https://dx.doi.org/10.24001/ijaems.3.6.11 *ISSN: 2454-1311*

A new structural similarity measure: clustering of multi-structured documents

Ali Idarrou

IRF-SIC Ibn Zohr University, Agadir Morocco
Email : ali.idarrou@uiz.ac.ma

Abstract— *This paper is part of the continuity of our work on the structural clustering of multi-structured multimedia documents. One of the major problems of our work is how to compare two multi-structured documents, and therefore to compare document structures to be able to identify the resemblances between structures and transformation rules of a structure to another (evaluation of a processing cost). We have defined a new structural similarity measure for identifying common substructures in two multimedia documents, taking into account constraints of such documents (relations between components, order of components, etc). In our previous work, we have studied the impact of the sub-process of "filtering" of our clustering process on the quality of the generated classes. In this work, we describe the sub-processes of transformation of a structure to another and we propose a measure for evaluating the cost of a structural transformation.*

We evaluate our approach on a corpus of documents extracted randomly from the INEX 2007 corpus and a corpus composed of the notices of books (in XML format) from the library of the Toulouse 1 Capitole University.

Keywords— multimedia document, structural clustering, structural similarity measure.

I. INTRODUCTION

The multimedia information is available in large quantities and in different formats (text, image, sound, etc.). However, this source of information would be useless if our ability to effectively access does not increase too [12]. It is therefore necessary to have automatic tools for quick access to desired information, thus reducing user effort. Automatic classification is a solution that allows you to organize a large collection of documents, to reduce the search space and consequently to improve the performance of the access to information process. The problem, given a set of documents, is how to group these documents in clusters form of similar documents? This problem causes several issues such as: How to represent these documents? How to evaluate the similarity between two documents?

A multimedia document is composed of several objects of various natures: image, text, sound, etc. It is essentially multi-structured, coming from the composition of several sub-documents, which are themselves more or less complex because each sub-document has one or more structures. The complexity of multimedia documents to multiple structures involves a problem related to their representation. Indeed, the multi-structurality induces complex and multiple relationships between the same two components of a document. In [15], the model *MVDM* "Multi Views Document Model" of [3] allows a rich representation of the mutli-structured documents and that this richness can be exploited to classify these types of documents. To classify structurally multimedia documents to multiple structures, we continue within *MVDM* and we consider that the document structure is sufficiently a discriminating factor for classification.

Comparing two documents requires modeling these documents in a formal manner and using an appropriate measure for evaluating the similarity between these documents. The chosen model must be able to express the maximum of information on the documents to compare effectively. In [17], more modeling of documents will be more sophisticated and the comparison of these documents will be accurate but difficult. We are interested in the representation of multimedia documents using graphs. Comparing structurally two documents is therefore comparing the graphs that represent them.

The geometric models and attribute-based models don't allow the comparison of structured objects [6] and [7]. We have defined a similarity measure for identifying common sub-structures in two multimedia documents, taking into account constraints related to this type of documents (relations between components, order of components, etc). Thus, the new similarity measure proposed is based on matching graphs. This is a structural measure and not a surface one. Indeed the surface measure is based on the descriptive properties of objects while the structural similarity measure between objects is based on the relationships between these objects [5].

In our previous work [9], we have studied the impact of the sub-process of "filtering" of our clustering process on the quality of the generated classes. In this work, we describe

https://dx.doi.org/10.24001/ijaems.3.6.11 *ISSN: 2454-1311*

the sub-processes of transformation of a structure to another and we propose a measure for evaluating the cost of a structural transformation.

This paper is organized as follows: In the next section, we present two examples of work that used tree transformations representing the documents. In the third section we present our similarity measure. In the fourth section, we present the *MVDM* model. We define, in the fifth section, our approach to structural classification of multi-structured multimedia documents and describe the sub-processes of transformation of one structure to another. Before concluding, we present in the sixth section our experiments.

II. RELATED WORKS

In their approach to structural classification of documents, the authors of [2] have used summary trees obtained by transformation (depth reduction, elimination of repeated nodes, etc.). However, these transformations may cause a loss of semantic and contextual information. For example, the reduction of the depth ("Fig.1") involves the elimination of components and relations between these components. Indeed, the relation (*A*, *P*) of T_1 can't play the same role as the relationship (*A*, *P*) of T_2.

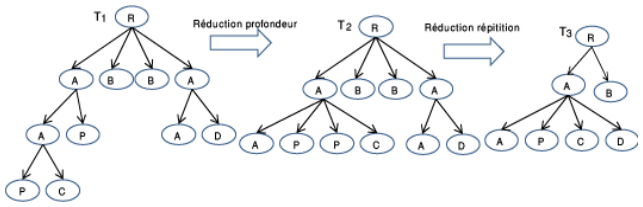

Fig. 1 - Extraction of structural summary [2]

In [16], XML documents are represented as a tree, which is considered as a set of paths. Thus, the classification is based on the calculation of the frequency of these paths. The idea of linearization trees proposed in this work is very interesting. In contrast, pretreatment steps, which include reducing the number of paths, and the filtering of tags that can cause loss of information, which can have a negative impact on the quality of the classifier.

In the next section, we present our similarity measure.

III. A NEW STRUCTURAL SIMILARITY MEASURE

Conventional systems comparison return a value indicating that the two compared objects are similar or not. However, in most applications, it is interesting to have more details on the proximity of the objects being compared. We are interested in the category of systems that evaluate the proximity between two objects from a continuous value to quantify the similarity and the difference between these two objects.

In [11], we have proposed a new measure of structural similarity based on the graph matching. This measure reflects the structure of graphs compared in the sense that comparing the paths graphs taking into account both the position of the nodes, the order of brother nodes and relationships between those nodes. In our context, we consider that the position of the nodes and the relationships between these nodes are two essential parameters in a process of structural comparison of multimedia documents. Thus, the weighting function that we proposed (formula [6]), upon which our similarity measure is based, reflects these two parameters "Fig.2".

In graph theory, the comparison of graphs is a combinatorial problem. To reduce this combination, we chose to consider a graph as a set of paths [9]. Comparing two graphs is therefore comparing the paths that compose them. In the example in "Fig.2", the graph G_2 is composed of paths: *A/B*, *A/D/E/K*, *A/D/E/H*, *A/D/A* and *W/H*.

To evaluate the structural similarity *Sim* (*G*, *G* ') between two graphs *G* and *G'* oriented, labeled and ordered (*G* = (*V*, *E*) and *G* '= (*V'*, *E* ')), we defined the following measure:

$$Sim(G,G') = 1 - Dist(G,G') \quad [1]$$

$$where\ \ Dist(G,G') = \frac{d_{GG'} + d_{G'G}}{2} \quad [2]$$

$$and\ d_{GG'} = \frac{1}{n} \sum_{i \in [1,n]} d_{Inc}(chm_i, G') \ \ and\ d_{G'G} = \frac{1}{n'} \sum_{j \in [1,n']} d_{Inc}(chm'_j, G) \quad [3]$$

where $d_{GG'}$ (resp. $d_{G'G}$) : is the alignment distance between *G* and *G'* (resp. between *G'* and *G*), *n* and *n'* (not nuls) are respectively the number of paths of *G* and *G'*.

d_{Inc} : allows to evaluate the degree of inclusion of a path in a graph.

$$d_{Inc}(chm, G') = \min_{k \in [1,n']} \left[\frac{\sum_{e_j \in chm} |P_e(e_j) - w_{j,k}|}{\sum_{e_j \in chm} P_e(e_j)} \right] \quad [4]$$

$$where\ w_{j,k} = \begin{cases} P_e(e'_h)\ if\ \exists\ e'_h \in chm'_k\ /\varphi_e(e_j) = e'_h\ ;\ chm'_k\ is\ a\ path\ of\ G' \\ 0\ \ otherwise \end{cases} \quad [5]$$

- *n'* : the number of paths of *G'*,
- φ_e : bidirectional alignment function of relations from *E* (resp. from *E'*) to *E'* (resp. to *E*) which allowing to align two similar arcs:

$$\varphi_e : E \rightarrow E'$$
$$a \mapsto \varphi_e(a) = a'$$; where the arc *a'* is similar to the arc *a*.

https://dx.doi.org/10.24001/ijaems.3.6.11 *ISSN: 2454-1311*

and P_e the weighting function allowing to weight the graphs, it's defined by:

$$P_e : E \rightarrow]0,1[$$
$$(u,v) \mapsto P_e(u,v)$$

$$P_e(u,v) = \begin{cases} 1-\frac{\alpha}{k} \text{ if } prof(v)=1 \\ P_e(x,u)-\frac{\alpha}{k^{prof(v)}} \text{ otherwise } ; x \in père(u) \end{cases} \quad [6]$$

- $x \in père(u)$: u can have many node fathers (eg "Fig.2", nodes H of G_2).
- $prof(v)$: depth of v : position in a path,
- k (*a power of* 10) is a parameter indicating the maximum number of son nodes for each node (*maximum number of son*<k) depending on the nature of the collection of documents processed,
- α is a parameter that depends on the type of node v:

$$\alpha = \begin{cases} 1 \text{ if } v \text{ is an attribute or metadata} \\ ord(v) \text{ otherwise } (ord(v) : \text{order of the node } v; \text{ its position relative to its brothers nodes}) \end{cases}$$

In comparison process document structures, we believe that the information provided by the structural relationships is of key interest and that two documentary structures composed of the same elements, do not necessarily mean that they are similar. According to the theory of mapping developed by [4], the best analogies are those based on relationships between entities rather than their descriptive properties.

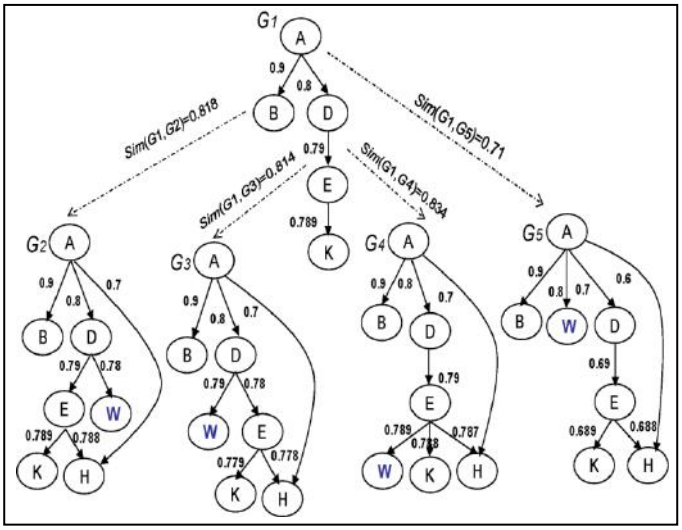

Fig.2 - Example of graph similarity

The structural similarity between each pair of graphs:

$Sim(G_1,G_2)$ = 0.818, $Sim(G_1,G_3)$ = 0.814, $Sim(G_1,G_4)$ =0.834 and $Sim(G_1,G_5)$ = 0.71.

In this example, the difference between $Sim(G_1,G_2)$, $Sim(G_1,G_3)$, $Sim(G_1,G_4)$ and $Sim(G_1,G_5)$ is explained by the fact that the proposed measure takes into account the distribution of structural elements in the graphs compared. We notice a difference, which becomes important in the case of $Sim(G_1, G_5)$, between the values of similarities as a result of differences in positioning some nodes, in particular the node W (different order or different level). This shows that the proposed similarity measure is taking into account of two parameters depth and order, penalizing differences in depth.

In the next section, we present the *MVDM* model.

IV. PRESENTATION OF *MVDM* MODEL

The *MVDM* model introduces the notion of view: set of structural nodes and relationships between nodes. A node can be simple or complex. In the latter case, the node can be considered as a sub-document itself can be split into a set of nodes and relationships between nodes. There may be more than one possible relationship between two components of a document. This allows materializing several organizations for this document. According to this model, the notion of document structure can be encompassed within a large concept which is that of "view". A specific view corresponds to a particular organization or a view of a document. It represents one of the structures of a multi-structured document [3]. For example in "Fig.3", the specific view Vsp_1 is a description by a speaker of an audio document while the specific view Vsp_2 is a description by emission of the same document. These two views are aggregated into a single logical structure of the document "*audio_doc*".

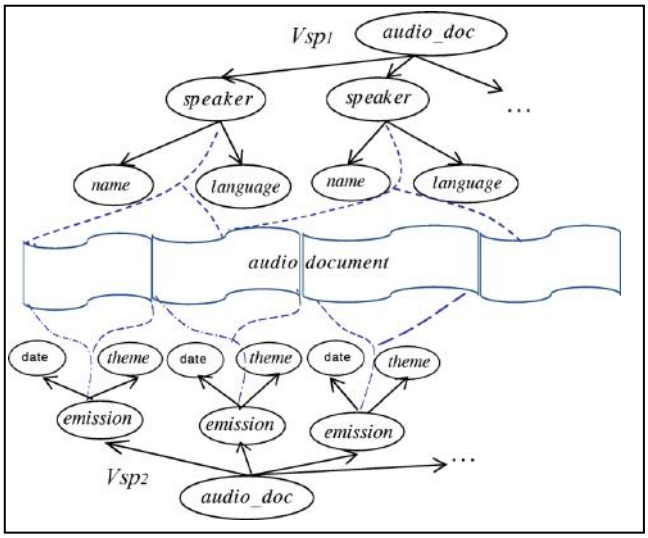

Fig.3 - Two descriptions (two views) of the same audio content

The *MVDM* model is composed of two layers: a specific layer (DW_{sp}: "Fig.4") where each specific view, characterizing the organization of a particular document, is represented in tree form and a generic layer (DW_g: "Fig.4")

https://dx.doi.org/10.24001/ijaems.3.6.11 *ISSN: 2454-1311*

where the generic views (comparatives clusters) are represented by graphs. A generic view (graph) represents a collection of specific views that are structurally similar ("Fig.5").

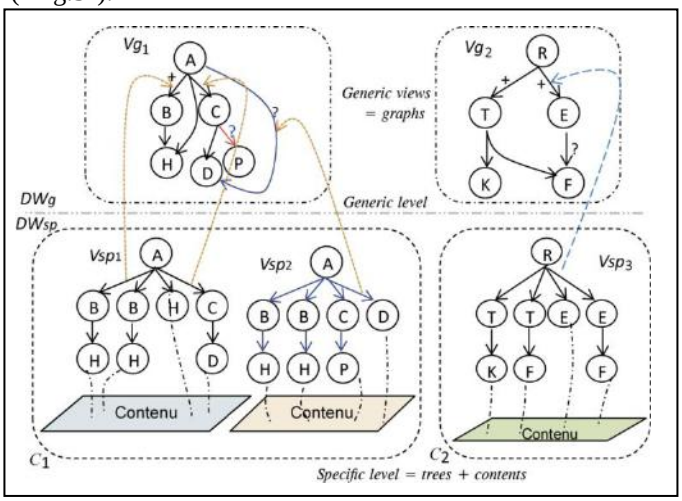

Fig. 4 - Example of documentary Warehouse: DW

The cluster representatives (Vg_i)) are indexes that allow interaction with a large collection of documents from various sources which are generally very heterogeneous. Indeed, access to the representative of a class can access a targeted manner to the sub-collection of documents DW_{sp}, represented by it.

In the next section, we present the definition of a documentary classification.

V. DEFINITION OF A DOCUMENTARY CLASSIFICATION

In the framework of *MVDM*, the problem of document classification results in a problem of attaching a given specific view to the generic view (the generic level: DW_g) structurally similar. The choice of the generic view to what the specific view must be attached is based on comparing it to all the generic views of the documentary warehouse.

We are interested in the structural classification (of document structures), considering that the structure is an interesting discriminating factor for classification. Thus, the structural classification in the sense that we understand [9] allows creating, in a documentary warehouse, clusters called generic views. A generic view is a tree superposition representing the structures of documents; it is enriched (transformation) as one goes along the classification. This superposition generates a structure of rooted graph ("Fig.4"). It is not a simple summary, as is the case of the works using summarized trees to represent the documents, but rather a rich description (without loss of information) representing a set of specific structures that are structurally similar. The question we address in this section is how to build the generic views (clusters) in the framework of *MVDM*?

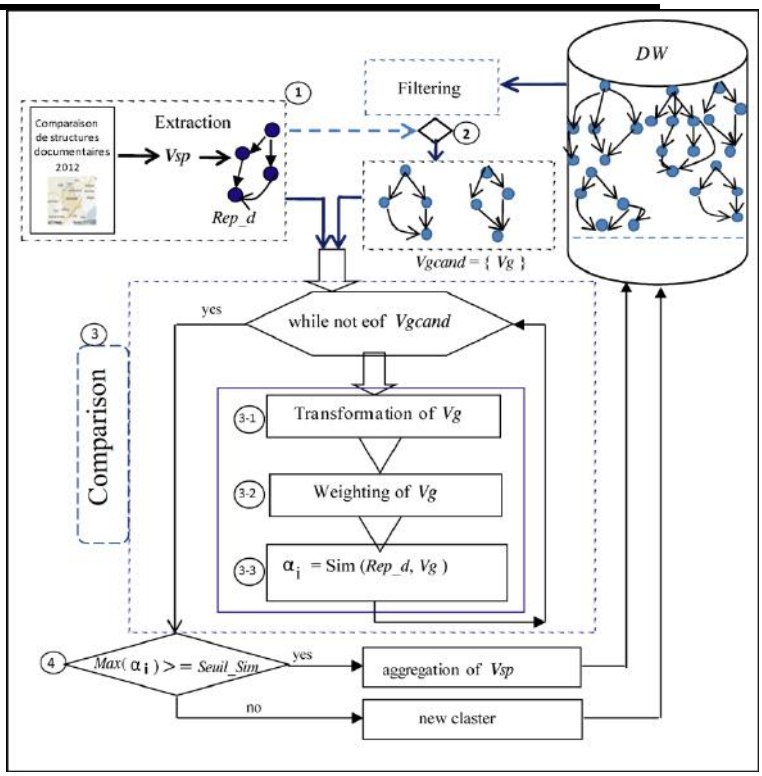

Fig.5 - Process of structural classification

Where *Vgcand* is a set of generic views of documentary warehouse is likely to be similar with the specific view *Vsp* in input. *Rep_d* is the representative of the specific view of the document to integrate. After the filtering process, the set of views *Vgcand* candidates for comparison will be used for the following stages of the comparison process.

The construction of generic views (clusters representatives) goes through a comparison process (step 3 in "Fig.5"). Before calculating the similarity between *Rep_d* view in input and the generic views of *Vgcand*, we apply a transformation process that enriches each *Vg* candidate and make it more generic (most representative).

In our previous work [8], we have described the steps of our clustering process. In the next section, we present the sub-process of transforming a generic view (graph) to another and we describe our transformation process and we study the impact of this transformation on classification.

Transformation of generic views

a) Principe

The aim of the transformation is to render the views representatives of clusters the most representative and therefore optimize the storage volume of documentary warehouse. This step allows bringing closer each generic view to the representative *Rep_d* of the specific view of the document to integrate. Possible additions of fragments of *Rep_d*, missing in each of candidate views, may be considered. Unlike approaches of [2] and [16], in our

https://dx.doi.org/10.24001/ijaems.3.6.11 *ISSN: 2454-1311*

approach we respect both the order of the nodes (and arcs) and the preservation of arcs (eg. "Fig.7") without loss of components of the graph transformed (without information loss).

An example of transformation of a structure represented using a graph

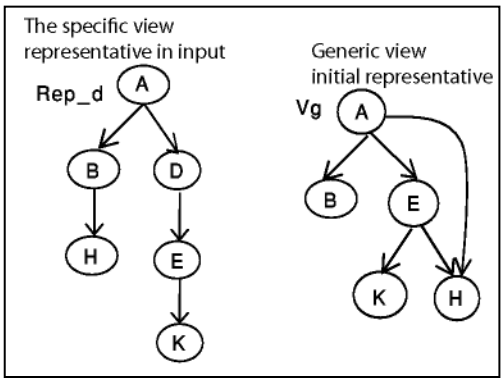

Fig.6: An example of graph comparison

In "Fig.7", the paths chm_1=*A/B/H* et chm_2=*A/D/E/K* of *Rep_d* have a degree of similarity respectively with the paths chm'_1=*A/B* et chm'_2=*A/E/K* of *Vg*. In this example, fragments (*B, H*), (*A, D*) and (*D, E*) of the view represented by *Rep_d* (missing in the generic view *Vg*) are added: adding nodes and arcs.

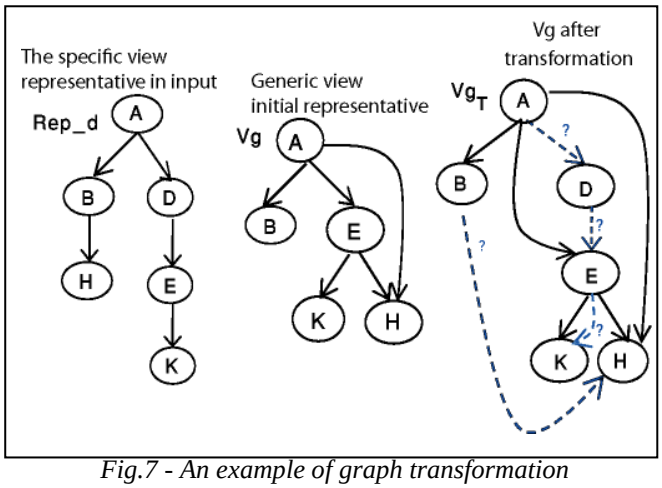

Fig.7 - An example of graph transformation

Two questions can be posed at this level: the first is how to enrich the generic views by adding nodes and arcs: (1) without losing information and (2) without disturbing the cluster whose representative has undergone a transformation? The second question is what is the impact of transformation on the quality of the clusters?

The insertion of nodes and arcs should not cause loss of information. For example, in the graph *Vg* of "Fig.7", when inserted the node *D*, we kept the relationship (*A, E*) and therefore kept the paths *A/E/K* and *A/E/H* of *Vg*. At the same time, we added the path *A/D/E/K*, in the same view Vg_T, because it must represent the graph *Rep_d*. More specifically, in this example, the aim of transformation is to obtain a structure capable of representing both of *Rep_d* and *Vg*.

The insertion of nodes and arcs must not disturb the cluster whose representative has undergone a transformation. Vg_T view should be similar to *Vg* (view before transformation) and must represent all documents already represented by *Vg* but at the same time, it must represent *Rep_d*. For this, we have proposed to use the notion of optional arc marked "? "As additional information (cardinality) of the arc (*A, D*) means that the arc is optional (denoted by (A, D)?).

Convention :

When comparing two paths, arcs (and nodes) that don't exist in one of the paths will not be considered.

In the example of "Fig.7", we show that the graph *Vg* is isomorphic to Vg_T:

Let *Rep_d* = (V_1,E_1), *Vg* = (V_2,E_2) et Vg_T = (V_3,E_3)
V_1 = {A,B,H,D,E,K}, E_1 = {(A,B),(B,H),(A,D),(D,E),(E,K)}
V_2 = {A,B,E,K,H}, E_2 = {(A,B),(A,E),(E,K),(E,H),(A,H)}
V_3={A,B,D?,E,K,H}
E_3={(A,B),(B,H)?,(A,E),(A,D)?,(D,E)?,(E,K),(E,H),(A,H)}
{*A/B/H, A/D/E/K*} the set of the terminal paths of *Rep_d*.
{*A/B,A/E/K,A/E/H,A/H* } the set of the terminal paths of *Vg*.
{*A/B/?H,A/E/K,A/E/H,A/?D/?E/?K,A/H*} the set of the terminal paths of Vg_T.

with *X/?Y* : means that the arc (*X,Y*) is optional.
We have $V_2 \subseteq V_3$ and $E_2 \subseteq E_3$ therefore *Vg* is a sub-graph of Vg_T ($Vg \subseteq Vg_T$).

As far as that goes, the graph Vg_T is a sub-graph of *Vg*:
In fact :

$V_3 \subseteq V_2$ (*D* is optional) and $E_3 \subseteq E_2$ ((*B,H*),(*A,D*) et (*D,E*) are optional)
Therefore Vg_T is a sub-graph of *Vg*, so the graphs *Vg* and Vg_T are similar.

As far as that goes, the graph *Rep_d* is a sub-graph of Vg_T.
In fact:
$V_1 \subseteq V_3$ and $E_1 \subseteq E_3$, so *Rep_d* is a sub-graph of Vg_T (*Rep_d* $\subseteq Vg_T$).

The transformation of a generic view *Vg* aims to enrich it by adding nodes and arcs of *Rep_d* that do not exist in *Vg*. This allows increasing the representativeness of *Vg*. After processing, the resulting graph Vg_T can represent both *Rep_d* and *Vg*.

Our contribution in this step, compared to the work of [13], is that adding nodes doesn't cause a loss of information (arcs).

https://dx.doi.org/10.24001/ijaems.3.6.11 *ISSN: 2454-1311*

The following figure shows an example of adding nodes (in the tree *T*: initial tree) using the approach of [13]. Adding node "Language" has resulted in the loss of the relationship (Speaker, Trans), relevant information in a process of comparing structures.

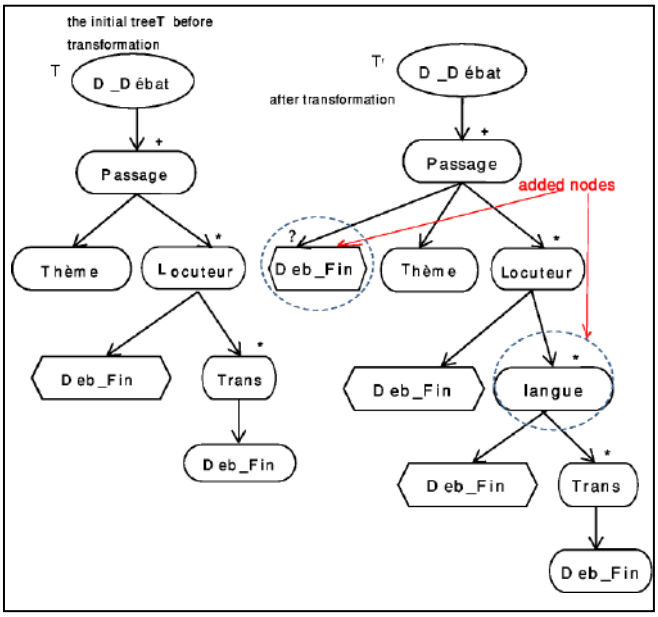

Fig.8 - Example of adding nodes according to the [13] approach

a) Transformation impact on the quality of clusters

The cluster quality depends on the coherence (homogeneity) of the individuals who compose it. This coherence is measured by the intra-cluster distance. The shorter the distance is between the individuals of the same class, the more homogeneous the cluster is.

After the classification of the set E of elements into a set $C = \{c_1, c_2, ..., c_m\}$ of m clusters, the generated clusters should check:

(1) $\forall\ c_i \in C$; $c_i \neq \varnothing$; one cluster represents at least the specific view of the document that generated this cluster.

(2) $\forall (c_i,c_j) \in C^2$; $i \neq j \Rightarrow c_i \cap c_j = \varnothing$; clusters are disjoint (separation)

(3) $E = \bigcup_{i=1}^{m} c_i$; The union of the classes is the set E (initial set)

The transformation problem of classes is related to the question: when to stop the transformation of a class representative?

The transformation can lead to a problem of rapprochement of clusters (decrease inter-cluster distance) and therefore disrupt the classification. Indeed, when the classes are very similar there may be ambiguity: one or several documents belonging to two different classes. This leads to a classification in which clusters are heterogeneous: inter-class distance decreases, however, the intra-class distance increases. When two classes are very similar so there isn't more interest in keeping its: they must be merged into a single class.

The clusters separation is one of the criteria to qualify a classifier. In [14], a wider separation of the clusters implies a better discriminatory power. In [1], two distant objects represent data belonging to different groups.

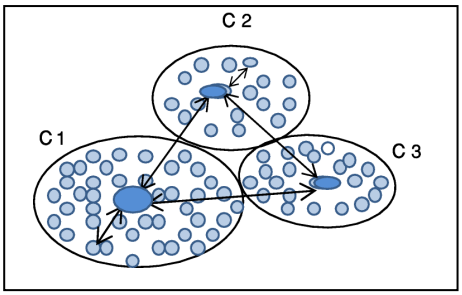

Fig.9 - Illustration of distance intra and inter-cluster

To maintain the stability of clusters and maintain their quality, we proposed to fix a priori a minimum inter-cluster distance (using an inter-cluster threshold). Increasing the separation between clusters can reduce the noise and increase the precision of classification. The use of this parameter is a solution to the problem of rapprochement of classes.

Our contribution at this level, compared to the work of [13] and [3], is taking into account the separation of clusters (a minimum inter-class distance). This allows verifying the inter-cluster distance (separation of classes, "Fig.9") of generic views before and after the transformation. Failure to ignore this parameter can result in continuing to transform (evolve) clusters constantly. At some point, one or more clusters may dominate (absorb) the other clusters. This can cause a disruption of clusters.

a) Transformation cost

Concerning the transformation cost of a graph into another, we proposed this sum of the costs of basic operations (addition of the fragments operations):

$$\sum_{i=1} coût_i \quad where\ coût_i\ is\ the\ cost\ of\ the\ add\ operation\ of\ the\ arc\ (u_i, v_i)$$

where $coût_i = \frac{\alpha_i}{k^{prof(v_i)}}$; α_i and k are two parameters (formula [6]) that reflect the hierarchical and contextual aspects of the structural elements of graphs compared.

In the example in "Fig.8" (k = 10) the cost of transformation *Vg* into *VgT* is:

https://dx.doi.org/10.24001/ijaems.3.6.11 *ISSN: 2454-1311*

$$coût_1 + coût_2 + coût_3 = \frac{1}{100} + \frac{2}{10} + \frac{1}{100} = 0.22$$

$coût_1$: *cost of the add operation of the arc* (*B,H*) (α_1 = *ordre(H)= 1 et prof(H)=2*),

$coût_2$: *cost of the add operation of the arc* (*A,D*) (α_2 = *ordre(D)= 2 et prof(D)=1*),
$coût_3$: *cost of the add operation of the arc* (*D,E*) (α_3 = *ordre(E)= 1 et prof(E)=2*).

A noted class C_i represented by Vg_i, representing *n* specific views can be formally defined as follows:
$C_i = \{Vsp_k / k \in [1,n] ; Sim(Rep\_d_k, Vg_i) >= Seuil\_Sim\}$

Where Vsp_k is a specific view attached to the generic view Vg_i ("Fig.4"), Rep_d_k is the representative of the specific view Vsp_k and *Sim* is the function of structural similarity that we defined in Section 3.

At the end of the transformation process, both the set of transformed generic views (and verifying the separation condition) and the cost of transformations of each of these views are retained for the final step (decision).

When all candidate generic views have been transformed, the system proceeds to the final step to make the final decision: determine, among the generic views of the documentary warehouse, the generic view most similar to the specific view of the document to be integrated.

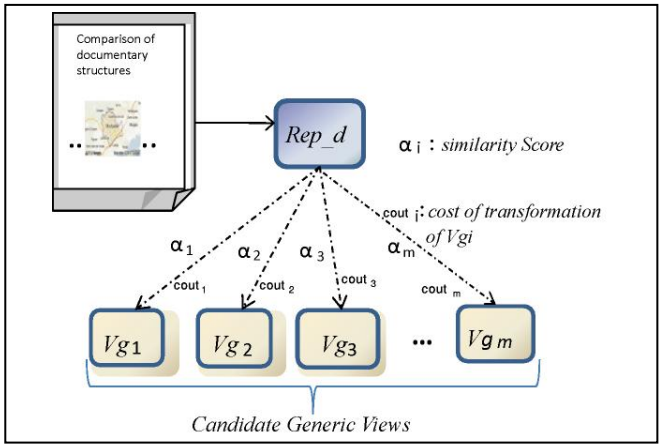

Fig.10 - Calculation of similarity scores between the document representative and the existing generic views

This step consists in extracting from the set of generic views the one whose degree of similarity with the specific view of the document to be integrated is the highest, then comparing this degree with the similarity threshold (*Seuil_Sim* a parameter set by experimentation). There are two cases:

(1) If this degree of similarity is strictly less than *Seuil_Sim*, a new class will be created from the representative *Rep_d* of the specific view of the document to be integrated,

(2) if this degree of similarity is greater than or equal to *Seuil_Sim* then two cases can be envisaged:

- a single generic view is similar to the specific view of the document to be integrated. In this case the specific view is attached to it.
- several generic views are similar to the specific view of the document to be integrated. In this case we choose the one for which the integration of this new specific view will require the least transformations (least cost).

VI. EXPERIMENTAL RESULTS

The used corpus

In our experiments, we study the similarity threshold impact on the quality of the classes generated by our classification process. For this, we conducted two sets of tests on the same corpus composed of 1278 documents extracted randomly from the *INEX 2007* corpus and a corpus composed of descriptive records of books in XML format from the library of University of Toulouse 1 Capitol (table 1).

Number of documents	**1278**
Total number of nodes	30236
Total number of elements	17427
Total number of attributes	12809
Average number of nodes/*Vsp*	23.66
Average number of paths /*Vsp*	8.72

Table 1 : Description of the used corpus

In both sets of tests, we fix the *filtering threshold* to 64% and we vary the *similarity threshold* to 78% and 82%. Tables of each of our experiences will show the following measures: *Nb_ Nodes:* the number of specific views related, *Nb_Paths*: the number of paths and Std_*Dev*: standard deviation intra-cluster.

With a *similarity threshold* of 78% (*classif78*), the 1278 document are grouped into 36 clusters:

Clusters	*Nb_Vsp*	*Nb_Nodes*	*Nb_Pahs*	*Std_Dev*
$C_{1\text{-}78}$	34	729	310	0.02
$C_{2\text{-}78}$	186	4222	1770	0.01
$C_{3\text{-}78}$	21	471	193	0.02
$C_{4\text{-}78}$	30	621	246	0.06
$C_{5\text{-}78}$	20	367	135	0.01
$C_{6\text{-}78}$	22	436	218	0.03
$C_{7\text{-}78}$	23	748	98	0.01
$C_{8\text{-}78}$	85	1514	607	0.01
$C_{9\text{-}78}$	40	940	364	0.03
$C_{10\text{-}78}$	105	2479	583	0.02
$C_{11\text{-}78}$	67	1056	419	0.01
$C_{12\text{-}78}$	13	319	121	0.02
$C_{13\text{-}78}$	42	1084	518	0.03
$C_{14\text{-}78}$	56	1244	395	0.02
$C_{15\text{-}78}$	30	654	251	0.03
$C_{16\text{-}78}$	18	467	181	0.02
$C_{17\text{-}78}$	6	143	70	0.01

https://dx.doi.org/10.24001/ijaems.3.6.11 *ISSN: 2454-1311*

$C_{18\text{-}78}$	33	810	327	0.02
$C_{19\text{-}78}$	29	523	194	0.02
$C_{20\text{-}78}$	26	478	174	0.01
$C_{21\text{-}78}$	18	425	186	0.01
$C_{22\text{-}78}$	34	827	248	0.01
$C_{23\text{-}78}$	30	539	216	0.02
$C_{24\text{-}78}$	29	529	189	0.01
$C_{25\text{-}78}$	7	133	63	0.01
$C_{26\text{-}78}$	13	281	86	0.00
$C_{27\text{-}78}$	22	474	156	0.01
$C_{28\text{-}78}$	8	191	53	0.01
$C_{29\text{-}78}$	29	645	220	0.00
$C_{30\text{-}78}$	42	1028	305	0.00
$C_{31\text{-}78}$	72	1399	508	0.02
$C_{32\text{-}78}$	10	232	124	0.01
$C_{33\text{-}78}$	12	255	130	0.02
$C_{34\text{-}78}$	44	1257	242	0.01
$C_{35\text{-}78}$	17	2170	991	0.01
$C_{36\text{-}78}$	5	546	288	0.02

Table 2: Classification (classif78) results

With a *similarity threshold* of 82% (*classif82*), the 1278 document are grouped into 39 clusters:

Clusters	Nb_*Vsp*	Nb_Nodes	*Nb_Paths*	*Std_Dev*
$C_{1\text{-}82}$	32	678	293	0.01
$C_{2\text{-}82}$	185	4198	1770	0.01
$C_{3\text{-}82}$	21	471	193	0.02
$C_{4\text{-}82}$	15	344	130	0.01
$C_{5\text{-}82}$	20	367	135	0.01
$C_{6\text{-}82}$	12	281	147	0.01
$C_{7\text{-}82}$	23	748	98	0.01
$C_{8\text{-}82}$	85	1514	607	0.01
$C_{9\text{-}82}$	40	940	364	0.03
$C_{10\text{-}82}$	105	2479	583	0.02
$C_{11\text{-}82}$	67	1056	419	0.01
$C_{12\text{-}82}$	13	319	121	0.02
$C_{13\text{-}82}$	41	1061	507	0.02
$C_{14\text{-}82}$	56	1244	395	0.02
$C_{15\text{-}82}$	16	361	124	0.01
$C_{16\text{-}82}$	17	445	170	0.01
$C_{17\text{-}82}$	6	143	70	0.01
$C_{18\text{-}82}$	33	810	327	0.02
$C_{19\text{-}82}$	29	523	194	0.02
$C_{20\text{-}82}$	26	478	174	0.01
$C_{21\text{-}82}$	18	425	186	0.01
$C_{22\text{-}82}$	34	827	248	0.01
$C_{23\text{-}82}$	30	539	216	0.02
$C_{24\text{-}82}$	29	529	189	0.01
$C_{25\text{-}82}$	7	133	63	0.01
$C_{26\text{-}82}$	13	281	86	0.00
$C_{27\text{-}82}$	22	474	156	0.01
$C_{28\text{-}82}$	8	191	53	0.01
$C_{29\text{-}82}$	29	645	220	0.00
$C_{30\text{-}82}$	42	1028	305	0.00
$C_{31\text{-}82}$	72	1399	508	0.02
$C_{32\text{-}82}$	10	232	124	0.01
$C_{33\text{-}82}$	12	255	130	0.02
$C_{34\text{-}82}$	44	1257	242	0.01
$C_{35\text{-}82}$	17	2170	991	0.01
$C_{36\text{-}82}$	5	546	288	0.02
$C_{37\text{-}82}$	5	78	46	0.02
$C_{38\text{-}82}$	9	194	89	0.01
$C_{39\text{-}82}$	30	531	189	0.02

Table 3: Classification (classif82) results

We gave each of the clusters its equivalent of clusters of *classif78*. After examining the results of the two classifications *classif78* and *classif82*, we have noted the emergence of three new clusters (Table 3) $C_{37\text{-}82}$, $C_{38\text{-}82}$, and $C_{39\text{-}82}$.

In comparison with the results of *classif78* (Table 2), we notice a considerable optimization of the standard deviation of intra-cluster that have changed: lines 1, 4, 6, 13, 15 and 16 of Table 3.

We have observed that increasing the value of threshold similarity imply the homogeneity of clusters. On the other hand, this allows the creation of an excess of classes. However, the decrease in this value reduces the number of classes (Table 2). More specifically, when the threshold similarity value decreases, the number of specific views attached to each cluster increases, which leads to a heterogeneity between individuals of the same cluster. We must find a compromise between the number of generated classes and intra-cluster homogeneity.

VII. CONCLUSION AND OUTLOOKS

Our classification approach is based on a measure of structural similarity that we have proposed. A measure based on matching graphs. It is based on a weighting function that reflects the hierarchical and contextual aspects of the components of graphs. It is parameterized by a threshold similarity to define a priori the degree of similarity between the representative of each generated cluster and individuals in this cluster. This ensures an intra-cluster coherence.

As we have evoked the clusters (generic views) can undergo transformations as one goes along the classification and that can cause the problem of approximation of clusters. During this series of tests that we have conducted, we noted that such a phenomenon is not produced. This is because we have fixed a priori for each test a *minimum inter-cluster*. The separation of clusters is one of the criteria to qualify a classifier. Taking into account the separation of clusters allows keeping the discriminating power of these classes to avoid overlapping and rapprochement. We have noted during the experiments that we have conducted that the problem of rapprochement of clusters doesn't arise.

When clusters are sufficiently separated, the problem of the belonging a specific view to more clusters will not be

https://dx.doi.org/10.24001/ijaems.3.6.11 *ISSN: 2454-1311*

envisaged. A problem we have not faced in the series of tests that we have conducted. Increasing the inter-cluster distance reduces noise and increases the precision of classification.

Our future works will be devoted to the study of the combination of the filter threshold and the similarity threshold.

REFERENCES

[1] Bisson G., « La similarité: une notion symbolique/ numérique. Apprentissage symbolique-numérique ». Eds Moulet, Brito, Cepadues Edition, 2000

[2] Dalamagas T., Cheng T., Winkel K-J, Sellis T.K., "A methodology for clustering XML documents by structure". Information Systems 31(3), 2006: pp.187-228

[3] Djemal Karim, « De la modélisation à l'exploitation des documents à structures multiples », Thèse de Doctorat de l'Université de Paul Sabatier. - Toulouse, France, 2010.

[4] Gentner D., "Structure-mapping: A theoretical framework for analogy", *Cognitive Science*, 7, 155-170. (Reprinted in A. Collins & E. E. Smith (Eds.), *Readings in cognitive science: A perspective from psychology and artificial intelligence*. Palo Alto, CA: Kaufmann), 1983.

[5] Gentner D. "*The mechanisms of analogical learning"*. *In* S. Vosniadou et A. Ortony (dir.), *Similarity and analogical reasoning*, p. 199-241. Cambridge: Cambridge University Press. 1989.

[6] Hummel, J.E. "*Where view-based theories break down: The role of structure in shape perception and object recognition",* In E. Dietrich and A. Markman (Eds.). Cognitive Dynamics: Conceptual Change in Humans and Machines. Hillsdale, NJ: Erlbaum, 2000.

[7] Hummel, J.E. "*Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition" Visual Cognition,* 8, p. 489-517, 2001.

[8] Ali Idarrou, Driss Mammass, Chantal Soulé-Dupuy, Nathalie Vallés-Parlangeau. "A generic Approach to the Classification of Multimedia Documents: a Structures Comparison". Dans : ICGST International Journal on Graphics, Vision and Image Processing, The International Congress for global Science and Technology, Vol. Volume 10 N. Issue VI, p. 13-18, december 2010.

[9] Ali Idarrou, Chantal Soulé-Dupuy, Nathalie Vallés-Parlangeau. *Classification structurelle des documents multimédias basée sur l'appariement des graphes (regular paper).* Dans : *INFormatique des Organisations et Systemes d'Information et de Decision (INFORSID 2012), Montpellier (France), 29/05/2012-31/05/2012*, Association INFORSID, p. 539-554, 2012.

[10] Ali Idarrou and Driss Mammass, "*Structural Clustering Multimedia Documents*: *An Approach based on Semantic Sub-graph Isomorphism*". *International Journal of Computer Applications* 51(1):14-21, August 2012. Published by Foundation of Computer Science, USA,Vol. 51 N. 1, August 2012. Accès: http://www.ijcaonline.org/archives/volume51/number1/8005-1343.

[11] Ali Idarrou and Driss Mammass, "A New Structural Similarity Measure for Clustering Multi--Structured Documents", In Journal of Theoretical and Applied Information Technology, 10th April 2016. Vol.86. N°.1 p 34-43, ISSN: 1992-8645 E-ISSN: 1817-3195. http://www.jatit.org/volumes/Vol86No1/5Vol86No1.pdf

[12] Sami Laroum, Nicolas Béchet, Hatem Hamza et Mathieu Roche, "Classification automatique de documents bruités à faible contenu textuel", Manuscrit auteur, publié dans RNTI : Revue des Nouvelles Technologies de l'Information 1, 2009.

[13] Mbarki M., Gestion de l'hétérogénéité documentaire : le cas d'un entrepôt de documents multimédias., Thèse de Doctorat de l'Université de Paul Sabatier, Toulouse 3 France, 2008.

[14] Oh,Il-Seok, Lee J., Suen C., Analysis of Class Separation and Combination of Class Dependent Features for Handwriting Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, No. 10, October 1999, 1089-1094.

[15] Pierre-Edouard PORTIER « Construction des Documents Multistructurés dans le Contexte des Humanités Numériques », Thèse de Doctorat de l'INSA De Lyon France, 2010.

[16] Vercoustre, A. M., Fegas, M., Lechevallier, Y., Despeyroux, T., et Rocquencourt, I. (2006). « Classification de documents XML à partir d'une représentation linéaire des arbres de ces documents ». Paris, France, 433–444.

[17] Sorlin S. « Mesurer la similarité de graphes », Thèse de Doctortat de l'Université de Claude Bernard Lyon I France , 2006.