The Add-Value of Cases on WUM Plans Recommendation

Cristina Wanzeller
Escola Superior de Tecnologia e Gestão
Instituto Politécnico de Viseu e CI&DETS
Campus Politécnico, 3505-510 Viseu, PORTUGAL
Email: cwanzeller@di.estv.ipv.pt

Orlando Belo
ALGORITMI R&D Centre
University of Minho
PORTUGAL
Email: obelo@di.uminho.pt

Abstract—**Web Usage Mining is nowadays extremely useful to a diverse and growing number of users, from all types of organizations trying hard to reach the goals of their Web sites. However, inexperienced users, in particular, face several difficulties on developing and applying this kind of mining processes. One crucial and challenging task is selecting proper mining methods to deal with clickstream data analysis problems. We have been engaged on designing, developing and implementing a case based reasoning system, specifically devoted to assist users on knowledge discovery from clickstream data. The system's main aim is to recommend the most suited mining plans, according to the nature of the problem under analysis. In this paper we present such system, giving emphasis to the retrieving of similar cases using a preliminary constructed case base.**

I. Introduction

WEB based technology is widespread, but implementing and administrating Web sites still are activities increasingly complicated and very time consuming to most of the organizations. The Web has matured and the users have diverse, rising and unusual requirements. Indeed, site usage differs very often from expectations, demanding deep decision support to orient improvements. Site promoters require objective feedback regarding site effectiveness evaluation and insights how to enhance the Web offers. Hence, knowing and understanding visitors' behaviour is strategic to achieve site goals and the maximize Web's potential.

Web Usage Mining (WUM) is related to the application of the general *Knowledge Discovery* (KD) processes to data related with the interaction activity between visitors and Web sites, known as clickstream or usage data. WUM is an important tool to a large diversity and increasing number of decision maker users along the organization, having very different levels of knowledge in this area. Inexperienced users face many difficulties, among them, the crucial challenge of selecting proper *Data Mining* (DM) methods to deal with clickstream data analysis problems. Conversely, skilled users hold acquired know-how that may be especially relevant to the remaining users, surrounded by the same environment and confronted to similar decision problems to solve. Ideally, this know-how should be available in order to be shared and reused across the organization, creating new forms of synergies and empowering potential analysts. Moreover, the knowledge obtained from these experiences may be reused along diverse organizations, enlarging its application and usefulness.

Having the referred issues in mind, we decided to build a system, named *Mining Plans Selector* (MPS), especially devoted to assist users on developing and applying WUM processes. The past successful WUM exercises of the organization are the base that sustains the followed approach, based on the *Case Based Reasoning* (CBR) paradigm. CBR is a learning and problem solving approach [1], [10], [17], [18]. Instead of relying solely on general knowledge of a problem domain or making associations along the generalized relationships between problem descriptors and conclusions, CBR is able to utilize the specific knowledge of previously experienced, concrete problem situations or cases [8] [18]. A new problem is solved by finding a similarity to the formal approach of the past case and reusing it in the new problem situation. A second important difference is that CBR also is an approach to incremental and sustained learning, since a new experience is retained each time a problem has been solved, making it immediately available for future problems [1].

The MPS system behaves as a corporative tool to capture, manage and reuse the previous WUM application cases. The system's main aim is to suggest the most suited WUM plans, according to the nature of the problem under analysis. The MPS also provides support to collect and organize the knowledge gained from the experience on solving WUM problems, bringing such knowledge up to date and promoting the system's sustained incremental learning. New WUM processes are stored on a collective case base, centralizing a key resource to the MPS's capacity to solve problems in a corporative knowledge base.

Assisting decisions within KD processes is not a new initiative. There are some works that explore the CBR paradigm to undertake related purposes. For instance, the Mining Mart project [13] represents several efforts regarding the reuse of successful data pre-processing processes, appealing to a case based metadata repository. However, to help the users on establishing the mapping between the problem at hands and the stored ones, this system doesn't explore the potential meta-model neither the typical CBR methods,. The main focus of active support lies on the adaptation of the selected case to the current problem. Furthermore, this project is cen-

978-83-60810-22-4/$25.00 © 2011 IEEE

tered in pre-processing activities, not in DM or KD processes.

Another example is the *MetaL* project [12]. This project involved multiple research and development initiatives, some of which based on the CBR paradigm (e.g. [6], [11]). The main aim was to assist the user in the model selection step of the KD process. The project attention focused mainly on the algorithms selection issue, within regression and classification problems. Contrariwise, our work has a different perspective and scope. MPS is devoted to the WUM specific domain and considers processes development at distinct levels. MPS previews assistance on models selection, comprising diverse DM functions and processes involving transformation operations and multiple stages. Besides, the system reaches a greater level of abstraction.

In this paper we explain the motivation of our work and the followed strategy, and we present briefly the MPS system. We are testing the system more exhaustively, particularly the retrieving of solutions to similar WUM problems using a preliminary constructed case base. This case base contains WUM application examples describing and reproducing experiences available. MPS is able to capture knowledge from experience, using a semi–automatic approach, and retrieves similar WUM processes, giving a specific target dataset and analysis requirements. In section II, we present the challenge we are trying to address. We describe our approach, particularly the case base (section III) and the MPS system (section IV) main characteristics. Additionally, in section V we present some of the most relevant issues involved on the construction of the preliminary case base, and in section VI we discuss the process of retrieving similar WUM application cases and the general results of its evaluation, using the preliminary case base.

II. Mining Clickstream Data

Clickstream or usage data is automatically logged by Web servers, being a very rich and valuable source of visitors' behavior information. This data provides a detailed record of every single action taken by the visitor, besides the outcome of the process, typically captured on traditional off-line interactions. Moreover, clickstream data is captured implicitly without questioning users directly, providing a non-intrusive way to obtain objective feedback. Therefore, exploring WUM to extract knowledge from this and related data (e.g. users' demographic and transactions' information) has potentially enormous benefits to organizations [20]. Some important and actionable areas of WUM exploration consist of Web personalization, business intelligence, system performance improvement and site content and structure enhancement [19]. For instance, known examples of Web personalization, namely automatic recommendation, include Amazon.com's personalized recommendations and music or playlist recommenders such as Mystrand.com commercial systems [14].

Naturally, electronic commerce sites get much attention, both in professional and research arenas. Electronic commerce is considered a "killer" domain for DM since many of the ingredients necessary for successful DM are easily satisfied, including [8] [9]:

(i) wide records, i.e. many attributes or variables;
(ii) many records, i.e. large volume of data;
(iii) controlled data collection (e.g. electronic data gathering);
(iv) results can be evaluated and return on investment measured;
(v) action can easily be taken (e.g. change the site, offer cross-sells).

In electronic commerce the underlying goal is quite objective, typically to increase sales and profit, and may be achieved by understanding properly customer access behavior. Some businesses exist only virtually on the Web and, obviously, improving offers and even previewing needs are crucial to all organization members.

As any other KD process, WUM is an open-ended, exploratory and participant driven process, involving several actions and decisions, which usually comprise [4]: (i) picking relevant data (dataset and variables); (ii) identifying proper DM functions; (iii) choosing suitable models or algorithms and setting its parameters; (vi) transforming data to improve its quality, to better fit the methods assumptions and to answer a concrete analysis problem. Those activities and decisions are not trivial. By the contrary. Selecting proper mining methods, i.e. functions and models, and applying them to the available data are known challenges of the KD process development. Among multiple issues, they require an appropriate reformulation of the practical problem into a DM problem and a deeper technical understanding of the methods, being also influenced by many kinds of factors, often complex and subjective, such as the characteristics of the available data and the process preference or success criteria. Besides, some methods overlap in terms of the problems they can solve. Consequently, KD activities are typically accomplished repetitively, following different directions and testing several variants of each direction. Examples of variants include trying and comparing different attributes selection, data transformations and models parameter's settings. In short, KD and WUM are complex and very time consuming processes, frequently not leading to useful results for a particular goal.

As expected, the Web environment and clickstream data characteristics increase even more the general challenge. Distil the important information from the irrelevant one, deal with too much particularities and rapid changing conditions and get meaning from the data, are only a few subset of such issues. Analysts must tackle (human and not human) visitors' behavior aspects, which, in the last case (not human), skew the results and tend to be progressively more varied and difficult to distinguish. In fact, most of the previously pointed successful ingredients of the electronic commerce domain are also present in other types of Web sites or activities and viewed as truth challenges, requiring greater efficacy on WUM processes. Namely, becomes necessary to treat systematically such huge, complex and constantly growing data source, counting with hard time constraints. Decision makers across the organization demand for fast transformation of this massive data into valuable and actionable knowledge, to orient new ways of acting and site's improvements leading to revenue. Additionally, in the specific WUM area the problem types, the kinds of mining activities, the related practical applications and the key data items are less studied and structured.

Our underlying goal is to promote a more efficient, effective, and synergetic use of the organization's resources, decreasing the effort and time required to derive useful knowledge, bringing up together multiple valuable contributions to overcome the main difficulties. The focus of our work lies on the WUM processes development challenge of selecting suitable mining methods to apply on a specific clickstream analysis problem. We gave more emphasis to the modeling phase of the WUM process, typically presuming the availability of sources containing pre-processed data, but considering also tasks of the remaining phases. Our idea is that arduous and intensive pre-processing tasks must be centralized in a previous stage, in order to make data available to all the potential analysts in more manageable forms. The primary target of our work is, precisely, the inexperienced analysts, facing problems that may be solved exploring WUM. Consequently, an implicit requirement is to support problem descriptions making use of abstractions related to the real problems to solve and, naturally, to establish direct relationships among such abstractions and the most promising DM methods and approaches.

III. Strengths and Opportunities of WUM Application Cases

The most important learned lesson from 2000 KDD Cup annual competition was the crucial role played by humans in WUM processes development, even when the only interesting success criteria was accuracy or score [8]. Human insight was strategic in tasks as feature selection and construction from hundreds of available attributes and in the choice of mining methods. Indeed, most of the success obtained by experts, when dealing with WUM problems, comes from their acquired know-how. Even they cannot provide general and consistent rules to support problem solving.

Building up WUM application cases has considerable strengths, mainly realized by structuring and memorizing the knowledge acquired from the experience. Hence, we decided to document, catalogue and store WUM past experiences, in a specific oriented knowledge base that could be applied over clickstream data analysis. Examples of past successful solved problems might be the most helpful and convincing form of aid in this scope, since 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.

A straight reuse of WUM solutions is quite possible in this scope, since recurrent problems are common. Still, becomes necessary to enable flexible means for relating new problems and the stored ones, to help users on identifying the most plausible strategies to address the problem at hands. The CBR paradigm brings a key opportunity to our knowledge base, providing inherently a proper way for attending this demand. CBR methods favor a flexible similarity based comparison, even if the involved features are not objective and precisely defined. CBR can cope with incomplete and subjective information and makes possible to consider only the relevant features and use specific importance levels, increasing the potential of answering the real user needs. Furthermore, CBR provides 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 CBR strength is of great importance to us, due to the constant evolution of WUM and the need to incorporate knowledge about new mining algorithms, tools, types of problems, solving approaches and kinds of discoveries applications.

Defining and representing cases are also crucial issues for CBR. A case may be defined as a contextualized piece of knowledge representing an experience that teaches lessons fundamental to achieving goals [8]. In MPS, a case is a WUM process described by a set of fundamental dimensions (D, T, P, A and K) and, combining the CBR principles, structured in terms of a domain problem and the applied solution (Fig. 1). A problem is essentially defined by:

- characterizations of the available `data` (D), at dataset and variables level;
- categorizations of the WUM problem `type` (T), mainly in terms of abstractions such as main underlying activity, analysis goals and practical application areas;
- `process` evaluation criteria (P) (not shown in Fig. 1).

The applied solution comprises: a sequence of `activities` (A), including transformation and modeling stages, the involved data and the model parameter settings; prior and derived `knowledge` (K), concerning to facts that affected the analysis, the extracted knowledge and the relations to such facts; and general information about the WUM `process` (P).

We also have a context description item to organize cases in terms of a Web site's perspectives or particular sections. This item is a logic container for cases description features. The initial idea was to avoid redundancy on descriptions, since we have several datasets and common features. Though, the context provides flexibility on retaining details from different parts of one Web site. The context may be associated with some aspects of problem description, namely dataset, activity, and specific and general facts.

The most important question concerning datasets is to capture the relevant properties to the particular purpose of DM methods selection. We need a consistent characterization, in order to be able to compare dissimilar datasets. In fact, we compare the metadata, not the clickstream dataset itself. Our strategy is based on a common data characterization approach [11]. This approach has been frequently and successfully used in Meta-Learning, to select adequate learning algorithms. In general terms, we adapted this approach to clickstream data characteristics.

The cases' problem part is used to describe WUM problems and to find out previous similar cases, both defined

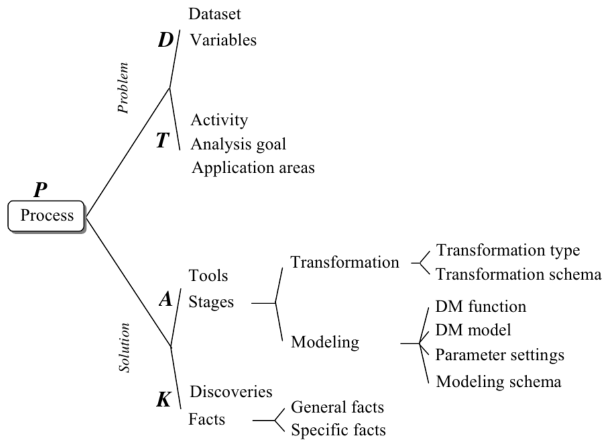

Fig. 1 WUM application case main elements

based on the common features (Table 1). The solution parts of the most similar cases retrieved are used to produce mining plans, forming the recommended solution to the submitted problem.

IV. Mining Plans Selector System

The tasks involved in CBR have been described as a cyclical process, comprising the 4REs i.e. retrieve, reuse, revise and retain [1]. We adapted this widely acknowledge cycle to the activities to perform by the MPS system, devising six constituent steps. These steps form a problem solving and learning from experience strategy, oriented to the WUM so special application domain. Fig. 2 shows the adopted cycle. The original steps from [1] are presented, at italic, to distinguishing them from the added ones.

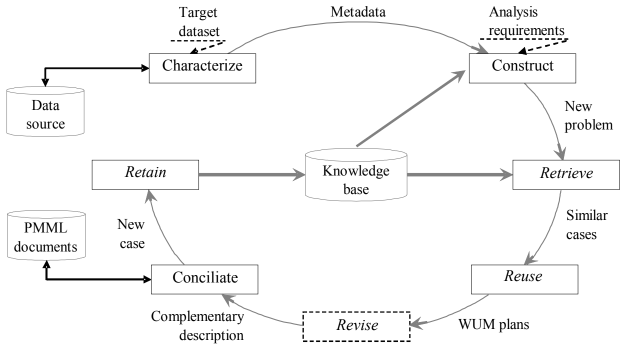

Fig. 2 Adopted CBR cycle

To solve a problem, the MPS system acts tacking as inputs the target dataset and the analysis requirements and delivers WUM plans appropriate to the current problem based on the cases kept on the knowledge base. The problem solving part of the system comprises five steps: Characterize, Construct, *Retrieve*, *Reuse* and *Revise*. One MPS's specific task is to characterize the target data, producing a systematic and consistent meta-representation, comprising different types of data sources. Another particular task is to construct a new WUM problem, guiding, gathering and organizing the user's explicit analysis constraints specification. The retrieve task is a typical one, being used to find out the cases most similar to the target problem. The reuse task generates WUM plans, mostly based on the mining methods and the levels of similarity of the retrieved cases, and considering also the evaluation criteria most important to the analyst. This step does not performs extensive adaptation of the solution to the current problem, namely in the wide sense intended by the original step. Nevertheless, it focus the main parts of the candidate cases that may be transferred to the target problems, by rec-

Table I.
List of Problem Description Features

	Category	Features	Similarity measure for:	
			Single values	Set values
P	Evaluation criteria	Precision, Time of reply, Interpretability, Resources requirements and Implementation simplicity	(NMDc)	
	Process date	Date	(NMD)	
T	Site's activity	(a set of) Activity	(SM)	(HMA)
	DM task	(a set of) Goal (a set of) Application area	(SM) (SM)	(HMA) (HMA)
D	Characteristics at dataset level:	-Number of lines and columns/variables -% of numeric, categorical, temporal and binary columns -Granularity (e.g. session) -Type of visitor's identification -Type of visitor's information recording -Access order and access repetition availability -Access data and hour availability	(NMD) (NMD) (E) (E) (E) (E) (E)	
	Characteristics at variable level:	(a set of) Variable: -Data type -Semantic category -Number of distinct values -Number of null values	(G) (SM) (SM) (NMD) (NMD)	(MA)

ommending mining methods instead of mere cases, preparing the reuse of the methods that constructed the solution. The revise step is accomplished outside of the system, using a KD tool.

Concerning the MPS learning perspective, the system operates accepting heterogeneous descriptions of new WUM processes and acquiring knowledge. MPS uses a semi-automated learning approach, in order to systematize and simplify such arduous activity. The steps included in learning are: Conciliate and *Retain*. The accepted incomings are: documents describing mining activities, generated by the KD tool in *Predictive Model Markup Language* (PMML) [15], a XML based standard to define and share statistical, and DM models across compliant applications; the process complementary description, which would be exhaustive, if the used tool does not supports the PMML standard. First, a conciliate task transforms and combines the heterogeneous descriptions items, supplied by user interaction and documents in PMML. Then the traditional retain task essentially augments the knowledge base with a new case, elaborated by integrating and structuring the incoming elements, considering the internal schema of the cases' representation.

V. Building up the Preliminary Case Base

Testing a system like MPS, specifically the problem solving point of view, requires an extensive case base. We must accept large and diverse input datasets, since clickstream data are huge and analyzed in distinct forms. More important, we need a wide set of successful and representative WUM processes, using the existent datasets. Such processes have to include different DM functions and methods and be applied to solve a comprehensive set of typical problems types.

The preliminary case base was build appealing to WUM application examples, based on real data and analyses that were available in the Internet, and some other research works. Some of these analyses reproduce WUM processes developed with such data, published together with the respective sources or in research papers. This option was made to attend the requirements previously explained, as well to overcome the issue of the discoveries success subjectivity, which in the case of simulated examples is relative and difficult to evaluate. So, we provide a greater level of success guaranty. Another advantage of this option was the greater diversity of situations.

About disadvantages, the system was not evaluated according to the previously established and idealized circumstances: pre-processed and quality data; analysis problems one of an organization. Preparing cases, in these conditions, is a more complicated and time consuming task. The analysis of each dataset is extended for longer periods, since it requires data transformation efforts and, mainly, the understanding of these data and its surrounding context. Furthermore, the case base profile changed slightly, since it concerns to more than one organization. However, first, this case base provides wide application to different kinds of organization and situations. Second, the system proved that was able to retain details from varied environments. The context description, previously described, was very useful to deal with this new scenario.

Regarding the prepared cases main characteristics, we emphasize the diversity among the series of original data, from which the used datasets were derived. The original data vary from Web servers logs, in its rude form, to data already preprocessed, and in some cases with distinct series of data devoted for the treatment of different problems. This is precisely the situation of the 2000 KDD Cup case study [8]. One used three datasets of this source and multiple analyses based on them. In this type of situation we mainly filtered and derived new features, taking into account data quality and relevance to the problem at hands. Other datasets were pre-processed and used to generate multiple datasets, such as, for example, page view or access level clickstreams, aggregations at session level and binary matrices (e.g. sessions X accessed pages). Other known and available used examples were the *msnbc* [5] and *ECML\PKDD 2005 Discovery Challenge* [2] datasets and reported experiences, both about clickstream data analysis.

The construction of the preliminary case base provided the way to conduct experimental tests of the semi-automated learning approach of the system. This approach proved to be very useful on decreasing the efforts on processes extensive descriptions and to reduce the dependency from WUM experts. The well known datasets we mentioned are very long, being very handy to have automatic ways for capturing dataset metadata. Thus, dataset characterization was tested under demanding conditions and the respective metadata was successfully captured. Besides, usually we have processes with several stages, including each one the selection of numerous variables and the specification of several values of parameter settings. The learning approach was used, with success for all the WUM processes from which was possible to obtain PMML documents, despite the need to complement the description through explicit user interaction. We may conclude that the learning approach is effective. The problem solving part experimental results are discussed in the next section.

VI. Retrieving Similar WUM Processes

The retrieve step plays a vital role on problem solving. This step selects the most plausible cases to found the construction of mining plans to recommend, according to the target problem. The variants of problem description that might be submitted to the system are diverse, but may be systematized into three main types, related with the previously mentioned dimensions: oriented by the target dataset (dimension D), by other kinds of constraints (dimensions P and T) or both (dimensions D, P and T). Table 1 shows the problem description features and the measures used to assess the level of the similarity (defined on Table 2).

The similitude assessment approach devised over WUM problems comprises the modelling of the following types of measures:

- local similarity measures for simple (single-value) attributes;

- local similitude measures for structured (multiple or set value) features (namely MA and HMA measures);
- global similarity measures (G) 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. It is applied at variable's and case's level. 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 (E) matches (e.g. for binary attributes) or is expressed in form of similarity matrices (SM), which establish each pairwise similitude level (e.g. for some symbolic descriptors). To compare numeric simple features, we adopted similarity measures mainly based on the normalized Manhattan distance.

We also need similarity measures for complex descriptors, modeled as set–value features, containing atomic values (e.g. application areas) or objects having themselves specific properties. For instance, variables have specific properties (e.g. data type) and may occur in different number for each dataset. Indeed, these needs were 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 (e.g. [3], [7], and [16]). Even so, we explored a number of them, for instance, the measures suggested on [7], and the comparative tests performed lead us into tailored or extended (MA and HMA) measures, better fitting our purposes, as reported in [21].

Concerning the retrieve evaluation, the general and specific tests performed so far demonstrate to the system's effectiveness. The specific tests included the comparison among distinct types of objects, such as series of variables and datasets. Regarding datasets, the system discriminates the most similar and dissimilar ones, based on the adopted features and proposed as the most relevant ones for selecting mining methods and approaches. The results conform to the intuitive notion of similarity among datasets, based on the general idea about each one. For instance, some identified trends were the following:

TABLE II.
LIST OF MAIN USED SIMILARITY MEASURES

Description	Measure
(G) Weight average (global similarity function)	$Sim_{global}(t,c)=\frac{\sum_{f=1}^{n} Sim_{local}(t.f,c.f)*w_f}{\sum_{f=1}^{n} w_f}$
(NMD Normalized Manhattan distance	$Sim_{Local}(t.f,c.f)= 1-\frac{\lvert t.f-c.f\rvert}{f_{\max}-f_{\min}}$
(NMDc) Normalized Manhattan distance changed	$Sim_{Local}'(t.f,c.f)=\begin{cases} 1 & c.f\geq t.f \\ 1-\frac{\lvert t.f-c.f\rvert}{f_{\max}-f_{\min}} & c.f<t.f \end{cases}$
(E) Exact (text or binary)	$Sim_{Local}(t.f,c.f)=\begin{cases} 1 & c.f=t.f \\ 0 & c.f\neq t.f \end{cases}$
(SM) Similarity matrix	
(MA) Maximums Average	$Sim_{MA}(A,B)=\frac{1}{n_A+n_B}\sum_{1}^{n_A}\max_{a\in A}(sim(a,b))+\sum_{1}^{n_B}\max_{b\in B}(sim(a,b))$
(HMA) Half maximums Average	$Sim_{HMA}(A,B)=\frac{1}{n_A}\sum_{1}^{n_A}\max_{a\in A}(sim(a,b))\quad A\subset\text{Target set}, B\subset\text{Case set}$
t, c – target and case (or part of them) *t.f, c.f* – values of each feature f *Sim_{local}* – local similarity measure *n* – number of features *w_f* – feature f importance weighting	*f_{max}, f_{min}* – maximum and minimum values (observed) on feature f *A, B* – two sets, such that a ∈ A and b ∈ B *sim(a,b)* – similarity between each pair of elements of the two sets *n_A, n_B* – cardinality of the sets A and B

- When the target dataset is a binary matrix: the most similar datasets are also binary matrices; the most dissimilarity datasets are common and, frequently, datasets having access granularity.
- When the dataset has access granularity: the most similar datasets also have access granularity; the most dissimilar datasets are usually the same.
- When the target data set has session or other granularity (e.g. visitor) and is not a binary matrix: there is not a simple and strait similarity pattern (justified by the variety of attributes gathered at these levels); the most dissimilar datasets have mainly access granularity.

In terms of general tests, the remain descriptors also reflect influent factors and affect cases retrieving, contributing for establishing the bridge between analysis requirements and suited mining methods and approaches. The system relates WUM processes based on similar intentions and applications, but not necessarily coincident. Since the data characterization descriptors are in majority, within the problem description, by default their relative importance is greater and the system tends to select processes based on similar datasets. This default behaviour accords to the intended one and is considered a good result. 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. Furthermore, the system provides means to change the default behaviour and to improve the problem specification, namely exact filtering criteria, specific descriptors importance levels and the exclusion of irrelevant (or unknown) descriptors.

VII. Conclusions

Rapidly changing conditions and the global competition have brought tremendous pressure into organizations way of life, demanding an effective presence on the Web and a more responsive and proactive attitude to realize its full potential. WUM is one crucial tool to bridge the gap between massive clickstream data and actionable knowledge, in order to devise Web site's opportune enhancements. However, WUM learning curve is a serious obstacle to inexperienced users, being pertinent to have a strategy showing the way how to proceed.

The proposed and developed work aims at promoting a more efficient, effective and synergetic exploration of WUM, decreasing the effort and time required to derive useful knowledge from clickstream data. To achieve this aim we designed, developed and implemented a prototype of a CBR system, specifically devoted to assist users on WUM processes, mainly on selecting proper mining methods and approaches to address analysis problems. The system also provides support to users on documenting and organizing the knowledge gained from the experience on solving new WUM problems, through a semi-automatic learning approach. The previous collected and stored WUM application cases are therefore the base that sustains the recommendation of mining plans to solve new problems.

We believe that the MPS system is a good tool for knowledge creation, sharing and reuse. The system is based on abstractions related to the real problems to solve, meaning that it could serve the particular needs of less skilled users, wishing to learn how to handle a concrete problem, being also useful to specialists interested in reusing successful solutions, instead of solving the problems from scratch. Data is always growing and is increasingly stored by organizations. DM tools are gaining more importance and KD processes are becoming more useful and widespread, although they remaining complex.

In this paper we described our system, focusing the retrieving of similar cases and its evaluation using a preliminary case base. These prepared cases reproduce WUM exercises descriptions publicly available, overcoming the need of a wide set of examples and the issue of the discoveries success subjectivity. The used datasets and the reproduced WUM processes are challenging and real applications of WUM, proving demanding conditions to test the MPS system. The cases contain some diversity of circumstances, which is beneficial to sustain the construction of a repository of this nature. The system's evaluation appealing to this preliminary case base also points to the system's effectiveness. A drawback to point out is the intentional generality of some abstractions used to categorize problems (e.g. analysis goals and application areas), which restricted their diversity. The potential of the approach has not been completely explored, since greater levels of abstraction might be achieved, enlarging the case base and developing further such categorizations.

For the future we plan to further evaluate the current implementation. This will be realized through the preparation of more cases and, particularly, within the context of a study case, based on a concrete target organization.

References

[1] A. Aamodt and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations and Systems Approaches" in Artificial Intelligence Communications (AICom), IOS Press, vol. 7, no 1, pp. 39-59, 1994.

[2] *ECML\PKDD 2005* conference web site. http://www.liaad.up-.pt/~ecmlpkdd05/. Access June 2011.

[3] T. Eiter and H. Mannila, "Distance Measures for Point Sets and their Computation", Acta Informatica, vol. 34, no 2, pp. 109–133, 1997.

[4] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communications of the ACM, vol. 39, no 11, pp. 27-41, 1996.

[5] S. Hettich and S. D. Bay, the UCI KDD Archive [http://kdd.ics.uci.edu]. Irvine, CA: University of California, Department of Information and Computer Science, 1999.

[6] M. Hilario and A. Kalousis, "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 pp. 180-191, 2001.

[7] M. Hilario and A. Kalousis, "Representational Issues in Meta-Learning", in Proc. of the 20th International Conf. on Machine Learning (ICML '03), AAAI Press, pp. 313-320, 2003.

[8] R. Kohavi, C. Brodley, B. Frasca, L. Mason, and Z. Zheng "KDD-Cup 2000 Organizers' Report: Peeling the Onion", SIGKDD Explorations, vol. 2 no 2, pp. 86–98, 2000.

[9] R. Kohavi and F. Provost, "Applications of Data Mining to E-commerce", (editorial), Special issue of the International Journal on Data Mining and Knowledge Discovery, 2001.

[10] J. Kolodner, "Case-Based Reasoning", Morgan Kaufman, San Francisco, CA, 1993.

[11] C. Lindner and R. Studer, "AST: Support for Algorithm Selection with a CBR Approach", in Proc. of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'1999), Springer, pp. 418-423, 1999.

[12] MetaL project http://www.metal-kdd.org/ Access June 2011.

[13] K. Morik and M. Scholz "The MiningMart Approach to Knowledge Discovery in Databases", in Intelligent Technologies for Information Analysis, Springer, 2004.

[14] B. Mobasher, "Data Mining for Web Personalization", lecture Notes in Computer Science, 4321, 90-135, 2006.

[15] Predictive Model Markup Language. Data Mining Group. http://www.dmg.org/index.html. Access June 2011.

[16] J. Ramon "Clustering and Instance Based Learning in First Order logic" PhD thesis, K.U. Leuven, Belgium, 2002.

[17] C. Riesbeck and R. Schank, "Inside Case-based reasoning", Lawrence Erlbaum, 1989.

[18] R. Schank, "Dynamic Memory: A theory of learning in computers and people", Cambridge University Press, 1982.

[19] J. Srivastava, R. Cooley, M. Deshpande and P.-N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", in SIGKDD Explorations, vol. 1, no 2, pp. 1–12, 2000.

[20] J. Srivastava, P. Desikan and V. Kumar, "Web Mining - Accomplishments and Future Directions", invited paper in National Science Foundation Workshop on Next Generation Data Mining, Baltimore, MD, 2002.

[21] C. Wanzeller and O. Belo, "Similarity Assessment in a CBR Application for Clickstream Data Mining Plans Selection" in Proc. of the 9th International Conference on Enterprise Information Systems (ICEIS' 2007), Funchal, Madeira, Portugal, 2007.