Journal of Theoretical and Applied Information Technology
10th April 2016. Vol.86. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A NEW STRUCTURAL SIMILARITY MEASURE FOR CLUSTERING MULTI-STRUCTURED DOCUMENTS

[1]ALI IDARROU, [1]DRISS MAMMASS
[1]IRF-SIC, IBN ZOHR University Agadir, Morocco
E-mail: ali.idarrou@uiz.ac.ma

ABSTRACT

Traditional methods of document comparison are based on the similarities called "*surfaces*": a model of similarity based on descriptive properties of objects without considering the relationships between these properties. We have proposed a *new structural measure*, based on sub-graph isomorphism, taking into account the distribution (order, position, etc) of components of the documents compared and the relationships between these components (preserve more sense). Our measure reflects both the contextual and structural aspects of documents compared. In this work, we will show in detail our similarity measure and study the impact of the *similarity threshold* (a parameter fixed previously) on generated clusters. We evaluate our approach on a corpus of multimedia documents extracted randomly from the INEX 2007 corpus and the corpus of descriptive records of books in XML format from the library of the University of Toulouse 1 Capitole.

Keywords: *Multimedia Document, Clustering, Sub-Graph Isomorphism, Structural Similarity*

1. INTRODUCTION

Automatic classification is a solution that allows organizing a large collection of documents. This allows reducing the search space and thus improves the performance of the information access processing in a large mass of data: increasing the accuracy and reducing noise.

Documents can be grouped together according to their structure and/or their content. We consider that the document structure is a sufficiently discriminating factor for classification. Thus, the structural classification in the sense that we understand [7] allows creating, in a documentary warehouse, clusters called generic views. A generic view is a superposition of trees representing document structures. It can be enriched (addition of fragments: transformation of the generic views) along with the classification. This tree superposition creates a rooted graph structure (example Figure 1). It is not a simple summary, as is the case of the works using the summary trees to represent documents, but rather a rich description (without losing information) representing a set of specific structures structurally similar.

Comparing two documents requires modeling these documents in a formal manner and using (or defining) an appropriate measure to evaluate the similarity between these documents. We are interested in representing multi-structured multimedia documents using graphs. To compare two documents is, therefore, to compare the graphs that represent them. The graph theory could be of great interest in the evaluation of the structural similarity. The induced sub-graph isomorphism allows showing that a graph is included in another, one, while the partial sub-graph isomorphism determines the intersection of the two graphs.

To evaluate the proximity between two graphs, we have proposed a new structural similarity measure based on sub-graph isomorphism that relies on a graph weighting function that we have introduced. The latter allows expressing hierarchical and contextual aspects of components (nodes and arcs), insofar as it takes into account the distribution of these components in the graph and the nature of the relationships between these components. Our graph weighting function allows reflecting both the structure and sense of the compared documents.

In our previous works [9], we have made a comparative study with manual classification, on the one hand, and with the approach of [11] on the other hand. We have also studied the impact of the filtering sub-process, of our clustering process, on the quality of the generated clusters. The aim of the filtering process is to optimize the space of graph comparison in order to improve our clustering process performance. In this paper, we will show in detail our similarity measure and we will study the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

impact of the similarity threshold (a parameter fixed a priori) of the resulting clusters.

In the next section we will give an overview, not exhaustive but representative, of the works which have used trees or graphs to represent documents. First we will begin this section with some basic notions on graphs. In the third section we will present the *MVDM* model "*Multi Views Document Model*" [4]. We will describe in the fourth section our similarity measure. In the fifth section, we will give a brief overview of our structural clustering process of multi-structured multimedia documents. Before concluding, we will present in the sixth section our experimental results.

2. RELATED WORKS

2.1 Basic notions on graphs

Let $G=(V,E)$ and $G'=(V',E')$ two graphs are defined by its set of nodes V (resp. V') and its set of edges E (resp. E').

Definition 1

G is a sub-graph of $G' \Leftrightarrow V \subset V'$ and $E \subset E'$.

Definition 2

G is isomorphic to a sub-graph of G' if and only if there is an injection f from V to V' such:

$\forall(u,v)\in V^2; (u,v)\in E \Rightarrow (f(u),f(v))\in E'$

2.2 Document representation

Several works have used trees to represent the documents to compare. In their approach of structural classification of documents, the authors of [3] use the tree summary obtained by transformations of trees (depth reduction, elimination of repeated nodes, etc). However, these transformations can cause a loss of semantic and contextual information. For example, the depth reduction involves the elimination of components and relations between these components.

The works of [2,10,13,15,17] have used the frequent sub-trees (sub-trees that appear frequently in the collections of trees considered) to classify documents. In their approach to semantic classification of XML documents [14] have proposed a model of data representation that exploits the notion of tree-tuple to identify the semantically coherent sub-structures in XML documents. 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. Thus, the classification is based on the calculation of the frequency of these paths. The idea of linearization of trees proposed in these works is very interesting. In the approach of [11], the semantic and logical structures of XML documents to be classified are represented as tree forms. [11] has proposed a measure to evaluate the degree of inclusion between two trees T and T' :

$$\mathrm{Sim}(T,T') = 1 - \frac{\sum_{v_j} \mathrm{Danc}(v_j)}{\sum_{v_j} \mathrm{Panc}(v_j)} \qquad (1)$$

- $Danc(v_j)$: represents the alignment distance of ancestors of node v_j.
- $Panc(v_j)$: represents the weight of the ancestors of node v_j.

Other works have used graphs to represent documents. In fact, graphs are data structures having the capacity to represent complex and structured objects. The mathematical theory of graphs could be of great interest to the evaluation of the similarity of documents, both in retrieval information and the documentary classification. In [18], the sub-graph isomorphism can be used to show the inclusion or equivalence of two graphs. In the works of [6], the graphs were used to represent images segmented to classify them. In [1], graphs have been used to represent objects for computer assisted design. The nodes of the graph represent the components of the object and the edges of the graph represent binary relations between these components. In [4] the graphs were used to represent the multi-structured documents within a documentary warehouse. He has proposed the *MVDM* model to describe the multi-structured documents.

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

3. PRESENTATION OF THE MODEL *MVDM*

The *MVDM* model introduced the concept of view: a set of structural nodes and relations between these nodes. A node can be simple or complex (for example, a fragment multimedia image). In this last case, the node can be considered as a sub-document itself can be fragmented into a set of nodes and relations between these nodes. There may be more than one possible relationship between two same components of a document. A document view allows materializing several organizations of this document. According to this model, the notion of document structure can be encompassed within a wider notion which is the view. A specific view corresponds to a particular organization of a document or a viewpoint on this document. It reflects one of the structures of a multi-structured document [4]. The *MVDM* model is composed of two levels: (1) a specific level (*DWsp*) where each

Journal of Theoretical and Applied Information Technology
10th April 2016. Vol.86. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

specific view, characterizing the organization of a particular document, is represented in tree form and (2) a generic level (*DWg*) where each generic view (cluster) represents a collection of the specific views structurally similar. The generic views are represented in graph forms (Figure 1).

We can write: $DW = DWg \cup DWsp$ (Figure 1) where DWg represents the generic level (clusters) of *DW* and *DWsp* represents the specific level of *DW*: the specific characteristics of each document (structure+content). Access to the cluster representative (Vg_i) allows targeted access to a sub-collection of documents of *DWsp* represented by it (by Vg_i).

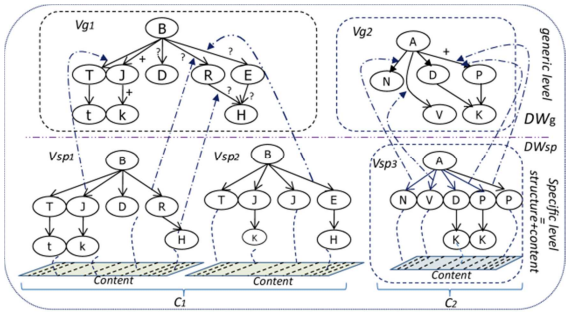

Figure.1: Architecture of the documentary warehouse

Formally, we can define as follows:

- $DW_g = \cup_{i=1}\{Vgi\}$ is the set of generic views, where each generic view Vg_i consists of a set of generic paths. We can write : $Vg_i = \cup_{j=1}\{path_j\}$ where each generic path $path_j$ is a set of generic relationships: $path_j = \cup_{r=1}\{e_r\}$;
- $DW_{sp} = \cup_{k=1}\{Vsp_k\}$ is a set of specific views, where each specific view Vsp_k is composed of a set of specific paths. We can write: $Vsp_k = \cup_{p=1}\{path_p\}$ where each specific path $path_c$ is a set of specific relationships: $path_c = \cup_{s=1}\{e_s\}$.

In the next section, we introduce a new structural similarity measure.

4. DEFINITION OF A STRUCTURAL SIMILARITY MEASURE

4.1 Weighting of a Graph

In a multimedia document, the relationships between the structural elements are additional information that can't be neglected. For example, in a television newscast, the audio and video must be synchronized (temporal relationships, special relationships, etc) to ensure consistency of the information.

In a process of structural comparison of documents, we think that the structural information is essential and that two documents composed of the same components doesn't imply they are similar. For example, the same image in two different documents may not express the same context.

According to the mapping theory developed by [5], good analogies are those based on relationships between entities rather than their descriptive properties. In that vein, we have defined a weighting model of graph on which will be based our similarity measure. According to this measure the weight of an arc must reflect the importance of a structural viewpoint of this arc in the graph. It must therefore take into account the relationships between different components of a graph and the position of each of these components; position in a path and order relative to the brother components.

We have chosen to consider a graph as a set of paths. This allows reducing the cost generated by combinatorial search of graph isomorphism; known problem in graph theory [8]. The comparison of two graphs is therefore the comparison of the paths that compose them.

Let $G = (V,E)$ a directed, labeled and ordered graph. The *weighting function* P_e of a given arc is defined by:

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

where

$$P_e(u,v) = \begin{cases} 1 - \frac{\alpha}{k} & \text{if } prof(v)=1 \\ P_e(x,u) - \frac{\alpha}{k^{prof(v)}} & \text{otherwise; } x \in father(u) \end{cases} \quad (2)$$

- $x \in father(u)$: u can have multiple parent nodes (Figure 2, in graph G, $father(H) = \{C,A\}$);
- *prof(v):* profoundness of v: position in a path;
- k *(*a power of *10)* a fixed parameter indicating the maximum number of son nodes (number of son nodes< k) for each node of the manipulated graphs depending on the nature of the document collection treated (profoundness of root node = 0*);*
- α is a parameter that depends on the type of node v*:*

$$\alpha = \begin{cases} 1 \text{ if } v \text{ an attribute or metadata} \\ order(v) \text{ otherwise} \end{cases} \quad (3)$$

In the formula (2), the number of digits of the fractional part of $P_e(u,v)$, which depends on k, indicates the profoundness of the arc (u,v) *extremity.*

Journal of Theoretical and Applied Information Technology
10th April 2016. Vol.86. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

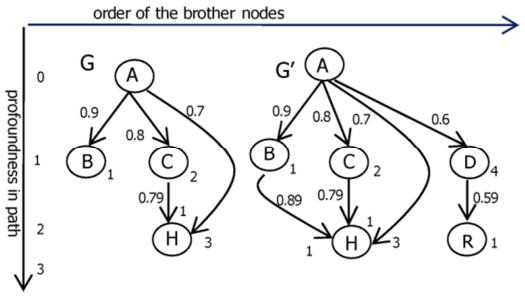

Figure 2: Example of the graph weighting

In the example (Figure 2), the graph G is composed of paths: "*A/B*", "*A/C/H*" *and* "*A/H*". We have:

$P_e(A,B) = 1 - \frac{\alpha}{10} = 1 - \frac{1}{10} = 0.9$,

$P_e(C,H) = P_e(A,C) - \frac{\alpha}{100} = 0.8 - \frac{order(H)}{100} = 0.8 - \frac{1}{100}$ $= 0.79$.

To classify documents, it is necessary to have an appropriate operator to evaluate the proximity between two documents. When the documents are represented in graph forms, to compare two documents structurally is therefore to compare the graphs that represent them. In graph theory, the problem of comparing graphs is reduced to the problem of looking for an isomorphism of (sub) graphs. The isomorphism of (sub) graphs allows showing that two graphs are structurally identical or one is included in the other. We situate our works in the framework of looking for an isomorphism of sub-graphs and we propose a new measure of structural similarity.

4.2 Sub-graph Isomorphism

We consider a graph as a set of paths and a path of a graph is a sub-graph of this graph. In the example of the Figure 2, $G = (V,E)$ where V is a set of nodes of G and E is a set of its arcs. The path denoted $path_2$ ="*A/C/H* " is a sub-graph of G. in fact, we can write $path_2 = (V_2,E_2)$ with $V_2 = \{A,C,H\}$and $E_2 = \{(A,C), (C,H)\}$. We have $V_2 \subset V$ and $E_2 \subset E$ therefore $path_2$ is a sub-graph of G.

Before defining our structural similarity measure, we first define the measure d_{Inc} which evaluates the inclusion degree of a given path in a given graph G':

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

where

$$w_{j,k} = \begin{cases} P_e(e'_h) \ \ if \ \exists \ e'_h \in path'_k / \ \varphi_e(e_j) = e'_h \ (path'_k \subset G') \\ 0 \ \ otherewise \end{cases} \quad (5)$$

- n': the number of G' paths;
- φ_e: an alignment function bidirectional from E (resp. E') to E' (resp. E) which allows aligning two similar arcs:

  $\varphi_e : E \rightarrow E'$

  $a \mapsto \varphi_e(a) = a'$; where the arcs a and a' are structurally similar.

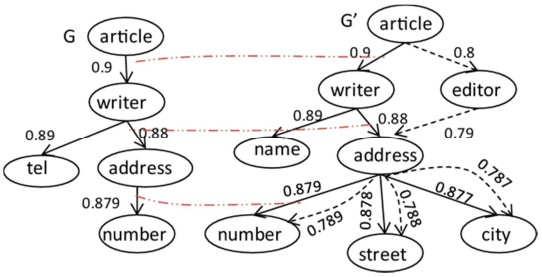

Figure 3: The inclusion of a path in a graph

In this example (Figure 3), the graph G' is composed of 7 paths:

- $path'_1$ = "*article/writer/name*";
- $path'_2$ = "*article/writer/address/number*";
- $path'_3$ = "*article/writer/address/street*";
- $path'_4$ = "*article/writer/address/city*";
- $path'_5$ = "*article/editor/address/number*";
- $path'_6$ = "*article/editor/address/street*";
- $path'_7$ = "*article/editor/address/city*".

The graph G is composed of two paths:

- $path_1$ = "*article/writer/tel*";
- $path_2$ = "*article/writer/address/number*".

Let G and G' two given graphs and *path* a path of G.

Theorem 1

$d_{Inc}(path,G') = 0$ if and only if *path* is isomorphic to a sub-graph of G' (*path* is structurally similar to a path of G': $path \subset G'$).

In fact:

$$d_{Inc}(path, G') = \sum_{e_j \in path} |P_e(e_j) - w_{j,k}| = 0$$

$\Leftrightarrow \forall \ e_j \in path; \exists \ k \in [1,n'] \ ; P_e(e_j) = w_{j,k}$

$\Leftrightarrow \forall e_j \in path; \exists \ e'_h \in path'_k / \varphi_e(e_j) = e'_h$

$\Leftrightarrow path \subseteq path'_k \ \ (path'_k \subseteq G')$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In Figure 3, $path_2$=" *article/writer/address/number*" of G and $path'_2$="*article/writer/address/number*" of G' are isomorphic.

In fact:

$$d_{Inc}(path_2, G') = min_{k \in [1,7]} \left[\frac{\Sigma_{e_j \in path_2} |P_e(e_j) - w_{j,k}|}{\Sigma_{e_j \in path_2} P_e(e_j)} \right] = 0$$

Theorem 2

$d_{Inc}(path, G') = 1$ if and only if $path \cap G' = \varnothing$

In fact:

$$d_{Inc}(path, G') = 1$$

$$\Leftrightarrow min_{k \in [1,n']} \left[\frac{\Sigma_{e_j \in path} |P_e(e_j) - w_{j,k}|}{\Sigma_{e_j \in path} P_e(e_j)} \right] = 1$$

$$\Leftrightarrow \forall e_j \in path \; ; \forall \; k \in [1,n'], \; w_{j,k} = 0$$

$$\Leftrightarrow \; path \cap G' = \varnothing$$

Theorem 3

Let G be a graph composed of n paths $\{path_1, path_2, ..., path_n\}$.

$\sum_{i \in [1,n]} d_{Inc}(path_i, G') = 0 \Leftrightarrow G$ *isomorphic to a sub-graph of* G'

In fact:

$$\sum_{i \in [1,n]} d_{Inc}(path_i, G') = 0$$

$\Leftrightarrow \forall i \in [1,n] \; ; \; d_{Inc}(path_i, G') = 0$

$\Leftrightarrow \forall i \in [1,n] \; ;$ $path_i$ isomorphic to a sub-graph of G'(theorem 1)

$\Leftrightarrow G$ isomorphic to a sub-graph of G'.

In this example (Figure 2), G is composed of 3 paths: $path_1$ = "*A/B*", $path_2$ = "*A/C/H*", and $path_3$ = "*A/H*".

$$\sum_{i \in [1,3]} d_{Inc}(path_i, G') = d_{Inc}(path_1, G') + d_{Inc}(path_2, G') + d_{Inc}(path_3, G')$$

$$= \left[\frac{(0.9-0.9)}{0.9} + \frac{(0.79-0.79)+(0.8-0.8)}{1.59} + \frac{(0.7-0.7)}{0.7} \right] = 0$$

Therefore the graph G (Figure 2) is isomorphic to a sub-graph of G'.

4.3 A new Structural Similarity Measure

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

We have proposed a new structural similarity measure based on sub-graph isomorphism. This measure reflects the structure of graphs compared in the sense that we compare the paths of graphs taking into account both the position of the nodes, the order of the brother nodes and the relationships between these nodes (example Figure 4). In our context, we consider that the position of nodes and the relationships between these nodes are two essential parameters in a process of structural comparison of multimedia documents.

To evaluate the structural similarity between two graphs G and G' noted $Sim(G,G')$, we have defined the following measure:

$$Sim(G, G') = 1 - Dist(G, G') \quad (6)$$

$$where \quad Dist(G, G') = \frac{(d_{GG'} + d_{G'G})}{2} \quad (7)$$

$$and \quad d_{GG'} = \frac{1}{n} \sum_{i \in [1,n]} d_{Inc}(path_i, G') \quad (8)$$

$$and \quad d_{G'G} = \frac{1}{n'} \sum_{i \in [1,n']} d_{Inc}(path'_i, G) \quad (9)$$

Where $d_{GG'}$ (rep. $d_{G'G}$) is the alignment distance between G and G' (resp. G' and G) and n and n' are respectively the number of paths of G and G'. The division by n (rep. n') allows normalizing the value of $d_{GG'}$ (rep. $d_{G'G}$) between 0 and 1.

Corollary 1

$d_{GG'} = 0 \Leftrightarrow G$ is isomorphic to a sub-graph of G'(theorem 1).

Corollary 2

$d_{GG'} = 0$ and $d_{G'G} = 0 \Leftrightarrow G$ and G' are isomorphic.

The similarity measure proposed is based on path matching of the graphs to compare. We show through the example of Figure 4, that it takes into account the distribution (*profoundness*, *order*, *hierarchy*) of the components (*nodes and arcs*) of the graphs compared.

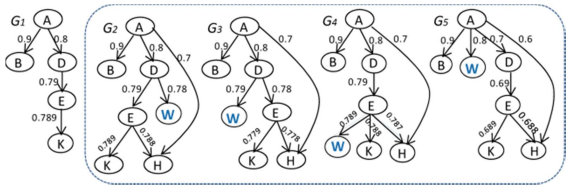

Figure 4: Example of similarity calculation between graphs

In this example, we have:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$Sim(G_1,G_2) = 0.82$, $Sim(G_1,G_3) = 0.81$, $Sim(G_1,G_4) = 0.83$ and $Sim(G_1,G_5) = 0.71$.

The difference between $Sim(G_1,G_2)$, $Sim(G_1,G_3)$, $Sim(G_1,G_4)$ and $Sim(G_1,G_5)$ can be explained by the fact that the proposed measure takes into account the distribution of structural elements in the graphs compared. We observe a difference, which becomes important in the case of $Sim(G_1,G_4)$, between the similarity values due to differences in positioning some nodes, in particular the node "*W*" (*different order or different profoundness*). This shows that the proposed similarity measure takes into account the profoundness and order, penalizing differences of the profoundness.

To compare the proposed measure with the existing measures, we have chosen two types of measures: a type based on the descriptive characteristics (*Jaccard's measure*) regardless of the relationship between the components of the objects compared and a type based on the structural alignment [11].

Table 1: Comparison between our measure, Jaccard's measure and [11]'s measure

Jaccard's measure	[11]'s measure	Our measure
$Sim(G_1,G_2)= 0.63$	$Sim(G_1,G_2)=1$	$Sim(G_1,G_2)=0.82$
$Sim(G_1,G_3)=0.63$	$Sim(G_1,G_3)=1$	$Sim(G_1,G_3)=0.81$
$Sim(G_1,G_4)=0.63$	$Sim(G_1,G_4)=0.99$	$Sim(G_1,G_4)=0.83$
$Sim(G_1,G_5)=0.63$	$Sim(G_1,G_4)=0.93$	$Sim(G_1,G_4)=0.71$

The graphs G_2, G_3, G_4, and G_5 consist of the same nodes. However, these nodes don't have the same distribution on 5 graphs. Specifically, these graphs aren't identical from a structural point of view. We note that the values shown by the lines in column 1 of Table 1 are the same. They don't depend on the organization of nodes of graphs G_1, G_2, G_3, G_4, and G_5. Unlike Jaccard's measure, our measure is structural and not a "*surface measure*", it takes into account the structural aspect of compared objects, which is clearly reflected through the values of the third column of Table 1. Indeed, our measure is based on a *weighting function* (2) taking into account the hierarchical and contextual aspects. Measuring of [11] calculates the degree of inclusion of a given graph in another. It doesn't evaluate the similarity between two graphs. The weighting function proposed by the author favors the son node (level n + 1) on the parent node (level n). We note that according to this measure, the similarity between a graph G and a graph G' which contains it is equals 1 and that whatever G' is (e.g, lines 1 and 2 in column 2 of Table 1. In fact, it is difficult to interpret the result as $Sim\ (G,G') = 1$. Unlike the measure of [11], our measure penalizes the non-matching components (nodes and arcs) of the graph G'. Specifically, our measure allows evaluating the inclusion in either direction ($G \subseteq G'$ *and* $G' \subseteq G$).

5. STRUCTURAL CLUSTERING OF MULTI-STRUCTURED DOCUMENTS

In [12], the model *MVDM* allows a rich representation of the mutli-structured documents and that this wealth can be exploited to classify multi-structured documents. Within the framework of *MVDM*, the problem of classification results in the problem of attachment (example, Figure 1) of a *specific view* of a given document to the generic view the most structurally similar. The choice of the *generic view*, of the documentary warehouse the most structurally similar to which specific view must be attached, is based on the comparison of the latter with all the *generic views* of the documentary warehouse.

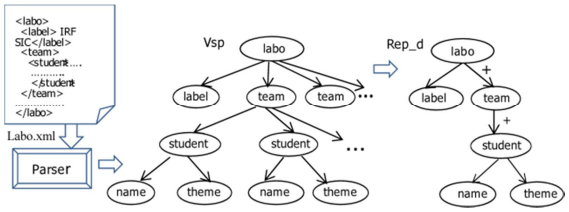

Figure 5: Example of extracting a specific view then the document representative

- *Vsp: is a specific view of the document "labo.xml",*
- *Rep_d*: *a generic representative* of *Vsp*,
- "?": *cardinality; means optional component,*
- "+": *cardinality; means one or more.*

In our previous works [7], we have presented the steps of our document integration process in the documentary warehouse. Due to the lack of space, we couldn't show in detail our approach to structural clustering, but we refer the reader to these works. The basic idea of our integration process of a new multimedia document is to extract the specific view *Vsp* of this document then its representative, which materializes the generic representation of *Vsp*. The representative *Rep_d* (e.g Figure 5) thus obtained is subsequently used in the comparison process. The comparison process consists of calculating the similarity between *Rep_d* and each *generic view Vg* of Dw_g (e.g Figure 1). Then, depending on the results of this step either aggregating *Vsp* in the cluster most similar (attach the specific components nodes and relationships of the document to the generic components structurally similar to *Vg*, example Figure 1) or create a new cluster. Clusters aren't defined previously; they are

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

created automatically along with the integration of documents.

Definition 3

Let G and G' be two directed labeled and ordered graphs. G and G' are structurally similar (according to our measure) if and only if: $Sim(G,G') >= Sim\_thresh$; where Sim_thresh is a similarity threshold (parameter fixed previously).
For example (Figure 2), with $Sim\_thresh = 0.82$ the graphs G_1 and G_4 are similar.

6. EXPERIMENTAL RESULTS

We have studied the impact of the similarity threshold on the quality of classes generated by our clustering process. To do this, we conducted three series of tests on the same corpus of 1606 documents extracted randomly from the INEX 2007 corpus and the corpus of descriptive records of books in XML format from the library of the University of Toulouse 1 Capitole.

Table 2: Description of the used corpus.

Number of documents	**1606**
Total number of nodes	38138
Total number of elements	21814
Total number of attributes	16324
Average number of nodes/*Vsp*	23.75
Average number of paths /*Vsp*	8.86
Average profoundness / *Vsp*	6.06

In the three series of tests, we varied the *similarity threshold* to 78%, 80% and 82%. The tables in each of our experiences will show the following:

- *NbVsp*: the number of specific views attached per cluster;
- *Nb_Nodes*: the number of nodes of type *elements* per cluster;
- *Nbpath*: the number of paths per cluster;
- *ProfMy*: average of profoundness of the specific views attached;
- *SimMy: average of the intra-cluster similarity;*
- *Ect_Typ: the standard deviation intra-cluster.*

With a *similarity threshold* of 78% (*classif78*), the 1606 documents are grouped into 40 clusters (Table 3).

Table 3: Clustering results (classif78)

C_i	Nb*Vsp*	Nb_Nodes	*Nbpath*	*ProfMy*	*Ect_Typ*
C_1	177	3850	1431	0.98	0.00
C_2	34	729	310	0.86	0.02
C_3	186	4222	1770	0.86	0.01
C_4	21	471	193	0.97	0.02
C_5	30	621	246	0.88	0.06
C_6	20	367	135	0.95	0.01
C_7	22	436	218	0.83	0.03
C_8	23	748	98	0.98	0.01
C_9	85	1514	607	0.98	0.00
C_{i0}	40	940	364	0.95	0.03
C_{i1}	105	2479	583	0.98	0.02
C_{i2}	67	1056	419	0.98	0.01
C_{i3}	13	319	121	0.96	0.02
C_{i4}	42	1084	518	0.95	0.03
C_{i5}	56	1244	395	0.97	0.02
C_{i6}	30	654	251	0.83	0.03
C_{i7}	18	467	181	0.89	0.02
C_{i8}	6	143	70	0.91	0.01
C_{i9}	33	810	327	0.95	0.02
C_{20}	29	523	194	0.98	0.02
C_{21}	26	478	174	0.98	0.01
C_{22}	18	425	186	0.95	0.01
C_{23}	34	827	248	0.99	0.01
C_{24}	30	539	216	0.99	0.02
C_{25}	29	529	189	0.98	0.01
C_{26}	7	133	63	0.98	0.01
C_{27}	13	281	86	0.99	0.00
C_{28}	22	474	156	0.98	0.01
C_{29}	8	191	53	0.98	0.01
C_{30}	29	645	220	0.98	0.00
C_{31}	42	1028	305	0.98	0.00
C_{32}	72	1399	508	0.96	0.02
C_{33}	145	3339	1273	0.96	0.02
C_{34}	10	232	124	0.91	0.01
C_{35}	12	255	130	0.95	0.02
C_{36}	44	1257	242	0.98	0.01
C_{37}	17	2170	991	0.98	0.01
C_{38}	5	546	288	0.97	0.02
C_{39}	2	204	115	0.99	0.01
C_{40}	4	509	241	0.95	0.01

With a *similarity threshold* of 80% (*classif80*), the 1606 documents of the corpus are grouped into 42 clusters (Table 4).

Table 4: Clustering results (classif80)

C_i	Nb*Vsp*	Nb_Nodes	*Nbpath*	*ProfMy*	*Ect_Typ*
C_1	177	3850	1431	0.98	0.00
C_2	34	729	310	0.86	0.02
C_3	186	4222	1770	0.86	0.01
C_4	21	471	193	0.97	0.02
C_5	15	344	130	0.95	0.01
C_6	20	367	135	0.95	0.01
C_7	12	281	147	0.94	0.01
C_8	23	748	98	0.98	0.01
C_9	85	1514	607	0.98	0.00
C_{i0}	40	940	364	0.95	0.03
C_{i1}	105	2479	583	0.98	0.02
C_{i2}	67	1056	419	0.98	0.01
C_{i3}	13	319	121	0.96	0.02
C_{i4}	42	1084	518	0.95	0.03
C_{i5}	56	1244	395	0.97	0.02
C_{i6}	16	361	124	0.96	0.01

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

C_{i7}	18	467	181	0.89	0.02
C_{i8}	6	143	70	0.91	0.01
C_{i9}	33	810	327	0.95	0.02
C_{20}	29	523	194	0.98	0.02
C_{21}	26	478	174	0.98	0.01
C_{22}	9	194	90	0.97	0.01
C_{23}	18	425	186	0.95	0.01
C_{24}	34	827	248	0.99	0.01
C_{25}	30	539	216	0.98	0.02
C_{26}	29	529	189	0.98	0.01
C_{27}	7	133	63	0.98	0.01
C_{28}	13	281	86	0.99	0.00
C_{29}	22	474	156	0.98	0.01
C_{30}	8	191	53	0.98	0.01
C_{31}	30	531	225	0.97	0.01
C_{32}	29	645	220	0.98	0.00
C_{33}	42	1028	305	0.98	0.00
C_{34}	72	1399	508	0.96	0.02
C_{35}	145	3339	1273	0.96	0.02
C_{36}	10	232	124	0.91	0.01
C_{37}	12	255	130	0.95	0.02
C_{38}	44	1257	242	0.98	0.01
C_{39}	17	2170	991	0.98	0.01
C_{40}	5	546	288	0.97	0.02
C_{41}	2	204	70	0.99	0.01
C_{42}	4	509	241	0.95	0.01

After examining the results represented in Table 4 (generated clusters and specific views component each cluster) of *classif80*, we have given for each cluster (of *classif80*) its homologous cluster of *classif78* (Table 3). Then we have noticed the emergence of two new clusters. In comparison with the results of *classif78*, we have noticed an improvement of the average similarity intra-cluster and a considerable optimization of the standard deviation (*Ect_Typ*) of intra-cluster of clusters which have undergone changes: lines 5, 7 and 16 of Table 4. In return, the number of generated clusters increases.

With a similarity threshold of 82% (*classif82*), the 1606 documents are grouped into 43 clusters (Table 5).

Table 5: Clustering results (classif82)

C_i	Nb*Vsp*	Nb_Nodes	*Nbpath*	*ProfMy*	*Ect_Typ*
C_1	177	3850	1431	0.98	0.00
C_2	32	678	293	0.86	0.01
C_3	185	4198	1763	0.97	0.02
C_4	21	471	193	0.97	0.02
C_5	15	344	130	0.95	0.01
C_6	20	367	135	0.95	0.01
C_7	12	281	147	0.94	0.01
C_8	23	748	98	0.98	0.01
C_9	85	1514	607	0.98	0.00
C_{i0}	5	120	46	0.86	0.04
C_{i1}	40	940	364	0.95	0.03
C_{i2}	105	2479	583	0.98	0.02
C_{i3}	67	1056	419	0.98	0.01
C_{i4}	13	319	121	0.96	0.02
C_{i5}	41	1061	507	0.95	0.02
C_{i6}	56	1244	395	0.97	0.02
C_{i7}	16	361	124	0.96	0.01
C_{i8}	17	445	170	0.89	0.01
C_{i9}	6	143	70	0.91	0.01
C_{20}	33	810	327	0.95	0.02
C_{21}	29	523	194	0.98	0.02
C_{22}	26	478	174	0.98	0.01
C_{23}	9	194	90	0.97	0.01
C_{24}	18	425	186	0.95	0.01
C_{25}	34	827	248	0.99	0.01
C_{26}	30	539	216	0.98	0.02
C_{27}	29	529	189	0.98	0.01
C_{28}	7	133	63	0.98	0.01
C_{29}	13	281	86	0.99	0.00
C_{30}	22	474	156	0.98	0.01
C_{31}	8	191	53	0.98	0.01
C_{32}	30	531	225	0.97	0.01
C_{33}	29	645	220	0.98	0.00
C_{34}	42	1028	305	0.98	0.00
C_{35}	72	1399	508	0.96	0.02
C_{36}	145	3339	1273	0.96	0.02
C_{37}	10	232	124	0.91	0.01
C_{38}	12	255	130	0.95	0.02
C_{39}	44	1257	242	0.98	0.01
C_{40}	17	2170	991	0.98	0.01
C_{41}	5	546	288	0.97	0.02
C_{42}	2	204	70	0.99	0.01
C_{43}	4	509	241	0.95	0.01

After examining the results in Table 5 (generated clusters and specific views component each cluster) of each clustering *classif80* and *classif82*, we have given for each cluster its homologous cluster of *classif80* (Table 4). Then we noticed the emergence of a new cluster that grouping 5 specific views (Table 5).

7. CONLUSION

This paper is a continuation of our works on the structural clustering of multi-structured multimedia documents. In this work, we have presented in detail our similarity measure and we have studied the impact of the similarity threshold (a parameter fixed a priori) on the resulting clusters.

Our clustering approach is not based on a "*surface similarity*", it's based on a structural similarity taking into account the relationships (*supplementary information*) between the components of the documents to compare. We consider that the relationship between two components is a crucial parameter in our structural comparison process. In fact, the sense of a multimedia document depends not only on the structural elements but also on the relationship between these elements. The proposed measure is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

based on a sub-graph isomorphism that relies on the path matching. This allows keeping the contextual and hierarchical aspects of the matched components.

Along with the construction of clusters, generic views may be transformed [7]. These transformations may lead to an approximation of those clusters or even to their overlapping. To maintain the discriminating power of clusters (*generic views*), we must ensure that they are sufficiently distant. In this context, we have proposed to use an *inter-cluster threshold*. This allows maintaining the cluster stability. Moreover, increasing the dissimilarity between clusters can reduce noise and increase the clustering precision.

In our future works, we will study the impact of inter-cluster threshold on clusters generated by our clustering process and we will show in detail the cost of transforming generic views (representatives of clusters).

REFRENCES:

[1] Champin P-A., Solnon C.,"Measuring the similarity of labeled graphs". Dans 5th Int. Conf. On Case-Based Reasoning (ICCBR 2003), Kevin D. Ashley and Derek G. Bridge ed. Trondheim (NO). pp. 80-95. LNAI 2689. Springer Berlin. 2003.

[2] Costa, G., G. Manco, R. Ortale, et A. Tagarelli, "A Tree-Based Approach to Clustering XML Documents by Structure". In PKDD, 2004, p 137–148.

[3] Dalamagas T., Cheng T., Winkel K-J, Sellis T.K. 2006, "A methodology for clustering XML documents by structure". Information Systems 31(3), 2006, p187-228.

[4] 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.

[5] Gentner D., "Structure-mapping: A theoretical framework for analogy", Cognitive Science, 7, (Reprinted in A. Collins & E. E. Smith (Eds.), Readings in cognitive science: A perspective from psychology and artificial intelligence. Palo Alto, CA: Kaufmann), 1983, p155-170.

[6] Harchaoui Z. et Bach F. "Image Classification with Segmentation Graph Kernels", Dans CVPR. IEEE, 2007.

[7] Idarrou, A., Mammass, D., Soulé-Dupuy, C., and Vallès-Parlangeau, N.,"A generic Approach to the Classification of Multimedia Documents: a Structures Comparison" ", In ICGST-ICISP Special Issue on *GVIP*, December 2010.

[8] Idarrou A., Mammass D. 2012, "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.

[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, Montpellier (France), 29/05/2012-31/05/2012, Association INFORSID, 2012, p. 539-554.

[10] Kutty S., Tran, T., Nayak, R., et Li, Y. "Clustering XML Documents Using Closed Frequent Subtrees" : A Structural Similarity Approach‖. Lecture Notes In Computer Science, , 2008, p183-194.

[11] Mbarki M. 2008, 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.

[12] Portier P-E, « Construction des Documents Multistructurés dans le Contexte des Humanités Numériques », Thèse de Doctorat de l'INSA De Lyon France, 2010.

[13] Razo F. D., A. Laurent, et Teisseire M., « Représentation efficace des arborescences pour la recherche des sous-structures fréquentes », In Actes de l'atelier Fouille de données complexes, Conférence Extraction et Gestion des Connaissances (EGC 2005), pp. 113.120.

[14] Tagarelli A., Greco S., "Semantic clustering of XML documents". ACM Trans. Inf. Syst. 28(1), 2010.

[15] Termier A., Rousset, M. C., et Sebag, M., "TreeFinder: a First Step towards XML Data Mining", Proceedings of the IEEE International Conference on Data Mining (ICDM'02), IEEE Computer Society Washington, DC, USA, 450.

[16] Vercoustre, A. M., Fegas, M., Lechevallier, Y., Despeyroux, T., et Rocquencourt, I. « Classification de documents XML à partir d'une représentation linéaire des arbres de ces documents ». Paris, France, 1006, p433–444 .

[17] Saleem Khalid. "schema matching and integration in large scale snario". Thèse de

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Doctorat de L'Université Montpellier II, France, 2008.

[18] Sorlin S., C. Solnon, "Reactive tabu search for measuring graph similarity". In 5th IAPR-TC-15 workshop on Graph-based Representations in Pattern Recognition, Luc Brun, Mario Vento ed. Poitiers, 2005, pp. 172-182. Springer-Verlag..