitors to the promoted contents, in order to reach the goals established for the web sites by their managers and administrators.

Data Mining (DM) and WUM tools are becoming increasingly important to a variety of users with different levels of knowledge in the area. Indeed, any user inside the organization can be, in general terms, an informal analyst. Nevertheless, such tools are usually too complex to be used without the aid of a specialist in the area. A common known challenge is related with the selection of suitable methods to apply on a specific data analysis problem, in order to improve the quality of results for a particu-

P. Perner (Ed.): ICDM 2006, LNAI 4065, pp. 161 – 175, 2006.

lar goal defined for a specific site. This challenge is the main motivation of our work, which aims at promoting a more effective, productive and simplified exploration of such data analysis potentialities. The way defended to achieve this goal consists in assisting the development and the application of DM processes, using the experience acquired in the past when we solved similar problems applying successful mining processes. This is a typical strategy based on the principles that we recognize that *Case Based Reasoning* (CBR) presents to us [1,11,19]. Based on such principles, we designed and implemented a prototype recommendation system, which is able to propose the most suited mining plans to a specific *clickstream* data analysis problem, given a high level description of the problem. The case based representation models can also act as exploration and sharing bases over knowledge repositories, promoting sustained learning and best practices adoption involving usage data exploitation.

Assisting decisions within DM applications and knowledge discovery processes is not a new initiative. There are some that explore also the CBR paradigm to undertake related purposes. The Mining Mart project [16], for instance, represents several efforts devoted to the reuse of successful data pre-processing processes, appealing to a case based metadata repository. However, this system does not explore the meta-model potentialities neither the typical CBR methods to help users on establishing the mapping between the problem that we have and the stored ones. Moreover, this project is centred in pre-processing activities, not in DM processes. Another example is the METAL project [14], which involved multiple research and development initiatives, some of them based on the CBR paradigm (e.g. [6,13]). Generally, the main aim of these initiatives was to assist users in model selection – one of the steps of a conventional knowledge discovery process –, and they focused mainly on the algorithms selection issue, within regression and classification problems. Conversely, our work has a different perspective and scope. The system implemented previews assistance on DM models selection, comprising diverse DM functions, and covers support to processes involving transformation operations and multiple stages, according to real-life applications requirements. Besides, the intended DM task specification reaches a greater level of abstraction. This paper explains our view of the mining methods selection challenge, discussing how it may be handled exploring the CBR paradigm and the implemented system. We also describe the system developed, identifying the main issues faced to fulfil the establish requirements and the approaches followed to accomplish this task.

2 The Challenge

Selecting the most suitable methods to apply on a specific data analysis problem is an important and known challenge of DM and WUM processes development. *Clickstream* data is a very rich and valuable source of information. It captures every trace of the interaction process, allowing revealing the behaviour of Web users, besides the traditional elements of the performed transactions. Potentially, this data can provide enormous discoveries, and the insights can easily be turned into actions [3]. Though, usage data brings up new issues. A huge amount of labour-intensive pre-processing is required to prepare it for mining. Even so, extracting meaning from this data is very difficult, due to their subtle nature, intrinsic complexity and large volume and number

of variables. Koutri et al [12] provide a survey very useful to explain the faced challenge. They discuss the major WUM techniques that can be applied for building adaptive hypermedia systems, identifying:

- the most prominent DM functions (clustering, association rule and sequential pattern mining);
- three specific aspects of adaptation, as the result of Web usage patterns application [15] (personal recommendation, dynamic adjustment and static page/site adjustment);
- some important types of usage patterns revealed from WUM (clusters of Web documents references, clusters of user visits, associations among Web documents and sequences of frequently accessed documents).

The above pattern types were also correlated with the identified adaptation aspects, describing the interpretation of each pattern type and the kind of decision support provided, in order to point out the most appropriate context(s) of each pattern use. In addition, the usage pattern types were compared based on the involved requirements and precision levels. For instance, the sequential patterns were considered the most accurate, since they are more informative, but they require richer datasets to capture the diversity of the behaviour of Web users and being the most difficult to obtain.

We described a few issues involved on one type of WUM application (or three more specific types of adaptation applications). The challenge increases by other kinds of issues faced in real-life scenarios. Among them are the requirements of identifying precisely the business problem, transforming some data to answer the problem and the technical understanding of the mining methods. Usually, DM and WUM tools provide a reduced and abstracted offer, as a set of available DM models. Yet, each model's characteristics constraint its applicability and impose distinct configurations. Furthermore, individual models of the same functions are, usually, more suited for distinct purposes. For example, within classification function, decision trees are descriptive models, being more appropriated for interpreting purposes than neural network models. So, if it is more important to understand the influent factors, than to develop an accurate predictive model, the analyst would prefer the former model.

Our real challenge is to find out some ways to empower analysts, in the sense they are able to serve themselves [3]. The followed approach consists in providing a system with the ability to assist them in two different ways:

1. organizing and storing on a shared repository the examples of successful WUM processes;
2. proposing the mining plans most suited to one *clickstream* data analysis problem, given a high level description of the problem.

Examples of past successful solved problems might be the most helpful and convincing form of aid in this scope. They may: (i) simplify the underlying complexity, providing at the same time the details of a tested and solved situation; (ii) yield context information, making possible to report the solutions along with the respective justifications and obtained discoveries; (iii) promote the mapping of the current problem, against the existent ones, when a more direct form of reuse is not possible. In fact, a straight reuse of one solution is quite possible, since recurrent problems are common on this domain. In addiction, the system can be incrementally improved by

adding new experiences. This particularity is of great importance, as new solving approaches, DM models, application areas, and WUM problems, are always coming up, being hereby automatically integrated.

To better show our point of view, we consider a typical and simple WUM example problem. This problem is centred on obtaining feedback about how visitants are using a Web site, to improve navigation convenience. Namely, the intended action is to add relevant links between some Web pages which are visited together. The analyst wants to find out which Web pages are the best ones to include as links and within which pages, tacking into account pages' importance from the visitants' point of view. In the following sections this example will be further developed, in terms of its treatment exploring the implemented system.

We previewed two exploitation scenarios for our system: the exploratory and the problem solving ones. The former involves the use of the system to gain insights about features of interest, typically through an incomplete description of the problem. For instance, the analyst might wish to know what kind of goals and intents or other categorizations have been used to describe related experiences, with the purpose to learn how to better specify the current problem. The analyst might also wish to find out which data elements or sources were used on similar problems, to decide whether usage data is enough (and on what granularity) or other related data must be integrated. Conversely, the problem solving scenario supposes the knowing of the current problem and its submission to the system using a more focused description, in order to obtain more selective solutions.

3 The Mining Plans Selection System

A CBR application can be described by a cycle comprising four processes, usually assigned by the four R's [1]: (i) retrieve cases similar to the current problem; (ii) reuse the information and knowledge of the retrieved case(s) to solve the problem at hand; (iii) revise or adapt the proposed solution to better fit the current problem, if necessary; (iv) retain the confirmed experience parts that might be useful in future problems solution. A case is usually viewed as a problem specification and a described solution of this problem. In the current scope, a case represents and describes one knowledge discovery process, in terms of two factor sets having higher influence in this scope: (i) analysis requirements and characteristics of the target dataset; (ii) experience about the application of DM functions and models and other operations. The items belonging to the first factor set define the CBR problem description, being useful descriptors to retrieve similar cases. The remaining factors belong to the data analysis problem solution, being retained and used to produce a solution description.

Fig. 1 shows the main functional components of the system and their interconnections, inputs and outputs. The CBR engine is the system's core component and the one that performs the inference processes. Currently, this component implements the retrieve and the retain processes and uses a Database Management System (DBMS) to manage the acquired knowledge - other CBR processes will be considered in future work.

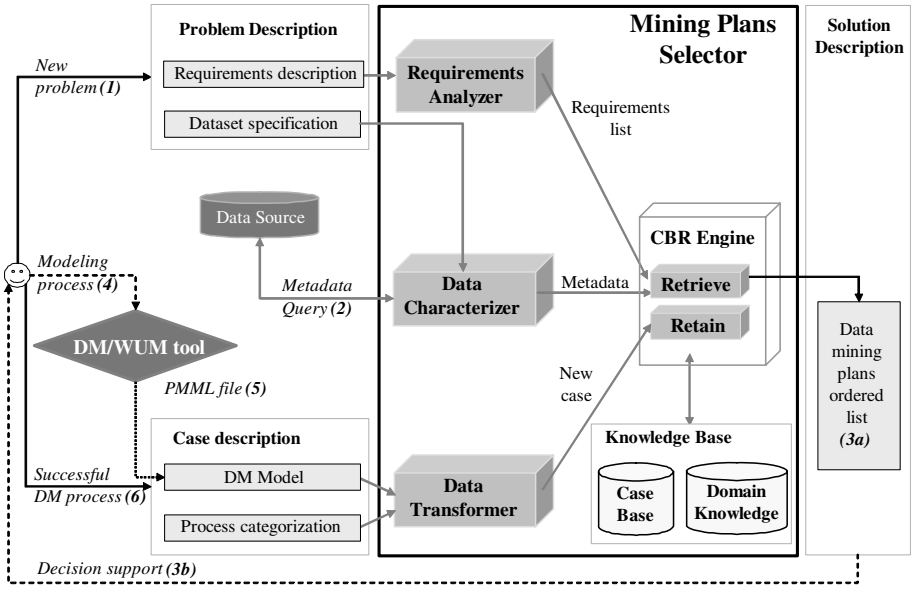

Fig. 1. The system's components

The most typical use of the system can be described by the following steps:

1. the analyst enters a new data analysis problem (1) through the dataset specification and the analysis requirements description;
2. the **data characterizer** component analyses the dataset and extracts the most relevant metadata (2), to the purpose of WUM processes selection;
3. the **requirements analyzer** component handles the analysis requisites description, to get and systemize the embedded constraints;
4. the **retrieve module** matches the incoming descriptions of the new problem against the (potential useful) existent cases, to find out the most similar cases;
5. the most similar cases are organized and then presented to the analyst, providing a suggested solution description, which comprises an ordered list of the most suited DM plans (3a);
6. the DM plans are reused to assist the analyst (3b), which develops and submits the DM process (4), appealing to a DM or WUM tool;
7. if the DM tool supports the Predictive Model Markup Language (PMML) standard [17], it can provide a DM model description through a file in the PMML format (5);
8. a successfully DM process (6) may become a new case to retain;
9. the **data transformer** component analyses and assembles the data of the new case specification, including the DM model description (e.g. PMML files) and the process categorization (complementary descriptions);
10. the **retain module** is then evoked to structure and store the new case, finishing the cycle.

The system was implemented as a Web application seated on typical client/server architecture, with three layers of services: interface, business and data. The options concerning the implementation were based on appealing, preferentially, to free software with open code and multi-platform, and to accepted standards, *Application Program Interfaces* (API) and the packages implementing these API. The technologies applied on the client side consisted, mostly, in HTML [21], supported by *Cascading Style Sheets* (CSS) [5], for the formatting, and by programming in JavaScript, for submission validation and browser behaviour and user interaction enhancement. The server side of the application was developed on Java environment, using the *Java 2 Platform Standard Edition* (J2SE 1.5.0) [7]. The business logic is in charge of Java components and the interface services employ the *Java Server Pages* (JSP) specification [10]. The publication and deployment of these services is assured by the JSP/Servlets container Apache Tomcat (version 5.5) [4]. The data services, developed on the Java platform, were implemented exploring different API, to support and abstract the access to different data sources. The *Java Database Connectivity* (JDBC) [9] protocol and API was used to deal with relational data sources access and manipulation. For XML/PMML document processing one used the *Java API for XML Processing* (JAXP DOM/SAX) [8] and the Crimson processor (Parser).

4 Mining Problems Description

Problem description involves the specification of the current problem's characteristics and a set of constraints, based on the analysis nature and on the analyst's preferences. The main elements of a WUM problem description are shown on Fig. 2 (in bold), along with the specification of some values of the example problem (referred before). The target dataset consists in *clickstream* data (a server log file), describing information (8 variables) at page view/access level (granularity). One important issue is to capture the relevant dataset properties to the particular purpose of DM methods selection. A common data characterization approach builds upon general measures, statistical (numerical attributes) and theoretical information (symbolic attributes) [13]. This approach has been frequently and successfully used in Meta-Learning, to select adequate learning algorithms. Though, those measures are numerous, complex and diverse and the proposals about their content have been used for a subgroup of DM functions. To accomplish datasets characterization we identified a simple set of descriptors, involving metadata automatically extracted by the system and properties values indicated by the analyst. These descriptors are of two main types: (i) DM generic characteristics, collected at dataset level and at individual variables level; (ii) WUM specific characteristics, obtained almost all at dataset level, except the variables' semantic category. The system also provides means to indicate the relevant items or the desirable proprieties, in terms of the dataset variables.

The major requirement regarding the WUM task is to support high level descriptions, through abstractions related to the real problems to solve. A description based on DM functions (or models) cannot abstract the complexity and might exclude processes involving other alternative functions (or models). So, this specification relies on the data analysis goals, which have two distinct perspectives in real-life applications: the business and the WUM ones. The business point of view is meant as the analysis

intention or possible uses of the discoveries, being assigned by application area. The WUM perspective stands for the mining result type to get and the sort of analysis approach to explore, in order to satisfy business goals. This point of view is called goal, since it reflects a kind of WUM problem, and implies that each new case to retain must be related to a specific goal and to one or more application areas. To allow several levels of subdivision in more specific sub-areas, and its definition according to organization needs, the application areas are organized in a dynamic hierarchy. Thus, the DM task specification becomes the selection of the most relevant goal(s) and application area(s), in levels of detail closer to the problem to solve.

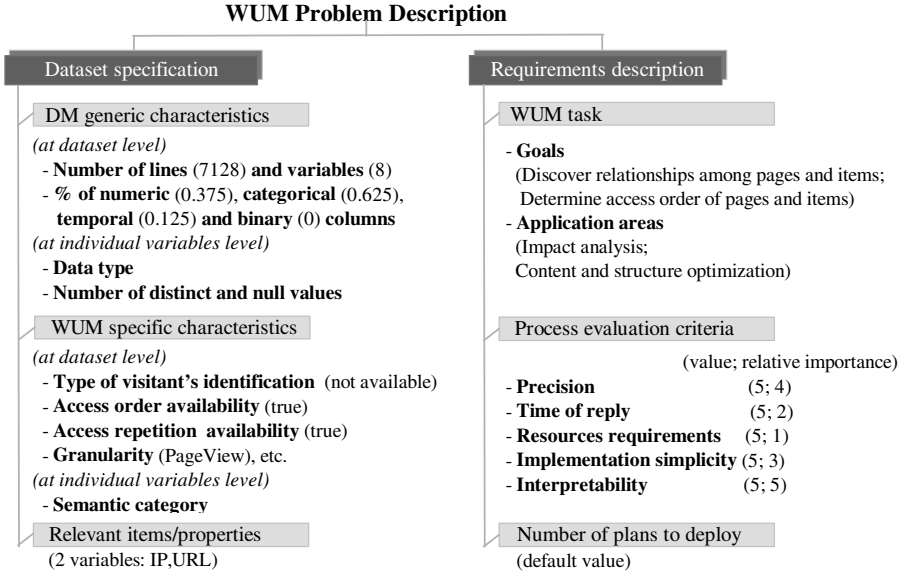

Fig. 2. Main elements of a WUM problem description

At the moment, our case base contains a small number of cases. This fact and the (current) intentional generality restrict the diversity of goals and application areas. Namely, the existent hierarchy of application areas comprises only two levels with three top areas (including few sub-areas): (i) adaptability (e.g. dynamic adjustment and personalization); (ii) business intelligence (e.g. profiling & targeting and initiatives evaluation & planning); (iii) quality of service (e.g. impact analysis and content & structure optimization). For instance, the quality of service area focuses actions of systems and Web site improvement, involving basic expectations and affecting all the visitants. Therefore, we selected the two exemplified sub-areas of the quality of service area to define the example problem, since they are both relevant and the closest ones to the intended actions. In terms of goals we used all the ones that might provide information about relationships among pages. Other existent goals include Distinguish visits based on target events and Identify & characterize different types of visits and visitants.

Concerning the evaluation criteria of DM processes, we adopted the usual performance indicators. As the most promising, the ones shown on Fig. 2 had been chosen. To simplify the specification of the intended evaluation criteria and to deal with its subjectivity, we establish an ordinal and limited scale for all indicators ([1-5]), where greater (>) always means better. Moreover, the values specified by the analyst are meant as lower bounds. In other words, what matters is a value being lower (worst) than the searched one. So, the analyst might describe what he considers acceptable, defining the lower bounds of the relevant indicators and imposing priorities among them through relative importance. Within the specification of the example problem we assigned the value 5 (maximum) to all the evaluation criteria and the greatest relative importance to interpretability followed by the other ones, as reported in the figure.

Finally, the problem description supports the specification of exact filtering criteria and descriptors importance levels, to enable the improvement of the problem specification. Furthermore, the analyst may exclude descriptors from the specification, describing only the values of the relevant (or known) ones. As the dataset metadata attributes are in majority, by default the system selects the processes with the most similar datasets. In fact, the dataset characteristics are always a crucial (predictive) factor, since models properties and assumptions, and even other factors (e.g. goals), frequently, demand for some specific data. We will explore this default mode to solve this problem, but applying the functionality of exact filtering to the goal descriptor, in order to use, at least, one of the possible ways of focusing our description. The results of the example problem specification are discussed in the following section.

5 Presenting a Mining Solution

A pertinent faced issue is how to organize the retrieved cases within the solution description output, tacking into account their utility to the analyst. The approach followed basis this organization on the case's model category. By model category we mean a representation of each distinct combination of (one or more) DM models – the most important applied methods – occurring among the stored cases.

Fig. 3 illustrates a (small) possible solution description for the example problem. The figure presents the mining plans of three model categories (column D), instantiated with the most similar case of the category (column A). The column (B) shows the similitude between the target and each instantiated case. The hyperlinks of the cases (A) provide direct access to the respective detailed information. The combo boxes (A) show the similarity with the remaining retrieved cases of the model category (expanded on E). These combo boxes allow to access further information about such cases, through the selection among the available options. The column C depicts the average values of each evaluation criteria, respecting to all the retrieved cases of the model category. The interpretability criterion is the first one, because previously, we gave it the greatest relative importance. Using the described organization, the analyst can see several alternative solutions of the same problem, as well several instances of one solution of particular interest. Hereby, we maximize the solutions utility simultaneously for two distinct purposes: diversity of alternative solutions and variety of instances of a solution of particular interest.

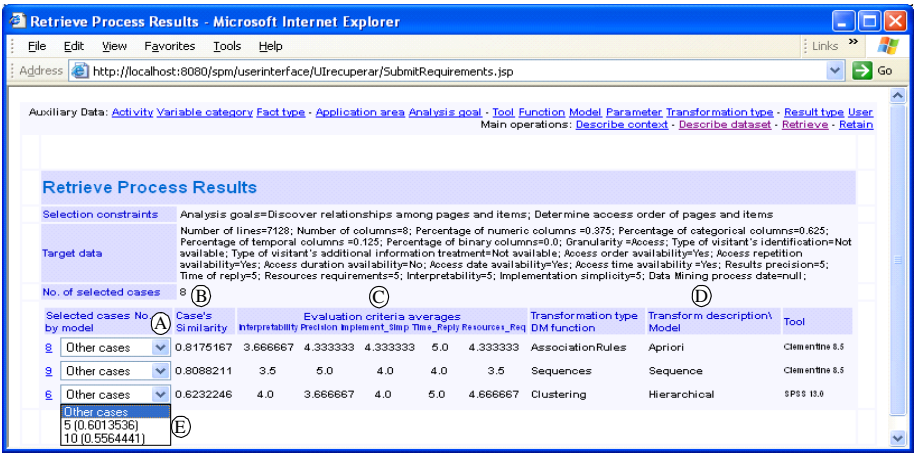

Fig. 3. Example of a description for a solution

All the suggested models are suited to the problem at hand. The association rules option is a good compromise between precision and coverage: it is more informative and precise (e.g. provides rules and the respective support and confidence) than the hierarchical clustering; generally, it provides better overall coverage than the sequence model, although being less informative than such model, which yields more fine-grained information (e.g. ordering among accessed pages). As expected, the system (on the default mode) gives emphasis to the similarity between datasets. The similitude of the cases from the hierarchical clustering is substantially inferior, since this model was applied to datasets with very different properties (e.g. binary matrix of pages × sessions). Conversely, the analyses from cases 8 and 9 were performed using datasets similar to the target one. However, the inclusion of the hierarchical clustering model within the solution is useful, since it is possible to transform the target dataset into the format commonly used to explore this model.

The strategy to undertake the retrieve process comprises the following major steps:

1. cases pre-selection, given the exact filtering criteria;
2. similarity estimation between the cases pre-selected and the target;
3. cases grouping by model category and determination of the evaluation criteria averages;
4. deployment of the firsts K groups, ordered (on first place) by the greatest similarity within the group and (on second place) by the evaluation criteria averages of the group.

Step 1 selects the WUM processes applicable on the current problem. Step 2 evaluates the proximity level of each retrieved case in relation to the target, pointing out the processes potentially more effective. Step 3 provides a global evaluation perspective of each model category, and, finally, step 4 allows the presentation of the K most promise mining plans, according to the similarity level and the model category evaluation criteria, which is most relevant to the analyst.

The similitude of each pre-selected case's problem to the target one is computed considering the correspondent feature values and the adopted similarity measures. The similitude assessment approach devised over WUM problems comprises the modelling of the following types of measures: (i) local similarity measures for simple and complex (multiple-value) features; (ii) global similarity measures defined through an aggregation function and a weight model. The global similitude combines the local similarity values of several features (e.g. through a weight average function), giving an overall measure. The local similarity measures are defined over the descriptors and depend mainly on the features domain, besides the intended semantic. Concerning simple (single-value) features, the local similitude of categorical descriptors is essentially based on exact matches (e.g. for binary attributes) or is expressed in form of similarity matrices (e.g. for some symbolic descriptors), which establish each pairwise similitude level. To compare numeric simple features, we adopted similarity measures mainly based on the normalized *Manhattan* distance.

We also need similarity measures for complex descriptors, modelled as set–value features, containing atomic values or objects having themselves specific properties. Indeed, this need was the main issue faced under the similarity assessment. For instance, it appears when matching the variables from the target and each case. We have to compare two sets of variables, with inconstant and possibly distinct cardinality, where each variable has its own features. There are multiple proposals in the literature to deal with related issues. Even so, we explored a number of them and the comparative tests performed lead us into tailored (extended) measures, better fitting our purposes.

6 Describing a Mining Experience

As shown in Fig. 4, case description comprises the DM model and the process categorization. This subdivision is justified by the intent to support the data model submission using files in PMML format. PMML [17] is a XML-based standard which provides a way to define statistical and DM model and to share them among PMML compliant applications. This standard is supported by a high and raising number of DM tools, even if with some limitations (e.g. versions supported). So, it represents an opportunity to automate some data gathering processes. Yet, it is necessary to obtain other data elements, about items unavailable in PMML files (e.g. configuration parameters and transformation operations), being required to provide a complementary form of data submission. Furthermore, the PMML file may not be available.

Despite we only show the most important elements, one concern was to capture a wide characterization of each WUM process, since it is essential to store the specific context required to find, interpret and evaluate the solutions. The DM model represents the modelling stages of the processes, where each instance comprises the major elements of the modelling description, extracted from PMML files or obtained directly from the analyst. A modelling stage involves the application of a DM model, belonging to a DM function, appealing to a DM tool, as well the configuration of a particular set of parameters and the use of a set of variables performing different roles. Categorization represents complementary information about the WUM processes, specifically, the data elements which can not be extracted from PMML files.

The dataset item includes the elements previously discussed, collected during the problem description specification. The transformation operations item respects to data preparation stages, described mainly in terms of the used tool, type of operation (e.g. derive new variable) and the set and roles of the involved variables (e.g. input and output variables). The discoveries item concerns to results provided by the process. Finally, the process classification regards to its categorization in terms of features such as evaluation criteria, application areas and analysis goals.

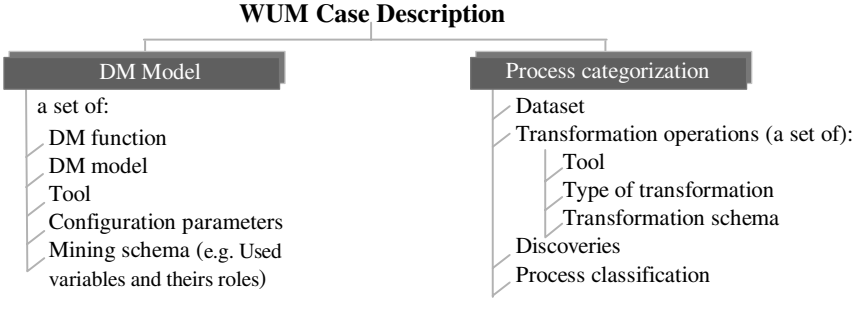

Fig. 4. Main elements of a WUM case description

7 Knowledge Representation

In CBR systems the primary kind of knowledge is contained on the specific cases, stored and organized in a case base. While traditionally viewed as data or information, rather than knowledge, concrete descriptions of past problem solving episodes became knowledge for CBR methods, since these methods are able to use cases for reasoning [2]. Besides case's specific knowledge, other and more general types of knowledge may be integrated within the CBR process, with varying levels of richness, degree of explicit representation and role [1]. Richter [18] introduced a model that identifies four knowledge containers in CBR systems: 1) vocabulary; 2) similarity measures; 3) solution transformations; and 4) case base. The first three containers represent compiled (more stable) knowledge, while the cases contain interpreted (dynamic) knowledge. This knowledge container view received wide acceptance, becoming the natural approach for knowledge representation structuring in CBR systems.

Our system's knowledge base integrates two components: the case base and the domain knowledge. The case base consists in a metadata repository, supported by a relational DBMS (involving about forty tables), where each case represents a successful DM process description. The domain knowledge concerns to knowledge about this particular scope of application, covering items as the specification of concepts and attributes (namely of problem description), mostly in terms of properties required to interpret, compare and retrieve the cases. This component provides knowledge items belonging to the vocabulary and similarity measures containers, since our system does not transforms solutions. The former container includes, for instance, items which define descriptor types and domains and establish the mappings between the relational

schema and the case descriptors. The knowledge involved in the retrieval and comparison approaches (e.g. weights and similarity functions) is held by the second type of container. Fig. 5 shows an excerpt of the cases representation conceptual metadata model, using a class diagram in *Unified Modeling Language* (UML) simplified notation.

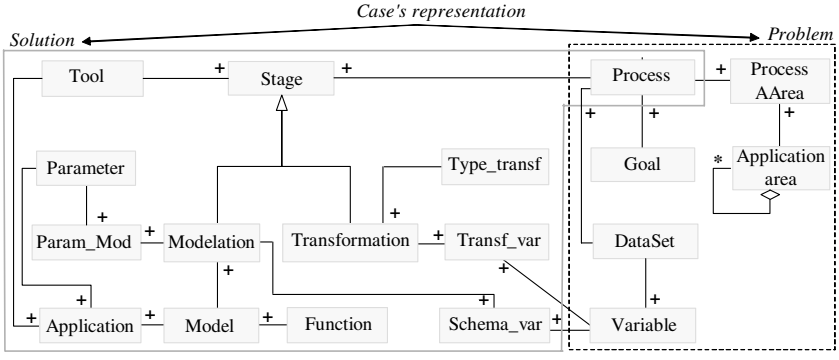

Fig. 5. Excerpt of the cases representation conceptual metadata model

The main components of a case's description are the following ones:

- *Process* - central class that represents an individual DM process and establishes the connection between classes concerning the problem and solution description. This class includes attributes that describe and characterize each process (e.g. evaluation criteria).
- *DataSet and Variable* – classes that embody the dataset and the variables characteristics.
- *Stage*, *Transformation* and *Modelation* - classes regarding to a DM process development phase; the superclasse Stage represents the shared parts or properties of each phase, while the Transformation and Modeling subclasses concern to the specific ones.
- *Model and Function* - classes that establish categorizations of the DM models and functions provided by tools.
- *Goal, Application_Area and Process_AArea* - classes covering, respectively, the analyses goals, the hierarchy of application areas and the association between processes and Application_Area.

The conceptual model presented provides the support to attend the established requisites, although being in continuous refinement. One of the refinements accomplished was the inclusion of the context of the datasets and the facts related to them and to the DM processes. This refinement intends to improve the cases description and to minimize the data redundancy. Other extensions consists on the inclusion of classes about sources (e.g. dataset and PMML files and database tables used), DM processes authors, DM processes discoveries and theirs relationships with existent facts.

8 Conclusions and Future Work

The WUM exploration is one important instrument to organizations involved in Web sites optimization and truly concerned on achieving their goals. Web site design, administration and improvement are complex tasks, which demand deep decision support, reaching an increasing ample and diversified population of users. However, DM and WUM tools and their concepts and techniques are too complex to be effectively explored without the aid of specialists on the area. The developed work aims at contributing to a more simplified, productive and effective exploration of WUM potentialities. As referred before, the main idea is to assist analysts through examples of solved similar analysis problems, promoting the reapplication of its best practices in the current situation. This approach seems to be the most opportune, according to accepted facts related to these processes nature. In DM and WUM domains, recurrent problems and methods repetitive use are quite common. Additionally, the experience and acquired know-how have a prominent value. Besides, examples of successfully solved problems are the most useful and convincing form of aid. To achieve this aim, we implemented a prototype system, which should suggest the mining plans more adjusted to one *clickstream* data analysis problem, given the respective description. This system is also based on abstractions related to the real problems to solve, meaning that it could serve the particular needs of less knowledge analysts, who wish to learn how to handle a concrete problem, being also useful to specialists interested in reminding and reusing successful solutions, instead of solving the problems from scratch.

The decision support involving discovering processes is an important working area, being, thus, the focus of multiple research projects. The CBR paradigm exploration is used too in efforts devoted to analogous purposes. Though, the work developed can be distinguished from the main related work on several features, namely, the support of multiple stage processes, the extended aid involving different DM functions selection, the integration of transformation operations (even if simplified) and, primary, the attempt to reach high abstraction levels in the intended DM task specification. Additionally, the system proposed is particularly devoted to the specific WUM domain and previews support over realistic exploration scenarios.

In this paper we described the system that we implemented to fulfil requirements and goals that we presented before, giving emphasis to their main characteristics and to the vision of its practical use to assist some steps within a WUM process development. A key factor to the system efficacy is a coherent and well structured definition of the analysis goals and application areas descriptors. The approach provided to support theirs definition and use is simple, flexible and effective. One drawback to point out is the treatment of the analysis goal descriptors, which is only suited to a moderate number of items. Even so, the potential of the approach has not been explored. Greater level of abstraction might be achieved developing further these descriptors and, thus, a better support is a possible future direction of work. This may be realised through some form of goals grouping, namely, an overlapping one. Additionally, the system has been tested using a small sample of simple WUM processes. In fact, the most exhaustive tests performed concern to the comparison between datasets and they point to the efficacy of the system. As previously mentioned we conducted a comparative study over several similarity measures and already integrated the ob-

tained results within our system. Nonetheless, the activities concerning the preparation of more cases, comprising WUM process with higher complexity are still occurring. Afterwards, a more systematic evaluation of the system becomes possible and necessary. Furthermore, other planed and related activity is to explore additional data mining algorithms, specifically, approaches able of better fitting the properties of Web usage data, preferentially appealing to free software.

Acknowledgments. The work of Cristina Wanzeller was supported by a grant from PRODEP (Acção 5.3, concurso nº02 /2003).

References

1. Aamodt, A. and Plaza, E.: Case-Based Reasoning: Foundational Issues, Methodological Variations and Systems Approaches. In Artificial Intelligence Communications (AICom), IOS Press, Vol. 7, No 1 (1994) 39-59.
2. Aamodt, A.: Knowledge Acquisition and Learning by Experience - The Role of Case Specific Knowledge. In Machine Learning and Knowledge Acquisition, Academic Press, Integrated Approaches (1995) 197-245.
3. Ansari, S., Kohavi, R., Mason, L. and Zheng, Z.: Integrating E-Commerce and Data Mining: Architecture and Challenges. In Proc. 2001 IEEE International Conf. on Data Mining, IEEE Comput. Soc. (2001) 27–34.
4. Apache Jakarta Tomcat. http://tomcat.apache.org/. Access April 2006.
5. Bos, B.: W3C. Web Style Sheets – Home Page. http://www.w3.org/Style/. Access April 2006.
6. Hilario, M. and Kalousis, A.: Fusion of Meta-Knowledge and Meta-Data for Case-Based Model Selection. In Proc. of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD '2001), Springer (2001) 180-191.
7. Java 2 Platform, Standard Edition (J2SE). Sun Microsystems. http://java.sun.com/javase/index.jsp. Access April 2006.
8. Java API for XML Processing (JAXP). Sun Microsystems http://java.sun.com/webservices/jaxp/. Access April 2006.
9. Java Database Connectivity, JDBC Data Access API. Sun Microsystems. http://www.javasoft.com/products/jdbc/index.html. Access April 2006.
10. Java Server Pages. Sun Microsystems. http://java.sun.com/products/jsp/. Access April 2006.
11. Kolodner, J.: Case-Based Reasoning, Morgan Kaufman, San Francisco, CA (1993).
12. Koutri, M., Avouris, N. and Daskalaki, S.: A Survey on Web Usage Mining Techniques for Web-Based Adaptive Hypermedia Systems. In S. Y. Chen and G. D. Magoulas (eds.), Adaptable and Adaptive Hypermedia Systems, Idea Publishing Inc., Hershey (2005).
13. Lindner, C. and Studer, R.: AST: Support for algorithm selection with a CBR approach. In Proc. of the 3rd European Conf. on Principles of Data Mining and Knowledge Discovery (PKDD'1999), Springer (1999) 418-423.
14. MetaL project http://www.metal-kdd.org/ Access April 2006.
15. Mobasher, B., Berendt, B. and Spiliopoulou, M.: KDD for Personalization. In PKDD 2001 Tutorial (2001).
16. Morik, K. And Scholz, M.: The MiningMart Approach to Knowledge Discovery in Databases. In N. Zhong and J. Liu (eds.), Intelligent Technologies for Information Analysis, Springer (2004).

17. Predictive Model Markup Language. Data Mining Group. http://www.dmg.org/index.html. Access April 2006.
18. Richter, M.: The Knowledge Contained in Similarity Measures. (Invited Talk) at the First International Conference on Case-Based Reasoning, ICCBR'95, Lecture Notes in Artificial Intelligence 1010, Springer Verlag (1995).
19. Riesbeck, C.K. and Schank, R.C.: Inside Case-Based Reasoning. Lawrence Erlbaum Associates, Hillsdale, NJ, US (1989).
20. Srivastava, J., Cooley, R., Deshpande, M. and Tan P.-N.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. In SIGKDD Explorations, Vol. 1, No 2 (2000) 1–12.
21. W3C HTML Working Group. HyperText Markup Language (HTML) – Home Page. http://www.w3.org/MarkUp/. Access April 2006.