Selecting clickstream data mining plans using a case-based reasoning application

C. Wanzeller[1] & O. Belo[2]
[1]*Departamento de Informática, Instituto Superior Politécnico de Viseu, Viseu, Portugal*
[2]*Departamento de Informática, Escola de Engenharia, Universidade do Minho, Braga, Portugal*

Abstract

Despite the increasing interest and wide use of data mining tools, extracting useful knowledge from the growing available data remains a complex task. The Web environment and clickstream data demand even more efficacy to knowledge discovery: such data is a rich, complex and huge source of information; the discoveries are easily actionable; and time constrains are typically hard. One general, crucial and very challenge issue of knowledge discovery is the selection of the right methods to apply according to the nature of the problem under analysis. This issue, particularly within the Web usage mining scope, is the focus of our work. This paper describes a case-based reasoning system especially oriented to assist users in the development and application of Web usage mining processes. This system takes as inputs the characteristics of the available data, the analysis' requirements, and, based on acquired experience from successfully processes, delivers a solution: a mining plan suited to the data analysis problem at hand.

Keywords: Web usage mining, case based reasoning, clickstream analysis, data mining assistance and data mining plans selection.

1 Introduction

Today, organizations are adapting their way of being in the electronic commerce arena. Some of them try hard to know how users act when visiting the sites they promote and sustain. Basically, they intend to gather information about the impact and effectiveness of their sites without questioning directly users. Regular

WIT Transactions on Information and Communication Technologies, Vol 37, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060231

visitors do not like to be faced with questionnaires about their preferences. They simple want to see what sites offer to them and, if possible, use it. So, the question is, how can organizations do it?

During the last years many techniques were used with significant impact and utility, contributing to know more about the behaviour of sites' visitors. However, *as we know,* Web Usage Mining (WUM) has been one of the most important and effective that was used by organizations in the establishment of behaviour patterns for Internet sites. It is not a surprise, since it provides us the means to extract valuable information (we may say in some sense, knowledge) from visitors' interaction process data. However, the complexity of WUM tools is often too high, which difficult a lot their correct exploitation and application. In order to attenuate this situation, we decide to design and develop a system with the ability to recommend adequate methods to explore site navigation data according the experience acquired by specialists in the past.

Basically, we intend to catalogue and store WUM past experiences in a specific oriented knowledge base that can be used latter by the system doing new mining plans recommendation that can be applied over sites' navigation data. This is not a novel approach. This problem solving strategy has been used for long time in Case-Based Reasoning (CBR) applications. Nevertheless, the combination of WUM and CBR techniques discovering navigation patterns it is not a conventional approach or an old fashion strategy, by the contrary. Such combination provides us new forms of approaching WUM problems and simplifies drastically the need for expertise in this so special application area. Users will have their WUM tasks simplified and will reach, potentially, new exploitation results. However, this approach has some real limitations: the unavailability of sufficient processes' descriptions, properly structured and applicable, according to the organization's own needs. Additionally, it requires an effective and flexible mechanism to help the matching of the current problem with the most promising mining processes. Even experts cannot provide comprehensive and reliable rules (or cases) for problem solving. Regarding this CBR applications often produce high quality results. Building up application cases by itself holds considerable benefits, realized by structuring and memorizing the knowledge acquired from experience, even when the solutions are not easily reusable. CBR methods favour a flexible similarity-based comparison, even if the involved features are not objective and precisely defined. Moreover, the possibility of considering only the relevant features and using specific importance levels increase the potentialities of answering the real user needs. In addiction, CBR is a sustained incremental learning approach, given that a new experience can be automatically integrated, each time a problem is solved, becoming immediately available to apply on future problems [1]. This aspect is of great importance, due to the constant evolution of this domain and the need to incorporate new knowledge about algorithms, tools and problems.

In this paper we describe the major characteristics and functionalities of the referred system in two different ways: (i) organizing and storing on a shared repository the examples of successful WUM processes; (ii) proposing the mining plans most suited to one *clickstream* data analysis problem, given a high level description of the problem.

2 Towards a more suitable WUM strategy

In general terms, WUM problem descriptions comprise data characteristics and analysis requirements. As any other Data Mining (DM) process, WUM involves as well identifying the proper *Data Mining* (DM) function (e.g. classification), choosing the DM model(s) or algorithm(s) (and setting its parameters) [4]. As we know, these tasks are very dependable from the given data and the preference criteria (e.g. interpretability), which must be take into consideration, obviously, if we intend to get some valuable results from a WUM process. Three major issues emerge within these tasks. First, even the former task is not as trivial as it seems to be, since it involves the reformulation of the business problem into a DM problem. Further, some methods (functions and models) overlap in terms of the problems they can solve. Second, the preference criteria may be conflicting and subjective. Third, a deeper technical understanding of the methods and the characterization of the available data are required to meet the application objectives. Each algorithm has specific properties and makes different assumptions on the data. Also, the functional environment is not stable, but actively influenced by several dynamic factors, as part of the development and application process [12]. Among these factors are the transformations that change the data properties, such as the ones crucial to answer the problem and others needed to better fit the methods' assumptions.

Clickstream data is a very rich and valuable source of information. It captures every trace of the interaction process and provides us the possibility to capture the behaviour of the visitor. The benefits of discovering visitors' preferences and navigation patterns are potentially enormous and the insights can be easily turned into actions. However, usage data brings up new issues that affect significantly mining tasks. The immediacy of data gathering, very often without analysis questions in mind, turns its pre-processing into a hard and labour-intensive task. Even so, extracting meaning from this data is very difficult, due to its subtle nature, intrinsic complexity and large volume and number of variables. Despite the great interest on this data, the WUM problem types, the kinds of mining activities and the related practical applications are less studied and structured, increasing even more the challenge of methods selection.

In order to improve more effectiveness in WUM processes and provide more suitable mining plans we decided to integrate knowledge acquired during the execution of previous WUM experiences. The plans will combine the cases that were retrieved, trying to maximize the solution utility essentially for two purposes: (some) diversity of alternative plans and variety of instances by plan. Preceding the systems' description we discuss the mining methods selection challenge. The task of selecting the most suitable methods to apply on a specific data analysis problem is an important and known challenge of the *Knowledge Discovery* (KD) process. But recommending a more suitable mining plan is not a very simple task, since it involves complex processes concerning with mining strategies, algorithms recommendation, and application requisites, just to name a few. A brief looking on current works in the area shows that algorithms recommendation [7, 8] is an active line of research in meta-learning, lying mainly on the algorithms selection issue, within regression and classification problems, following typically a data driven approach. However, as already referred, some other initiatives consider also the influence of application requisites and explore too the CBR paradigm [5, 6].

The algorithms selection issue is important, mostly in classification, but has a limited scope. First, some high level design aspects and initial stages of the application development are not supported [12]. Second, practice problems are often too complex to be handled by a single method. A broader scope of the challenge considers processes development at different levels and involving the application of multiple methods, according to real-life scenarios. This scope covers distinct approaches that may support different KD steps. For instance, the Mining Mart project [9] concerns to the pre-processing step, exploring a case-based metadata repository, while in [3] is proposed an ambitious approach of a user-guidance module to support the entire KD process. The system IDEA [2] offers similar assistance, but following a different direction, based on the fact that it is very difficult to discern the one best plan, as DM results can be unpredictable and users' goals and desired tradeoffs might not be easily or completely specified at the onset of an investigation.

All these issues were considered when we designed and developed the system presenting in this paper. However, we gave more emphasis to the DM step, typically presuming the availability of sources containing pre-processed data, but always considering data transformation operations. Our current focus lies in suggesting the most promising set of plans. We argue that the factors that are uncertain or difficult to specify should not affect the plans content. Instead, all factors should be used to find out and reduce the set of candidate plans, among the excellent existing DM processes, and to support the final selection. A plan is a distinct chain of methods and should be instantiated with examples of such methods application, in order to provide multiple kinds of informative details - context, settings and discoveries. In the next sections we will present a general overview of the mining plans selector system developed so far, giving particular attention to its main architecture's components, knowledge representation, mining problems description, and system's outcomes.

3 The MPS system

The *Mining Plans Selector* (MPS) system was conceived to serve particular needs of less knowledge users, who intent to recognize potential applications of WUM or, simply, 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 system provides a semi-automated approach for knowledge acquisition along the organization, and a comprehensive support for documenting WUM processes, covering the description of:

- sequences of activities along with explanations and justifications, comprising transformation stages and modelling stages based on the PMML standard [10], together with the applied models, used variables and the specific parameters settings;
- processes' context, sources, authors, discoveries and categorizations, obtained from practical situations, and under evaluation criteria used in the area;
- problems defined through the most influent aspects, involving data characterizations, which are wide in terms of the different DM methods but also focused on the specific WUM issues, as well abstractions related to the real problems to solve, which may be establish according to the organization's own particular requirements.

The system was especially oriented to assist users in the development and application of WUM processes, helping them in the establishment of more suitable strategies (mining plans) to mining clickstream data accordingly their organizations' criteria and supported by cases created from other WUM experiences (application cases) performed with success in the past. The mining plans are generated based on the characteristics of the clickstream data, the analysis' requirements (data and problems descriptions), and the cases that it keeps in its own knowledge base.

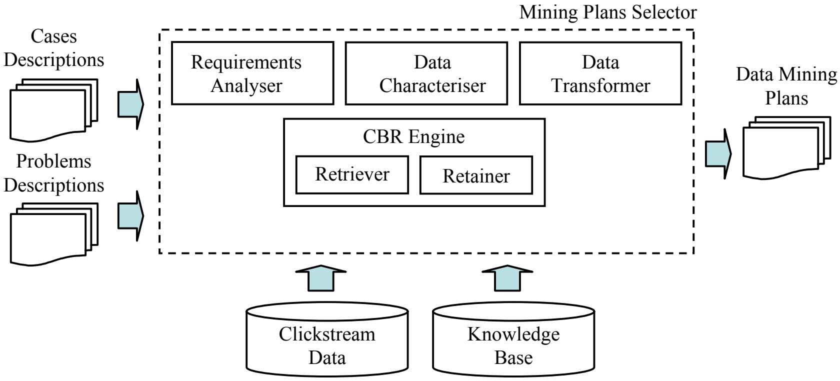


Figure 1: The MPS's functional architecture.

In Figure 1 we can see the basic MPS's functional architecture. It is constituted by four main functional components, which are responsible to perform the most significant tasks of the system, namely the:

a. Requirements Analyser, that is responsible to handle the requirements of the analysis processes, to get and systemize the embedded constraints.
b. Data Characteriser, this component analyses the dataset and collects the most relevant metadata, to the purpose of methods selection.
c. Data Transformer that is responsible to analyse and assemble the description of the cases, including the description of the mining model and the categorization of the process;

d. CBR Engine, which performs the inference and implements the retrieve and the retain processes; this component also implements the retriever and the retainer models, using a conventional database management system to manage the acquired knowledge; the retriever module uses the dataset metadata and the systematized requirements of the new problem to match him against the (potential useful) existent cases, while the retainer is responsible to store new cases resulting from successfully DM processes.

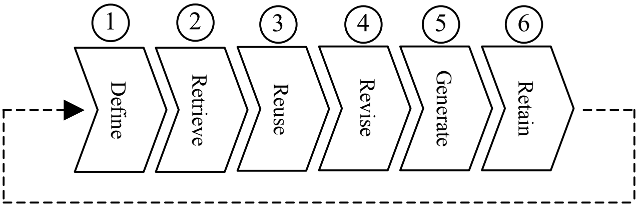


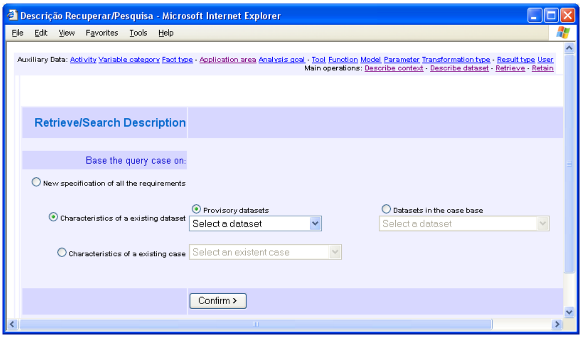
Figure 2: The MPS's life cycle.

The selection of a new mining plan is performed by the system in six distinct steps (Figure 2):

1) Define. A new data analysis problem is informally described, specifying the target dataset (source and additional information) and interacting with the system to provide the explicit requirements (Figure 3). Then, the system extracts and collects metadata, to produce and store a (provisory) dataset characterization, and analyses the requirements, to translate and systematize the embedded constrains, defining a new problem.
2) Retrieve. The incoming new problem, reflecting the inherent and explicit user's constraints and preferences, is compared to the already existing cases indexed by such factors, in order to find out the most similar ones.
3) Reuse. The most similar cases are ranked and organized, according to the level of similarity, the user preferences and the different technical solutions that they represent, forming a suggested solution, holding a ranking of DM plans, which is presented to the user.
4) Revise. The selected plan(s) assist the user during the DM process development using a DM/WUM tool. The user adapts the plans' details to the current situation to achieve a final revised process. If the tool supports the PMML standard, it may supply the model(s) representation of the revised process in such format.
5) Generate. A revised process may become a new case, transforming and combining the PMML file(s) (if available) and the remaining information of the successful process into a comprehensible new case.
6) Retain. The new case defined is structured and stored in the case base, according to its internal schema, finishing the cycle.

As a typical CBR based system, MPS keeps its primary knowledge in the knowledge base, which is organized, essentially, in two main components:

- the domain knowledge, which covers, mostly, the items required to interpret, compare and retrieve the cases, belonging to the vocabulary and similarity containers [11], since our system does not transforms solutions;
- the case base that consists in a relational metadata repository (of about forty tables), holding detailed examples of successful WUM processes, described in terms of the domain problem and the respective applied solution.


Descrição Recuperar/Pesquisa - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Auxiliary Data: Activity Variable category Fact type - Application area Analysis goal - Tool Function Model Parameter Transformation type - Result type User
Main operations: Describe context - Describe dataset - Retrieve - Retain
Confirm >

Retrieve/Search Description

Base the query case on:

○ New specification of all the requirements

● Characteristics of a existing dataset — ● Provisory datasets: Select a dataset — ○ Datasets in the case base: Select a dataset

○ Characteristics of a existing case — Select an existent case

Figure 3: Defining the retrieve query specification.

4 Using MPS in a WUM task

To show how MPS works, we selected the WUM problem of improving visitants' navigation convenience, to be realized by adding relevant links between some Web pages visited together. Basically, we can assume here that our primary goal is to find out which Web pages are the best ones to include as links and within which pages. The target dataset used consists in a server log file, describing data (8 variables) at page view/access level. One key issue is to capture the relevant dataset properties, to the purpose of methods selection. A common data characterization approach, build upon general, statistical and theoretical measures [6], has been often and successfully used in meta-learning. Though, these measures are numerous, complex and are devoted to a subgroup of DM functions. To accomplish datasets characterization we identified a simple set of metadata, involving items automatically extracted by the system and others indicated by the user, comprising: (i) DM generic characteristics, collected at dataset level and at individual variable level; (ii) WUM specific properties. The data characteriser component extracts and collects this metadata, which reflect inherent constraints. The requirements analyser deals with the explicit requisites.

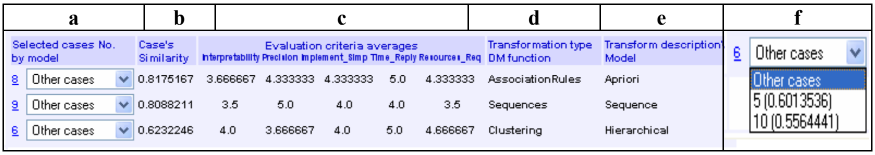
The WUM task must be defined using proper abstractions, as much as possible. A description based on DM functions (or models) cannot abstract the complexity and might exclude processes applying other alternative methods. Thus, it relies on the analysis goals, which has two distinct perspectives:

- Business, meant as the analysis intention or possible uses of the discoveries - assigned application area.

– WUM, standing for the mining result type desired and the sort of analysis to explore - called goal, since it reflects a kind of WUM problem/goal.

Presently, the hierarchy of application areas contains only two levels with three top areas: adaptability, business intelligence and quality of service. We selected two sub-areas of the last one to define the example problem, since both are relevant and the closest ones to the intended actions. For goals we used all the ones regarding to relationships among pages and we applied exact filtering to exclude irrelevant cases. For the process evaluation criteria, we adopted usual performance indicators in this scope. To simplify its specification and to deal with subjectivity, we establish an ordinal common scale for all criteria ([1–5]). The user might describe what he considers acceptable, defining the lower bounds of the relevant items and imposing priorities among them through relative importance. Within the example problem we used the maximum value of 5 as lower bound in all the criteria and the relative importance(s).

The retrieve process strategy comprises four main 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 evaluation criteria averages' determination; (4) deployment of the firsts K groups, ordered (firstly) by the greatest similarity of the group and (secondly) by its evaluation criteria averages. Step 1 filters the applicable processes Step 2 evaluates the proximity level of the retrieved cases, selecting the ones potentially more effective. Step 3 gives a global evaluation of each model category. By model category we mean a distinct combination of (one or more) DM models, occurring in the stored cases. Step 4 allows the presentation of K mining plans, according to the similarity level and the model category evaluation criteria, which is most relevant to the user.



Selected cases No. by model		Case's Similarity	Evaluation criteria averages Interpretability	Precision	Implement_Simp	Time_Reply	Resources_Req	Transformation type DM function	Transform description Model
8	Other cases	0.8175167	3.666667	4.333333	4.333333	5.0	4.333333	AssociationRules	Apriori
9	Other cases	0.8088211	3.5	5.0	4.0	4.0	3.5	Sequences	Sequence
6	Other cases	0.6232246	4.0	3.666667	4.0	5.0	4.666667	Clustering	Hierarchical

6 Other cases
Other cases
5 (0.6013536)
10 (0.5564441)

Figure 4: A solution description excerpt.

In Figure 4 we can see an excerpt of the solution description for the example problem, including the mining plans of three model categories (e) and the respective DM functions (d). Each model category is instantiated with the most similar case (case's number hyperlink on the left side of a). The case's hyperlink provides direct access to its detailed information. The similitude to the target is given on (b). The combo boxes (right side of a) allow to access further information about other cases of the model category, through the selection among the available options. The combo boxes also show the case's number and the respective similarity when expanded (f). The average values of the evaluation criteria are presented on (c) and refer to all the retrieved cases of the model category. The interpretability criterion is the first one, since it was establish as the most important. Using this organization, we maximize the solutions utility simultaneously for two purposes: diversity of alternative plans (multiple model categories) and variety of instances of a plan of particular interest (several instances for each model category).

The system (by default) gives emphasis to the similarity between datasets. The analyses from cases 8 and 9 were performed using the datasets most similar to the target. This fact and the intentional selection of the second goal justify the presentation of the sequential model, with lower effectiveness to the described problem. The similitude of the cases from the hierarchical clustering model is substantially inferior, since it was applied to datasets with very different properties (e.g. binary matrix of pages x sessions). Still, the inclusion of this plan is useful: (i) the model is suited to the problem at hand, although lesser than the association rules model, which is more informative and accurate (i.e. provides rules with support and confidence); (ii) it is possible to transform the target dataset into the format commonly used to explore this model.

5 Conclusions and future work

The proposed and developed work aims at contributing to a more simplified, productive and effective exploration of WUM potentialities. To achieve this aim, we implemented a CBR system to assist users in WUM processes development and application. We believe that such system is a useful tool for corporative knowledge creation, sharing and reutilization, not only from development experiences but also from the obtained discoveries. We previewed two major exploitation scenarios for the system: (i) the exploratory, to gain insights about features of interest, typically through an incomplete description of the problem; (ii) the problem solving, supposing the knowing of the current problem and its submission to the system using a more focused description, in order to obtain more selective solutions.

The system was implemented as a Web-based application. The investment on the interface optimization was considerable, since its quality and robustness became essential to tackle some basic functional requirements. The hypermedia-based interface proved to be very suited to deal with the underlying complexity of the domain, by streamlining access and providing rich navigation features around related information.

We conducted preliminary general and specific evaluation tests. The last ones concerned to a comparative study of several similarity measures, mostly devoted to forms of similarity assessment on sets of values. This need arises on the comparisons of items such as the (sets of) variables from two datasets. This study is out of scope of this paper, but we can say that the system seems to be effective in retrieving similar objects. The general tests performed until now point to the system's effectiveness. By default the system selects the processes with the most similar datasets. This default behaviour is considered a good result. In fact, the dataset characteristics are always a crucial factor. Moreover, we provide means to refine the problem description (e.g. exact filtering, importance levels). However, the system has been tested using a limited sample of (simple) WUM processes. For the future we plan to further evaluate this implementation, which might be realized through the preparation of more (complex) cases and in the context of a study case based on a concrete target organization.

Acknowledgements

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. *Artificial Intelligence Communications*, IOS Press, 7(1), pp. 39-59, 1994.

[2] Bernstein, A. and Provost, F., An intelligent assistant for the knowledge discovery process. *Proc. of the IJCAI–01 Workshop on Wrappers for Performance Enhancement in KDD*, Morgan Kaufmann, 2001.

[3] Engels, R., Lindner, G. and Studer, R., A guided tour through the data mining jungle. *Proc. of the 3rd Int. Conf. on Knowledge Discovery in Databases*, pp. 14–17, 1997.

[4] Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P., The KDD process for extracting useful knowledge from volumes of data, *Communications of the ACM*, 39(11), pp. 27-41, 1996.

[5] Hilario, M. and Kalousis, A., Fusion of meta-knowledge and meta-data for case-based model selection. *Proc. of 5th Conf. on Principles and Practice of Knowledge Discovery in Databases*, Springer, pp. 180-191, 2001.

[6] Lindner, C. and Studer, R., AST: Support for algorithm selection with a CBR approach. *Proc. of the 3rd European Conf. on Principles of Data Mining and Knowledge Discovery*, Springer, pp. 418-423, 1999.

[7] MetaL http://www.metal-kdd.org/.

[8] Michie, D., Spiegelhalter, D. and Taylor, C., Machine Learning, Neural and Statistical Classification, *Series Artificial Intelligence*, *Ellis Horwood,* 1994.

[9] Mining Mart http://www-ai.cs.uni-dortmund.de/MMWEB/index.html.

[10] Predictive Model Markup Language. http://www.dmg.org/index.html.

[11] Richter, M., The knowledge contained in similarity measures. (Invited Talk) at the *1st Int. Conf. on Case-Based Reasoning*, Lecture Notes in Artificial Intelligence 1010, Springer Verlag, 1995.

[12] Verdenius, F., and Engels, R., A process model for developing inductive applications. *Proc. of the 7th Belgian-Dutch Conf. on Machine Learning*, pp. 119-128, 1997.