SIMILARITY ASSESSMENT IN A CBR APPLICATION FOR CLICKSTREAM DATA MINING PLANS SELECTION

Cristina Wanzeller[1] and Orlando Belo[2]
[1]*Escola Superior de Tecnologia, Instituto Politécnico de Viseu, Campus Politécnico,Viseu, Portugal*
[2]*Escola de Engenharia, Universidade do Minho, Campus de Gualtar, Braga, Portugal*

Keywords: Web usage mining, clickstream data mining assistance, case based reasoning and similarity measures.

Abstract: We implemented a mining plans selection system founded on the Case Based Reasoning paradigm, in order to assist the development of Web usage mining processes. The system's main goal is to suggest the most suited methods to apply on a data analysis problem. Our approach builds upon the reuse of the experience gained from prior successfully mining processes, to solve current and future similar problems. The knowledge acquired after successfully solving such problems is organized and stored in a relational case base, giving rise to a (multi-) relational cases representation. In this paper we describe the similitude assessment devised within the retrieval of similar cases, to cope with the adopted representation. Structured representation and similarity assessment over complex data are issues relevant to a growing variety of application domains, being considered in multiple related lines of active research. We explore a number of different similarity measures proposed in the literature and we extend one of them to better fit our purposes.

1 INTRODUCTION

Selecting the most suitable methods to apply on a specific data analysis problem is an important and known challenge of Data Mining (DM) and Web Usage Mining (WUM) processes development. 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, focusing, specifically, on the WUM domain. Our approach relies on the reuse of the experience gained from prior successfully WUM processes to assist current and future similar problems solving. In (Wanzeller and Belo, 2006) we described a system founded on the Case Based Reasoning (CBR) paradigm, implemented to undertake our purposes. This system should assist the users in two main ways: (i) restructuring and memorizing, on a shared case based repository, the knowledge acquired after successfully solving WUM problems; (ii) proposing the mining plans most suited to one clickstream data analysis problem at hand, given its high level description.

The case based representation model must support a comprehensive description of WUM experiences, regarding the nature and requirements of such process development. The CBR exploitation relies on the similarity notion, based on the assumption that similar problems have similar solutions (Kolodner, 1993). Thus, a strictly related and essential concern is to devise a similarity model able to cope with the provided representation. Though, we were faced with the issue of defining a similarity model over WUM processes, mainly because we used a (multi-) relational cases representation. This issue arises due to the one to many relationships appearing among components of cases description.

Structured representation and similarity assessment over complex data are issues nowadays common and crucial to various areas, such as CBR, computational geometry, machine learning and DM, particularly within distance-based methods. Important tasks of the last area include the grouping of objects and the classification of unseen instances. Several research efforts aim to handle and learn from more expressive and powerful data representations, than the classical propositional setting. Inductive logic programming and multi-relational data mining fields traditionally symbolize the approaches to deal with more complex and intuitive settings directly. Some prominent examples are the RIBL (Relational Instance Based Learning) (Emde and Wettschereck, 1996), the RIBL2

Wanzeller C. and Belo O. (2007).
SIMILARITY ASSESSMENT IN A CBR APPLICATION FOR CLICKSTREAM DATA MINING PLANS SELECTION.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - AIDSS*, pages 137-144
DOI: 10.5220/0002396201370144

(Bohnebeck, Horváth and Wrobel, 1998) and the RDBC (Relational Distance-Based Clustering) (Kirsten and Wrobel, 1998) systems. For instance, RIBL constructs cases from multi-relational data and computes the similarity between arbitrary complex cases (by recursively comparing the first-order components, until fall back into propositional comparisons over elementary features). Besides, great attention is being focused on upgrading some propositional learning algorithms based on typed representations. A typed approach often simplifies the problem modeling (Flach, Giraud-Carrier and Lloyd, 1998). Namely, distance-based methods can be easily extended by embedding similarity measures specifically defined over common structured data types as lists, graphs and sets.

The CBR domain also embodies a relevant and closely related line of research to address the discussed issues. Cases are often represented by complex structures (e.g. graphs, objects), requiring tailored forms of similarity assessment (e.g. structural similitude). These measurements are usually computationally expensive, but more relevant cases may be retrieved. Object-oriented case representations generalize simple attribute-value settings and are useful to represent cases from complex domains. This kind of representation is frequent, being particularly suitable when cases with different structures may occur (Bergmann, 2001). The objects' similitude is determined recursively in a bottom up fashion and has been extended by a framework, to allow comparing objects of distinct classes and considering the knowledge implicit in the class hierarchy (Bergmann and Stahl, 1998).

There are multiple approaches proposed to deal with the faced issues. In this paper we describe the ones considered and applied to accomplish the similarity assessment within our system. Section 2 concerns to the main attributes of problem description and, so, the ones comprised on the similitude estimation. Section 3 covers the similarity model adopted in the retrieval process. In sections 4, 5 and 6 we define the similarity measures considered, according to the similarity model adopted. Section 7 reports comparative tests of some similarity measures and section 8 ends the paper with conclusions and proposals of future work.

2 MATCHING WUM PROBLEMS

Our system's relational case base consists in a metadata repository, containing detailed examples of successful WUM processes, described in terms of the domain problem and the respective applied solution. Figure 1 shows the core components of cases representation, through a class diagram in Unified Modeling Language simplified notation. The central class Process represents each WUM process and interconnects the classes of problem and solution description. The transformations of the dataset, the modeling steps, together with the applied models, theirs configuration parameters and the involved variables, build up the major solution part of a case. The case's problem part comprises the features that characterize the problem type. Those features are useful to specify new problems and can be organized into two categories: the dataset specification and the requirements description. The former regards to characterization metadata, gathered at dataset and individual variable levels. The last stands for a set of constrains, based on the analysis nature and the analyst preferences.

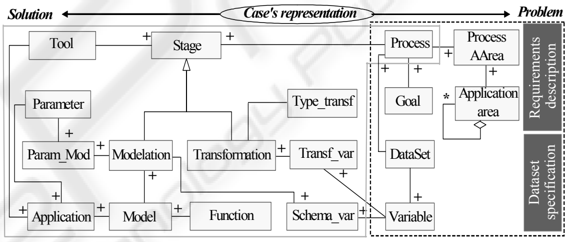

Figure 1: Cases representation conceptual model.

The conceptual metadata model reveals the cases multi-relational representation. Focusing on the problem description part, each instance of the central Process class is (directly or indirectly) related to several instances from some classes (e.g. Variable and ProcessAArea). The one-to-many relationships also occur in the specification of new problems within the target object. Table 1 illustrates the main features of the target object, along with theirs values type and organized by description categories and subcategories. The values type can be: (i) simple (atomic) in what respects to single-value attributes; (ii) a set of elements, modelling the one-to-many relationships as composed or structured multiple-value features. The table also shows the type of comparison (e.g. 1-1, N-1) when matching the correspondent features of the target and each case. The structured features give raise to comparisons between sets of finite elements, with inconstant and possibly different cardinality. We highlight three kinds of comparisons between sets:

- N-1 matches, between the sets of (symbolic) values from the Goal(s) feature, since the target might include the selection of one or more goals, although each case is related to only one instance of the Goal class;

- N-M matches, between the sets of (symbolic) values from the Application area(s) attribute, as both the target and case might be related to several instances of the ProcessAArea class;
- N-M' matches, between sets of elements with theirs own features, when comparing dataset variables from the target and each case.

Besides the target object, the specification of new problems might comprise additional information: (i) degrees of importance assigned to each feature, expressing which attributes are more or less relevant to the user; (ii) exact filters that define hard constraints for the features values of the available cases.

3 RETRIEVING SIMILAR CASES

A key process of CBR systems is retrieving the most similar case(s) that might be useful to solve the target problem. A core task to undertake is to measure the similarity of the target problem to the previous described problems, stored on the case base, along with theirs known solutions. In the most typical application of CBR, the similarity assessment of the cases is based on theirs surfaces features. Such features are the qualitative and quantitative attributes held as part of the cases' description, usually represented using attribute-value pairs. The similitude of each case's problem to the target problem is computed considering the correspondent feature values and the selected similarity measures. The similarity measures are a critical component in any CBR system, containing by itself knowledge about the utility of an old solution reapplied in a new context (Bergmann, 2001).

A general principle to orient the similitude assessment splits its modelling, defining two types of measures: the local and the global ones. The local similarity measures are defined over the simple attributes. The global similitude considers the whole objects and gives an overall measure, according to some rule that combines the local similarities (an aggregation function) and a weight model. This useful principle has to be extended to tackle our requirements. Since the structure and the features considered are the same to all cases, we do not have to handle the issues of comparing objects with arbitrary or distinct structure, neither measuring deeper forms of similarity (as the ones covered in Bergmann and Stahl, 1998; Emde and Wettschereck, 1996). We can explore a simpler and intermediate approach, based on the similarity assessment over structured data types, namely set valued features. The adopted approach comprises the modelling of the following items:

- global similarity measures defined through an aggregation function and a weight model (Sim_{global});
- local similarity measures for simple (single-value) attributes ($Sim_{local\ SINGLE}$);
- local similarity measures for structured (set-value) features ($Sim_{local\ SETS}$).

The basic procedure of similarity assessment using the previous model (omitting the weight factor) can be described by the following algorithm.

```
Input: target t and case c, whose problems are described by a set of features f.
Output: the similitude value between t and c.
1. Features loop: For each feature f Do:
      If f is simple then       /* Similarity for simple/single-value features  */
         $sim_f \leftarrow Sim_{local\ SINGLE}$ measure applied to the f values from t and c;
      Else                      /* Similarity for structured/set-value features */
         If the elements of the sets are simple then
            Use a $Sim_{local\ SINGLE}$ measure between each pair of values from t and c;
            $sim_f \leftarrow Sim_{local\ SETS}$ measure applied to all the previous values;
         Else                   /* sets' elements have their own inner features */
            Use $Sim_{local\ SINGLE}$ measures to determine the similarity between the
              correspondent inner features, for each pair of elements from t and c;
            Employ $Sim_{global}$ to aggregate the similarities of the inner features,
              deriving a similitude value for each pair of elements from t and c;
            $sim_f \leftarrow Sim_{local\ SETS}$ measure applied to all the previous values;
2. Overall similarity between t and c: Apply a $Sim_{global}$ measure to all the $sim_f$
   values.
```

Table 1: List of (sub)categories, features and values type of the target problem description.

	Subcategory	Features	Value (comparison) type
Requirements description	Evaluation criteria	Precision; Time of reply; Interpretability; Resources requirements; Implementation simplicity	Simple ordinal (1-1)
	WUM process date	- Process date	Simple continuous (1-1)
	DM task	-Goal(s) -Application area(s)	Set of symbolic (N-1) Set of symbolic (N-M)
Dataset specification	Characteristics at dataset level: - DM generic	-Number of lines and columns/variables -Percentage of numeric, categorical, temporal and binary columns/variables	Simple continuous (1-1)
	- WUM specific	-Type of visitant's identification -Access order and access repetition availability -Granularity (e.g. session), etc.	Simple Boolean and symbolic (1-1)
	Characteristics at variable level: - DM generic	(a set of) -Data type -Number of distinct values -Number of null values	Set of variables (N-M'): Simple symbolic and continuous
	- WUM specific	-Semantic category	

4 MEASURING GLOBAL SIMILARITY

Global similarity measures are applied at different levels, namely at case and sub-object levels, to aggregate the local similitude values from simple or structured features. The adopted and implemented measure consists on the traditional weight average function, defined by equation (1), where *t* and *c* denote the target and the case objects (or part of them), *t.f* and *c.f* are the correspondent values of each feature *f*, Sim_{local} is a local similarity measure, *n* is the number of features used in the comparison and w_f is the importance weighting of the feature *f*.

$$Sim_{global}(t,c) = \frac{\sum_{f=1}^{n} Sim_{local}(t.f, c.f) * w_f}{\sum_{f=1}^{n} w_f} \quad (1)$$

A global similarity measure is task oriented and contains utility knowledge (Bergmann, 2001). The adopted measure reflects the case features relevance, based on the respective assigned weights, allowing them to have varying degrees of importance. Our perspective is that defining features relevance levels (e.g. important, very important) is useful to better specify the problem and belongs to its description. Thus, our weight model (currently) consists essentially in representing the mappings between the available relevance levels and the used (internal) weighting values.

5 MEASURING SIMILARITY ON SINGLE-VALUE FEATURES

Local similarity measures contain domain knowledge, reflecting the relationship between the values of a feature. The adoption of a local similarity measure depends mainly on the feature domain, besides the intended semantic. The local similarity between categorical features is based on exact matches (e.g. for binary and text attributes) or can be expressed in form of similarity matrices (e.g. for symbolic attributes), which establish the similitude level. These matrices are held by relational tables.

To compare numeric features we adopted a common similarity measure, based on the normalized *Manhattan* distance (i.e. Similarity=1-Distance), defined by the equation (2), where *t.f*, *c.f* denote the target and case values of feature *f* and f_{max}, f_{min} are the maximum and minimum values (observed) on feature *f*, used to normalize the result.

$$Sim_{Local}(t.f, c.f) = 1 - \frac{|t.f - c.f|}{f_{\max} - f_{\min}} \quad (2)$$

The local similarity measure for the evaluation criteria features is an exception, being build upon a constraint imposed to equation (2), given by the rule:

$$Sim_{Local}{}'(t.f, c.f) = \begin{cases} 1 & c.f \geq t.f \\ 1 - \dfrac{|t.f - c.f|}{f_{\max} - f_{\min}} & c.f < t.f \end{cases} \quad (3)$$

The constraint imposed (3) was applied since the target values of evaluation criteria are meant as lower bounds of the features. In addiction, greater (>) always means better than the searched value, using the ordinal scale defined to these features. By this reason, the similarity should be estimated only if the case value is worst (lower).

6 MEASURING SIMILARITY OVER SET-VALUE FEATURES

The cases structured representation brings up issues to the similarity estimation, not addressed by the previous measures. We need to define similarity measures between sets of elements, containing atomic values or objects having themselves specific properties. The similitude between each pair of elements was modeled through: similarity matrices, defined over the distinct values of a feature (e.g. for goals); specific features that might have different levels of importance (e.g. for dataset variables).

There are already many proposals in the literature for measuring the similarity or distance between sets of objects (Eiter and Mannila, 1997; Gregori, Ramírez, Orallo and Quintana, 2005; Ramon, 2002). One widely used measure is the *Jaccard* coefficient. This coefficient, however, is not suited for our application, since only exact matching between elements is taken into account. We want to consider inexact matches, namely the proximity that can be derived by the similitude between different elements of the sets. Hence, this coefficient, theirs variants and alternative measures with the same behavior will not be considered in the sequel.

In (Hilario and Kalousis, 2003) three similarity measures are used to handle the comparison between sets, namely between dataset variables sets. Those measures rely on notions from clustering, where such type of comparison is a common task. Two of the used similarity measures are inspired in ideas developed in agglomerative hierarchical clustering algorithms (Duda, Hart and Stork, 2001). The first one is based on the *single linkage* algorithm, being defined as the maximum similarity observed between all pairs of elements of the two sets (as explaned in Table 4). Given two sets A and B, such that a ∈ A and b ∈ B, this measure is defined by the following equation:

$$Sim_{SL}(A,B) = \max_{a,b}(sim(a,b)) \qquad (4)$$

where *sim*(a,b) is the similarity between each pair of elements of the two sets (determined as explained in section 3). The other (second) similarity measure comes up from the *average linkage* algorithm. This measure is defined as the average similarity between all pairs of elements (a,b) from the two sets A and B, being given by:

$$Sim_{AL}(A,B) = \frac{1}{n_A * n_B} \sum_{1}^{n_A * n_B} sim(a,b) \qquad (5)$$

where n_A and n_B denote the respective cardinality of the sets A and B. The third measure is used on the scope of similarity-based multi-relational learning (Kirsten, Wrobel and Horvath, 2001), being based on the measure of RIBL (Emde and Wettschereck, 1996). The similarity between two sets is defined on equation (6) as the sum of the maximum similarities of the set elements with lower cardinality with the elements of the set with the greater cardinality, normalized by the cardinality of the greater set.

$$Sim_{MR}(A,B) = \begin{cases} \dfrac{1}{n_B} \displaystyle\sum_{1}^{n_A} (\max_{a \in A} sim(a,b)),\ n_A < n_B \\ \dfrac{1}{n_A} \displaystyle\sum_{1}^{n_B} (\max_{b \in B} sim(a,b)),\ n_A \geq n_B \end{cases} \qquad (6)$$

In (Eiter and Mannila, 1997) the authors review various distance functions proposed on sets of objects, among them the known *Hausdorff* distance (and metric) and the sum of minimum distances (SDM). The *Hausdorff* distance has some quite attractive properties (e.g. being a metric), but it does not seem to be suitable as similarity measure for our application, since it relies too much on the extreme values of the elements of both sets. Thus, after testing its inadequacy, this measure has not considered any more.

Given the measures considered as the most promising to our purposes and based on the comparative tests performed (and exemplified in section 7), we combined the best ideas from some measures to define two refined similarity measures more suited to our approach. The first one (*Sim*MA) is alike the SDM measure (after transforming the minimum into maximum), which maps every element of both sets to the closest element in the other set. The *Sim*MA measure (i.e. “Maximums Average”) takes into account the maximum similarity between each element and the other set, and averages these values, being defined by:

$$Sim_{MA}(A,B) = \frac{1}{n_A + n_B} \sum_{1}^{n_A} \max_{a \in A}(sim(a,b)) + \sum_{1}^{n_B} \max_{b \in B}(sim(a,b)) \qquad (7)$$

where *sim(a,b)* is the similarity between each pair of elements and n_A, n_B are the cardinality of the sets. In the second proposed measure the similarity between the target and the other set is defined as the average of the maximum similarities of the elements from the target set with the elements of the case set. This measure is asymmetric and is a variant of Sim_{MA} (7), based on Sim_{MR} (6). It uses less information than Sim_{MA}, namely the half concerning directly to the target, and gives emphasis to the greatest similarities of one set, like the Sim_{MR} measure. The measure is defined by the equation:

$$Sim_{HMA}(A,B)=\frac{1}{n_A}\sum_{1}^{n_A}\max_{a\in A}(sim(a,b)) \quad \mathrm{A}\subset \text{Target set} \qquad (8)$$

where *sim*(a,b) is the similarity between each pair of elements and n_A is the cardinality of the set A, contained by the target.

7 COMPARING SIMILARITY MEASURES FOR SET-VALUED FEATURES

In order to compare and evaluate some previously described measures, we implemented them to estimate the similarity between datasets variables. The reported tests are not a systematic evaluation. They only exemplify results involving some critical situations to justify our decisions. We used six variables sets (VS_A, VS_B, VS_C, VS_D, VS_E and VS_F), whose main characteristics are shown in Table 2. For each variable we selected two features: the data type (symbolic) and the number of distinct values (numeric). The similitude between each pair of variables was determined applying the global measure (1) to aggregate the values derived, using: a similarity matrix for the symbolic feature; the normalized *Manhattan* similarity measure (2) for the numeric feature. To minimize the factors involved, the tests were performed using: (i) the same weight value 1 for all features; (ii) a similitude value 0 between distinct values of the data type feature. To obtain two resembling datasets, VS_B was derived from VS_A removing one variable. VS_D shares one (integer) similar variable with VS_A and VS_B, while VS_E shares one (string) similar variable with VS_A and VS_B. VS_C has variables of common data type with VS_A, VS_B and VS_D, although with different properties. One difference between the two groups {VS_A, VS_B} and {VS_D, VS_E} is the cardinality. VS_F does not share similar variables with VS_A and VS_B.

Table 3 sums up the results of the tests, using VS_A, VS_B and VS_C as target. For all the measures, the similitude is determined after computing the similarity between all pair of elements from the two sets. The difference among them comes afterwards, as shown on Table 4. The criteria used to evaluate the measures included the following requirements:

- R1 - the measure should reflect equal variables sets, i.e. the reflexivity property (sim(x,x)=1) should hold;
- R2 - the measure should distinguish pairs of sets that are far apart from pairs of sets that are closer to one another.

As expected, the *single linkage* algorithm's similarity measure (Sim_{SL}) is not suited to the current purpose, since it fails requirement R2. If we have at least one equal variable (same data type and number of distinct values), the differences among the other variables are not reflected (e.g. Table 3 cells (a,D) and (f,E)). The *average linkage* algorithm's measure (Sim_{AL}) is not a better option. First, it generally fails criteria R1. This measure provides low similarities, even using equal variables sets, since it includes in the average value the differences between all variables pairs. Besides, this influence might be greater than the one from equalities (e.g. cells (b,A) and (g,B)). Second, and by the same reason, the differences in similarity between variables sets very diverse and equal are not substantially significant (e.g. between (b,A) and (b,F) and between (g,B) and (g,F)), failing too the R2 requirement.

The Sim_{MR} measure is intentionally sensitive to cardinalities differences. The idea beyond is to achieve perfect similarity only if the cardinality of both sets is equal. To our application this property is a disadvantage. This fact is exemplified by the very low similarities of (c,D),(c,E),(h,D) and (h,E) cells, particularly when compared with the VS_F similarity, which should be lower. The approach does not fail the basic requirements, but provides results in an unexpected order and penalizes a difference that we do not want to emphasize. On one hand, in several DM methods the inclusion of a distinct number of variables is not a relevant factor to methods selection. On the other hand, we have a feature at dataset level to reflect this difference.

The similarity measures Sim_{MA} and Sim_{HMA} fulfill the requisites and seem to be the ones that most adequately reflect the intended semantic. The idea of Sim_{MA} is to provide the closest mappings using information from both sets. $Sim_{HM\mathrm{A}}$ aims at reflecting the best matches of the target set with every considered case, being intentionally asymmetric. Its main characteristic is the orientation by the target, which embodies the relevant properties that we want to retrieve. $Sim_{HM\mathrm{A}}$ is easier to compute and generally provides more distant similarity values. Sim_{MA} uses more data than Sim_{HMA}, but not necessarily more informative. An exception occurs when the case set overlaps the target set. Namely, Sim_{HMA} does not distinguish VS_A from VS_B, if VS_B is the target. VS_B-VS_A (j,A) is maximal since VS_A overlaps with everything found in VS_B. Conversely, the VS_A-VS_B (d,B) similarity is less, as VS_B only contains part of what is found in VS_A. This result accords to the intended one, in what respects to features as application area. A user looking for application area x would be satisfied by analysis including x and y. However, this kind of result is not proper when comparing dataset variables. Having common variables is a possible situation, since the same dataset is typically used in other alternative analyses, being important to distinguish variables sets as VS_A and VS_B. Hence, the Sim_{MA} measure was used to compare variables while Sim_{HMA} was adopted to compare the remaining set features.

Table 2: Dataset variables sets main properties.

Variable set (VS)	VS_A	VS_B	VS_C	VS_D	VS_E	VS_F
Number of variables	3	2	6	43	453	8
Data types	2 integer 1 string	1 integer 1 string	integer	1 integer 42 boolean	1 string 452 boolean	boolean
Number of distinct values	452, 44 42	452 42	8,240,171, 23, 60,83	452 2	42 2	2
Similar variables	2 common variables			1 with VS_A,VS_B	1 with VS_A,VS_B	

Table 3: Results of the similarity measures between variables sets.

Target	Similarity Measures	VS_A	VS_B	VS_C	VS_D	VS_E	VS_F	
VS_A	Sim_{SL} (4)	1	1	.99	1	1	.73	a
	Sim_{AL} (5)	.79	.76	.82	.65	.65	.65	b
	Sim_{MR} (6)	1	.67	.33	.06	.005	.24	c
	Sim_{MA}(7)	1	.95	.93	.74	.73	.71	e
	Sim_{HMA} (8)	1	.92	.87	.83	.76	.65	d
VS_B	Sim_{SL} (4)	1	1	.88	1	1	.73	f
	Sim_{AL} (5)	.76	.76	.75	.62	.61	.61	g
	Sim_{MR} (6)	.67	1	.20	.04	.003	.15	h
	Sim_{MA} (7)	.95	1	.81	.74	.73	.71	i
	Sim_{HAM} (8)	1	1	.81	.86	.76	.61	j
VS_C	Sim_{MA} (7)	.93	.81	1	.76	.74	.73	k
	Sim_{HMA} (8)	.95	.80	1	.80	.71	.69	l
		A	B	C	D	E	F	

Table 4: Similarity measures schematic example.

Sim(a,b) matrix				Sim_{SL}	Sim_{AL}	Sim_{MR}	Sim_{MA}	Sim_{HMA}
Target set	Case set			max $(u,v,$ $w,x,$ $y,z)$	$\frac{u+v+w+x+y+z}{3*2}$	$\frac{C1+C2}{3}$	$\frac{L1+L2+L3+C1+C2}{3+2}$	$\frac{L1+L2+L3}{3}$
		b_1 b_2	Max by line (L_i)					
	a_1	u v	L_1					
	a_2	w x	L_2					
	a_3	y z	L_3					
	Max by column (C_j)	C_1 C_2						
Target set	Case set			=	=	$\frac{L1+L2}{3}$	≡	$\frac{L1+L2}{2}$
		b_1 b_2 b_3	Max by line (L_i)					
	a_1	u v w	L_1					
	a_2	x y z	L_2					
	Max by column (C_j)	C_1 C_2 C_3						

8 CONCLUSIONS

The proposed and developed work aims at contributing to a more simplified, productive and effective exploration of WUM potentialities. The practice shows that often it is more efficient to solve a problem starting from a tested successful solution of a previous similar situation, than to generate the entire solution from scratch. This fact is particularly truth in the DM and WUM domains, where recurrent problems are quite common. To achieve this aim, we implemented a system, essentially founded on the CBR paradigm, which should suggest the more suited mining plans to one *clickstream* data analysis problem, given its high level description.

In this paper we described the similarity assessment approach, followed within the retrieval process, in order to cope with the multi-relational case representation. Structured representation and similarity assessment over complex data are important issues to a growing variety of application domains. It is a known fact that there is a trade-off between the expressiveness of the representation languages and the efficiency (complexity) of the learning method. The strategy of extending distance-based propositional methods through structured and typed representations, able to simplify the problem modelling, and treating the features and theirs properties in the similarity measures is advantageous. It is simple, enables to benefice from the research and the efficiency from these methods, exploring at the same time the greater expressiveness of such representations. Since this strategy is suited to our current demands, it was adopted to handle the faced issues.

We considered specifically the issue of measuring the similarity between sets of elements. There are multiple proposals in the literature to deal with this issue, but an ideal and general approach, appropriate to several purposes such as the intended semantic and properties, does not exist. Consequently, we explored a number of different already defined similarity measures and we extended one of them to better fit our purposes. This extension gave raise to two measures suited to the similarity assessment of features with different properties.

We are currently working on the construction of more cases, comprising WUM process with higher complexity. Afterward, a more detailed and systematic experimental evaluation of the system is necessary. Moreover, one future direction of work concerns the weights assignment improvement, based on a comprehensive evaluation of the features relevance and discriminating power.

ACKNOWLEDGEMENTS

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

REFERENCES

Bergmann, R., 2001. Highlights of the European INRECA projects. In *ICCBR'01, 4th International Conference on CBR,* Springer-Verlag, 1-15.

Bergmann, R., Stahl, A., 1998. Similarity Measures for Object-Oriented Case Representations. In *EWCBR'98, 4th European Workshop on Case-Based Reasoning*. Springer-Verlag, Vol. 1488, 25-36.

Bohnebeck, U., Horváth, T., Wrobel, S., 1998. Term Comparisons in First-Order Similarity Measures. In *8th International Conference on Inductive Logic Programming*, Vol. 1446, Springer-Verlag, 65-79.

Duda, R., Hart, P., Stork, D., 2001. *Pattern Classification and Scene Analysis*, chapter Unsupervised Learning and Clustering. John Willey and Sons.

Eiter, T., Mannila, H., 1997. Distance Measures for Point Sets and their Computation. *Acta Informatica*, 34(2), 109–133.

Emde, W., Wettschereck, D., 1996. Relational Instance-based Learning. In *13th International Conf. on Machine Learning*, Morgan Kaufmann, 122-130.

Gregori, V., Ramírez C., Orallo, J., Quintana, M., 2005. A survey of (pseudo-distance) Functions for Structured-Data. In *TAMIDA'05, III Taller Nacional de Minería de Datos y Aprendizaje*, Editorial Thomson, CEDI'2005, 233-242.

Flach, P., Giraud-Carrier, C., Lloyd, J., 1998. Strongly Typed Inductive Concept Learning. In *8th International Workshop on Inductive Logic Programming*, Springer-Verlag, Vol. 1446, 185-194.

Hilario, M., Kalousis, A., 2003. Representational Issues in Meta-Learning. In *ICML'03, 20th International Conf. on Machine Learning* , AAAI Press, 313-320.

Kirsten, M., Wrobel, S., 1998. Relational Distance Based Clustering. In *8th Int. Conf. on Inductive Logic Programming*, Vol. 1446, Springer-Verlag, 261-270.

Kirsten, M., Wrobel, S., Horvath, T., 2001. Relational Data Mining. *Distance Based Approaches to Relational Learning and Clustering*, Springer-Verlag, 212-232.

Kolodner, J., 1993. *Case Based Reasoning*. Morgan Kaufmann, San Francisco, CA.

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

Wanzeller, C., Belo, O., 2006. Selecting Clickstream Data Mining Plans Using a Case-Based Reasoning Application. In *DMIE'06, 7th International Conference on Data, Text and Web Mining and their Business Applications and Management Information Engineering*, 223-232.