A Customizable Similarity Measure Between Histological Cases
Marie-Christine Jaulent PhD1, Adil Bennani1, Christel Le Bozec M.D. PhD1, Eric Zapletal1 and Patrice Degoulet MD. PhD1

1 SPIM, ERM 202 INSERM, Faculté de Médecine, 15 rue de l’école de Médecine, 75005 Paris, France

ABSTRACT

IDEM, a computerized environment dedicated to pathologists, includes a Case Based Reasoning (CBR) procedure to retrieve similar histological cases in the database. The relevancy of a retrieved case strongly depends on the similarity measure comparing case descriptions. The present work deals with the definition of a similarity measure in the context of IDEM. In a first step, a theoretical measure (relational, numerical and informed), based on the domain constraints, was selected. In a second step, the theoretical measure is optimized according to the current case base. Results are presented for a database of 33 cases of breast tumors. The contribution of this work is to give to pathologists an interactive environment that optimizes the similarity measure between histological cases. This work is also a contribution to the CBR cycle life since the similarity measure can be adapted while new cases are added to the base.

INTRODUCTION

In pathology, diagnoses are based on the identification of diagnostic criteria in histological images. Given the abundance of histological patterns, pathologists frequently face complex decision making problems which require expert knowledge, based on personal and long-term professional experience. IDEM is an integrated computerized environment dedicated to pathologists including a case-based reasoning mechanism [1]. The expert knowledge is embedded within a case base (the knowledge base). A case is composed of images, along with a structured description of the morphological features present in images and the diagnosis. Case Based Reasoning (CBR) systems reuses the solution of a solved case to build a solution of a new problem. This approach has been shown to provide an “intelligent” retrieval of reference images and diagnostic clues [2]. According to the generic model of CBR proposed in [3], the main task is “Reason with past cases” and includes several sub-tasks: “Elaborate a case”, “Retrieve a case”, “Reuse a case”, “Revise a case” and “Retain a case”. When the system is given a new problem to be solved (the target problem), it elaborates its description and retrieves a case from a case base (the source case). Then, the system reuses the solution of the source case by adapting it for the target problem and retains the new case in order to make it available for further problem solving. The outcome of the reasoning processing depends mostly on the retrieved source case.

The problem related to the relevancy of the source case is twofold:
- Is the case description model adapted to the retrieval of the most relevant source case? It concerns the description of the case itself and the description of the implicit and explicit links between the description and the diagnosis (4)
- Is the similarity measure adapted to the retrieval of the most relevant source case? It concerns the definition of an optimal measure for a given case base.

The present work deals specifically with the definition of a similarity measure between histological cases and is a contribution to the IDEM environment.

Unfortunately, there is no universal rule to choose the best similarity measure for a given application [5]. Moreover, in pathology, the determination of a similarity measure is complicated because of inherent problems such as 1) the multiplicity of diagnostic criteria leading to the same diagnosis, 2) the lack of reproducibility in identifying, localizing and labeling relevant morphological features in images corresponding to diagnostic criteria. For these reasons, we think that context-sensitive measures could be appropriate and we propose a methodology in two phases to determine this measure. In a first phase, a theoretical measure is defined only depending on the case description model of IDEM. However, this measure do not provide necessarily optimal results with different case bases potentially available in the IDEM environment. In a second phase, the measure is optimized using the results of the theoretical measure for a given case base. A general interface has been designed to easily fine tune parameters of the similarity measure while the case base is evolving. Results from fifty-three cases of breast pathology are presented and discussed to illustrate the gain obtained by the optimized measure compared to the theoretical one.

BACKGROUND

CBR systems typically retrieve cases from the case base by applying similarity measures. The measures are usually constructed in an ad hoc manner depending on the structure of the case [6]. For instance, Sebag and al. have proposed a rule-based similarity measure when cases are described as rules [7]. In [8], the authors define a weighted similarity to compare structured cases. Concerning similarity between relational structures, Bisson has mentioned the necessity to determine the weights using context knowledge [5]. Weights can be automatically determined using statistical or symbolic machine learning approaches [9]. We adopt the terminology found in [10]:
A similarity function is modeled by a 4-uple \((L_d, L_s, \Gamma, F_s)\).
Let \(L_d\) be the knowledge representation language for the description of cases.
Let \(L_s\) be the representation language for the similarities.
Let \(\Gamma\) be some meta-knowledge about the domain
Let \(F_s\) be the similarity function such that: \(F_s = L_d \times L_d \rightarrow L_s\).
The instantiation of the similarity parameters \(L_d, L_s, \Gamma\) leads to different classes of similarity measures:

**Propositional versus relational similarity**
Propositional similarity applies when \(L_d\) is an attribute-value representation language. The similarity between two cases is then an aggregation (eventually weighted) between similarities computed at the level of each attribute.
Relational similarity applies when \(L_d\) is an object-oriented representation language. In this representation, a case is a hierarchical set of objects (a graph). The similarity measure is then an aggregation between a surface similarity between the description of two objects, characterized by specific attributes, and a structure similarity based on graph matching procedures.

**Symbolic versus numeric similarity**
Symbolic similarity applies when \(L_d\) is a discrete set such as \{yes, no\} or \{little, intermediate, much\}. Numeric similarity applies when \(L_d\) is a subset of \(\mathbb{R}\), often \([0, 1]\). The value 1 expresses a full resemblance and the value 0 a full dissemblance.

**Informed versus non-informed similarity**
Additional knowledge \((\Gamma)\) may be available on the context.
Two similarity families are distinguished according to the fact that this knowledge is used or not.
Non-informed similarity applies when \((\Gamma)\) has a neutral effect on the similarity function \((F_s)\). Informed similarity applies when \((\Gamma)\) influences the similarity function \((F_s)\).
This can be done in two ways. On one hand, \((\Gamma)\) may be useful to select \(F_s\) according to the goal of the application.
On the other hand, \((\Gamma)\) may be directly used by the similarity function \(F_s\).

Each MF is characterized by three attributes which universes of discourse are defined within \(L_d\) as subsets of the controlled vocabulary of the Systemized Nomenclature of MEDicine (SNOMED) [11]. Each MF is associated to a free linguistic description -DS- (e.g. malignant epithelial cell), a standardized anatomic term -DA- (e.g. epithelial cell) and/or a standardized pathologic term -DP- (e.g. malignant). Anatomic terms are selected from the Topography (anatomy) nomenclature axis of SNOMED, and pathologic terms are selected from the Morphology (pathology) axis. The figure 1 shows an example of a hierarchical description of the case.

**MATERIAL**
In pathology, a case is the basic element of the Patient Medical Record. A pathologic case is associated, 1) to the set of relevant images and 2) to the pathologist(s) description(s) of this set of images.
A pathologist description consists in a collection of relevant **Morphological Features** (MFs) observed in the set of images of the case.
The MF is the key element of the case description. Each MF is associated to one or several **regions of interest** (ROIs) in images. One specific MF may be seen at different places in one image or in several images (with for example, different magnifications on several images).

Knowledge about similarities between anatomic (resp. pathologic) terms is embedded in a heuristic anatomic (resp. pathologic) numerical similarity table [1]. The DS attribute is not taken into account in the computation of the similarity between histological cases. The figure 2 provides a partial view of the similarity table between pathological terms (DP).

**Figure 1. Example of the hierarchical description of a case.**

**Figure 2. Partial view of the DP similarity table**
METHODS

Theoretical measure in the context of IDEM

From the material and the previously described formalism of similarity measures, we define a relational, numerical and informed similarity as the theoretical measure in the IDEM framework. $L_d$ is here an object-oriented representation language. $L_S$ is the sub-set $[0,1]$ of $R$. The meta-knowledge $\Gamma$ will be precise in the following.

The similarity measure has a recursive construction. This construction is illustrated by the following example. Let $N$ be a new problem and $C$ a case in the case base. The figure 3 draws the tree structure of these two cases. Let $Sim_{Case}$ be the global similarity between two cases and $Sim_{MF}$ be the similarity between two MFs. The global similarity is defined by:

$$Sim_{Case}(N,C) = Sim_{MF}(N_0, C_0)$$

This analysis includes fifteen subjects (5 novices, 5 intermediates, and 5 experts) selected from a total of thirty subjects who were recruited for participation. Each subject examined 4 cases of breast pathology.

At the level of the MF, the similarity is made of a surface similarity, $Sim_{Str}$ – and a structure similarity $Sim_{St}$ –:

$$Sim_{Str}(N_i, C_j) = OP(Sim_{St}(N_i, C_j), Sim_{Str}(N_i, C_j))$$

where OP is an aggregation operator between the surface and structure similarity. OP can be a minimum operator, a weighted sum operator, etc. It is part of the knowledge $\Gamma$.

The Surface similarity between two MFs is an aggregation of the similarity between the two corresponding anatomic terms (DA) – $Sim_{DA}$ – and the similarity between the two corresponding pathologic terms (DP) – $Sim_{DP}$ –. The values of $Sim_{DA}$ and $Sim_{DP}$ are directly extracted from the similarity tables:

$$Sim_{Str}(N_i, C_j) = OM(Sim_{DA}(DA_i, DA_j), Sim_{DP}(DP_i, DP_j))$$

Where $DA_i$ and $DP_i$ (resp. $DA_j$ and $DP_j$) are the anatomic term and pathologic term associated to $N_i$ (resp. $C_j$) and where OM is an aggregation operator between the DA and DP similarity. OM can be a minimum operator, a weighted sum operator, etc. It is part of the knowledge $\Gamma$.

The Structure similarity between to MFs is based on a tree matching algorithm. The procedure consists in finding the best matching between the hierarchical structures below the two MFs. We define $t$ as the number of levels that the algorithm should take into account below the two MFs under consideration. When $t$ is equal to 1, only the direct children of the MFs are integrated in the structure similarity. The value of $t$ can be interpreted as a tolerance degree and is part of the knowledge $\Gamma$.

The algorithm is illustrated from the example in figure 3 with $t = 2$. For the two MFs, $N_0$ and $C_0$, the different matching solutions are:

$\{(N_1, C_1), (N_2, C_2), (N_1, C_1), (N_2, C_1), (N_1, C_22)\}, \{(N_2, C_1), (N_1, C_2)\}}$. The last solution is the one illustrated in figure 3.

More generally, given two MFs $N_i$, $C_j$, let $A_{ij}$ be the number of elements in each matching solution (in the example $A_{0,0} = 2$), let $K_{ij}$ be the number of the different solutions (in the example $K_{0,0} = 4$, for $t = 2$) and let $n$, $c$ be two variables used to designed any couple of children from $N_i$ and $C_j$ for a given solution indicated by $k$. The structure similarity between $N_i$, $C_j$ is defined by:

$$Sim_{St}(N_i, C_j) = \frac{K_{ij}}{\max_k \left( \frac{\sum k_{ij}}{A_{ij}} \right)}$$

Finally, $\Gamma = \{OP, OM, t\}$

Optimization of the similarity measure

The recursive approach presented above has the drawback of combinatorial complexity with increasing values of $t$. Some general knowledge about the domain was added to $\Gamma$ to eliminate impossible matching solutions in order to alleviate the process. For instance, it is useless to compute the similarity between an MF with DA = cell and DP = proliferation and an MF with DA = gland and DP = dilatation since these two MFs has no semantic connection. This kind of knowledge is directly used by the similarity function $F_s$.

The previous theoretical measure has been designed to answer the constraints of the IDEM domain. However, it has to be optimized according to the diagnosis classes represented in the case base. Indeed, for a given base, the best similarity measure is the one that maximizes the intra-class similarity and minimizes the inter-class similarity within the base. While adding new solved cases in the base, there is no guarantee that the similarity remains the best one. We designed a general interface to allow an optimization procedure based on the interactive choice of the OP and OM operators. For the two operators, the different possible choices are the minimum, the maximum and the mean. These operators can be weighted operators in order to take into account differences between the influence of DA and DP in the surface similarity (OM) or between the influence of surface similarity and structure similarity in the global similarity (OP). In the current stage of the work, the user chooses the operator type and the associated weights directly through the interface.
However, for the OM operator, statistical approaches or machine learning approaches could be used to extract automatically these weights according to all the values of DA and DP in the current base [9].

RESULTS

53 pathological specimen were selected to constitute the case base. These cases correspond to confirmed reference cases in a french histopathology department (Institute Gustave Roussey). They include 8 disjoint diagnostic categories : A ("adenosis"), B ("ductal carcinoma"), C ("radial scar"), D ("papillary carcinoma"), E ("benign epithelial hyperplasia"), F ("fibroadenosis"), G ("benign phylloide tumor") and H ("tubular carcinoma"). In the current state of the work, the optimization process of the similarity measure has been tested on the cases of the base itself, that is, each case is considered in turn to be the new problem to solve. The idea was at first to verify that a case is most similar to the cases having the same diagnosis than to those having other diagnosis using the theoretical similarity measure. The results are registered in the table 1. A number in the table is a global degree $S_{inter\_class}$, ranging from 0 to 1 and defined as follows. Let $n$ (resp. $p$) be the number of cases C$_a$ (resp. C$_b$) in class A (resp. B). The degree $S_{inter\_class}(A,B)$ corresponds to the mean of the global similarity obtained for the cases of the two diagnostic classes A and B taken two by two and is expressed by :

$$S_{inter\_class}(A,B) = \frac{\sum_{i \in (1,n), j \in (1,p)} Sim_{case}(C_{ai}, C_{bj})}{np}$$

For every class A, one can define the intra class similarity by :

$$S_{intra\_class}(A) = \frac{\sum_{i,j \in (1,n), i \neq j} Sim_{case}(C_{ai}, C_{aj})}{(n-1)!}$$

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.66</td>
<td>0.55</td>
<td>0.44</td>
<td>0.6</td>
<td>0.48</td>
<td>0.54</td>
<td>0.62</td>
</tr>
<tr>
<td>B</td>
<td>0.55</td>
<td>0.91</td>
<td>0.55</td>
<td>0.79</td>
<td>0.44</td>
<td>0.81</td>
<td>0.71</td>
</tr>
<tr>
<td>C</td>
<td>0.44</td>
<td>0.55</td>
<td>0.78</td>
<td>0.52</td>
<td>0.47</td>
<td>0.55</td>
<td>0.6</td>
</tr>
<tr>
<td>D</td>
<td>0.6</td>
<td>0.79</td>
<td>0.52</td>
<td>0.8</td>
<td>0.44</td>
<td>0.7</td>
<td>0.64</td>
</tr>
<tr>
<td>E</td>
<td>0.48</td>
<td>0.44</td>
<td>0.47</td>
<td>0.44</td>
<td>0.72</td>
<td>0.41</td>
<td>0.46</td>
</tr>
<tr>
<td>F</td>
<td>0.54</td>
<td>0.81</td>
<td>0.55</td>
<td>0.7</td>
<td>0.41</td>
<td>0.88</td>
<td>0.65</td>
</tr>
<tr>
<td>G</td>
<td>0.62</td>
<td>0.71</td>
<td>0.6</td>
<td>0.64</td>
<td>0.46</td>
<td>0.65</td>
<td>0.8</td>
</tr>
<tr>
<td>H</td>
<td>0.58</td>
<td>0.8</td>
<td>0.69</td>
<td>0.76</td>
<td>0.52</td>
<td>0.7</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 1 : The mean similarities for the diagnostic classes

The results show that for each class, the intra-class similarity is higher than its inter class similarity with the other classes. The second aspect was to tune the similarity measure through the interface (figure 4) and to verify that the intra class similarity is enhanced by the optimization of the measure. The figure 5, illustrates the gain in of the intra_class similarity for the diagnosis "benign epithelial hyperplasia" after the optimization of the measure.

The choice of a relational measure is naturally adapted to our domain where cases are treelike structures. Relational measures were used in several systems [12] [10]. Moreover, the similarity measure has to be recursive since...
the tree structure of a case is not fixed. Classical algorithms on tree searching do not apply and we introduced a graph matching inspired by Sanfeliu and Fu [13]. The introduction of a level of flexibility in the matching process is mandatory in a domain where two experts may have a different vision for the structure of a case. This last point explains the main limit of this approach, namely the combinatorial explosion for large structures. Indeed, the number of matching increases with the depth of the tree-like structure and the required level of flexibility. Nevertheless, this theoretical measure remains acceptable in the context of IDEM for cases do not contain more than 15 or 20 MFs.

On one hand an optimization of performances by the elimination of useless matchings is realized. This elimination is based itself on knowledge connected to the types of anatomical and pathological attributes. Such knowledge do not modify the theoretical measure but simply reduce the time of calculation of the similarity.

On the other hand an optimization of results can be obtained through the developed interface, in setting directly the parameters of the similarity measure. The procedure allows to improve in real time the relevancy of the similarity according to the current case base.

An another important parameter is the similarity table. An heuristic quantitative representation of proximities between qualitative terms is always disputable. An approach using the notion of semantic distance between medical terms could be useful to valid the similarity tables contents [14].

The present work is also a contribution to the CBR cycle life since the similarity measure can be adapted while new cases are added to the base. Some automation is needed to define the weights by learning. One possible approach is the following : when an example is well-classified by the CBR system (in comparison with the real class), the weights associated to the attributes which tended to predict this class are slightly increased [15],[16],[12].

The main result of this work is to supply to the potential users an interactive environment allowing them to optimize the similarity measure according to the case base. The theoretical similarity measure is defined in a general way for the studied category of problems and only the optimization becomes an ad-hoc procedure.

REFERENCES


ACKNOWLEDGEMENTS

This work was supported in part by a grant from the Association pour la Recherche sur le Cancer