Classification of Multi-structured Documents: A Comparison Based on Media Image

Ali Idarrou[1,2], Driss Mammass[2],
Chantal Soulé Dupuy[1], and Nathalie Valles-Parlangeau[1]

[1] IRIT, groupe SIG, Université Paul Sabatier
118, route de Narbone, 31062 Toulouse cedex 9 France
[2] IRF – SIC Université Ibn Zohr Agadir Maroc
idarrou@irit.fr, driss_mammass@yahoo.fr,
{chantal.soule-dupuy,nathalie.valles-parlangeau}@univ-tlse1.fr

Abstract. This paper focuses on the structural comparison of multimedia documents. Most of the systems treating the multimedia documents exploit only the text part of these documents. However, the text is no longer the only means to carry information. The major issue is to extend these systems to the other modality notably to the image that constitutes one of the basic components of multimedia documents. The complexity of multimedia documents, multi-structured in essence, imposes not only a structural representation in the form of trees, but rather in the form of graphs. The graphs are in appropriateness to the description of these documents. For example, one will be able to describe the components of a scene of an image, the relations between these components, their positions (spatial relations), etc.

We propose a new similarity measure of graphs, based on a univocal matching between the graphs to compare. In our approach, we will take account of structural information and specificities of multimedia information. We evaluate our measure on a corpus of multi-structured documents from the INEX 2007 corpus.

Keywords*:* Documents, multimedia information, image, clustering, classification, similarity, matching of graphs.

1 Introduction

The technological evolution, of digitalization, coding, compression, storage and communication, contributes to the massive provision of numerical multimedia information (text, image, audio, video …) of which the exploitation necessitates appropriate methods, models and tools. The numerical multimedia documents are complex in essence. Coming up from the composition of different documents, themselves more or less complex, information is not as directly accessible as in text documents, for example. That puts us in the obligation to use tools and techniques allowing a synthetic and complete description of this type of multi-structured documents. Each multimedia document has its own structure, and each document composing it has also at least one structure if not several. Their management and, in

A. Elmoataz et al. (Eds.): ICISP 2010, LNCS 6134, pp. 428–438, 2010.

particular, their storage imposes that we no longer consider these documents as simple arborescence forms but rather in the form of graphs. In fact, the multi-structuralism induces more complex relations (there may be multiple relations between the same two nodes).

The graphs are mathematical tools that allow modelling structured objects; they are in appropriateness with multimedia numerical document representation in XML format and of sub-documents that they contain. For example, one will be able to describe the objects and the sub-objects which are components of a scene of an image, the relations between these objects, and their positions (spatial relations, etc).

Unlike the classical methods of image description, such as the descriptors of textures, the color histograms, etc, the graphs allow the modelling of different facets the images contents, like the nature of objects, the spatial relations, etc.

Most of the systems treating the multimedia documents exploit only the textual part of these documents. The major issue is to extend these systems to other modalities notably to the image that constitutes one of the basic components which is very frequent in the multimedia documents. The description of the text and the image (and other modalities) by using the same model of representation will allow combining jointly several methods in order to exploit multimedia information easily. In the research for information for example, the need of the user can be a fragment of text (or image), an image or the text + the image. The result can be one or more sub-graph(s).This technique will allow access to pertinent information in the multimedia documents for example. It will equally allow access to the image by context (using the neighboring nodes of the image).

The documentary classification is a solution which allows optimizing the access time in the relevant information in a large mass of data. We are interested in the structural clustering as an organization of documents in the form of classes of typological documents having similar structures.

In this paper, we are interested in the "text" and the "image" media, but our approach can be applied to the other types of media (audio and video). We will take account of structural information and multimedia specificities and will present a new similarity measure of graphs, based on a univocal matching between the graphs. This matching will take into account both the ontological context and the hierarchical context of the components of graphs to match.

In the following section we will present some examples of applications using graphs to represent the objects and we present the measures of graphs similarity most used in literature. In the third section, we will present our approach of multimedia document comparison, and we will finish by experimentation on the INEX 2007 corpus.

2 Representation of Objects Using Graphs: Comparative Study

The graphs are widely used in many applications such as in information retrieval, in image processing, in organic chemistry to represent molecules, etc. If objects are represented by graphs, measuring their similarity means matching their graphs so as to find their common points and their differences. The set of techniques and mathematical tools developed in graph theory, allows easily demonstrating the

properties, to deduce the characteristics of an object from a similar object, etc. Several previous works have used the graphs as models of representation of objects for analyzing and comparing. The graphs are used in various domains to represent objects. In organic chemistry, for describing the structures of the molecules where every atom is represented by a node of graph and each connection between atoms by an arc. The nodes and arcs have labels symbolizing the object they model. This allows to synthetize molecules and to show for example that a molecule enters the composition of another molecule [4]. In CAD, the multi-labelled graphs are used to represent objects of computer-aided design [8], [9]. Each graph node represents a component of the object and each arc of the graph represents a binary relation between two components. Labels are assigned to each node and to each arc in order to express the different types of components and relations. The objective in this context is to identify automatically all the objects already designed and similar to a new object to be designed. The matching used in this application is a multivoque matching: a component of one graph is put in correspondence with more than a component of the other graph (two components of a graph, for example, can play the same role as one component of the other graph). Let us note that two objects can be similar in spite of the fact that they are not identical. The labelled graphs are also used to represent the images to compare [12]. The images are often segmented in regions (nodes represent regions (labelled by their properties: color, size of the scene,...) and arcs represent the various binary possible relations between regions).

In our previous works [1], [2], we used the graphs to represent the multimedia documents in order to classify them within the framework of a documentary warehouse. It was a question of finding, among a set of multimedia documents, the most similar to a given document.

Different similarity measures of graphs exist in the literature. We present in what follows an overview on the most related measures:

- To calculate the similarity of two XML documents represented by trees, [6] has introduced the following measure:

$$Sim(T,T') = 1 - \frac{\sum_{vj} Danc(vj)}{\sum_{vj} Panc(vj)} \tag{1}$$

Where Danc(vj) : represents the distance of the ancestors of v_j.
and Panc(vj) : represents the weight of the path v_j.

- To evaluate the similarity of trees T_1 and T_2, representing XML documents, [3] had proposed the following measure:

$$Similarité(T_1,T_2) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} Sim(e_{1i},e_{2j})}{Max(|T_1|,|T_2|)} \tag{2}$$

Where $|T_1|$: the size of T_1, $sim(e_{1i},e_{2j})$ represents the similarity of nodes e_{1i} and e_{2j} belonging respectively to T_1 and T_2.

- In [8], [9], the similarity of two labeled graphs G and G', representing the objects to be designed, is based on a multivalent matching of peaks of these two graphs: a component of an object can play the same role of several other components of the subject:

$$Sim(G,G') = \max_{M \subseteq VxV'} \left[\frac{f(descr(G) \cap_M descr(G')) - g(splits(M))}{f(descr(G) \cup descr(G'))} \right] \quad (3)$$

Where f and g are two functions that depend on the application.

Splits(M) : Indicates the set of nodes of the graph, matched to more than one node of the other graph. Descr(G) $\cap_M$ descr(G') : Indicates all the common characteristics in G and G' with regard to the matching M.

- In [10], the authors showed that the measure (3) is not directly suited in certain contexts and they proposed a new measure of similarity of graphs that represent images. They showed that this measure is generic because it is parameterized by two functions, Sim_v and Sim_e, which depend on the considered application:

$$Sim(G,G') = \max_{m \subseteq VxV'} \left[\frac{\sum_{v \in V \cup V'} Sim_v(v, m(v)) + \sum_{(u,v) \in E \cup E'} Sim_e((u,v),(m(u,v)))}{|V \cup V'| + |E \cup E'|} \right] \quad (4)$$

Where G = (V, E) and G '= (G', V '), V (resp. V') set of nodes of G (resp. G '), E (resp. E') set of arcs of G (resp. G ') and m is a graph matching.

The above examples have discussed techniques for comparing objects represented by graphs, based on:

- The research for a univocal matching: a node and/or an arc of a graph can be put in correspondence with, at maximum, one node and/or an arc of the other graph (isomorphism of graphs, isomorphism of sub graphs, larger sub common graph to two graphs).
- The research for a multivocal matching: a node and/or an *arc* of a graph can be put in correspondence with one or several nodes and /or *arc*s of the other graph.

We noticed that these similarity measures are all based on the research for a better matching between the graphs to compare and that the preferences and the constraints imposed by the desired matching, for every measure, depending on the application. They are not defined in the same context which imposes, each time, constraints on the desired matching between graphs.

These constraints depend on the problem to be resolved and differ from one application to another. It is thus necessary to find each time the best adequacy between the objective and the effective behaviours of the defined measure.

Therefore, it is difficult to compare and apprehend the results of many studies that have addressed this topic.

However, the measure (2) takes into account neither the preservation of arcs nor the order of matched components. Nevertheless, two nodes can be synonymous but they have different semantics in the two trees.

On the other hand, the measures (3) and (4) do not take into account the distribution of the components of both graphs (depth level, order of components).

Concerning the measure (1), it just allows to evaluate the number nodes of T in T' (does not consider the relations between nodes).

In the next section, we will present our approach based on media image.

3 Structural Comparison of Multi-structured Documents

Unlike conventional approaches of comparison of documents, based on similarity of surface and which exploit only the textual part of documents, we are going to take into account in our approach, the structural information and the multimedia specificities to give more efficiency to the process of documentary comparison.

In this paper, we focus on the "text" and the "image" media, but our approach can be applied to other media (audio and video).

Generally, the images are segmented by regions. Every node (of graph) represents then a region and the arcs represent the various possible relations between regions. The image segmentation is not our objective but is our point of departure. Whatever is the used facet, the image can be represented by graph. The description of the text and the image (and the other modalities) by using the same model of representation is going to allow to combine jointly several modalities in order to exploit easily the multimedia information. In information research for example, one user can need a fragment of text (or image), image or text + image. The result returned can be one or more sub graph(s). This description also allows access to the image by context (using nodes neighbours of the image).

The choice of graphs as a tool of description is consistent with the representation of multimedia documents whose structural complexity is high. These tools allow a rich modeling of objects (and sub-objects) and of their relations. To compare two documents, simply compare their graphs. This matching is going to allow evaluating the number of nodes and arcs that they have in common. More precisely, the isomorphism of (sub) graphs can show one of the graph is included in an other (or both graphs are structurally identical) while the intersection between graphs can show the commonalities between objects represented by these graphs.

In this paper, the documentary structures are described by the meta-model MVDM [5]. This model consists of two levels: a generic level where the generic views are represented by graphs (a class represents a set of typologically similar documents while keeping their specific characteristics) and a specific level where the specific views are represented by trees.

Our integration process of a document in the documentary warehouse is composed of the following steps:

a) Extraction of the Generic View of a Document

To extract the generic view of an XML document, we use two parsers. The first one is based on Sax (Simple API for XML) which returns all the tags encountered in the

document. Then the semantic tags are intercepted by a second parser that can analyze, filter and make changes to get the generic view of the document. The generic view thus obtained is used as representative of a document in our matching process.

b) Filtering of the Generic Views

This step consists in selecting the generic views Vgi of the documentary warehouse that have one degree of resemblance with the generic structure Vrep of the previous step. The problem thus consists in finding an univocal matching φ_n, (a node and either arc of a graph is matching in at most a node and/ or arc of the other graph), between Vrep and Vgi who allows to make correspond the maximum of nodes and /or arcs of Vrep and Vgi. To preserve semantics (spatial relations, etc.) φ_n has to preserve arcs between components matched by Vrep and Vgi. We propose the function C_f, which calculates the coefficient of filtering C_f (Vrep, Vgi) from which we can select a generic view Vgi of documentary warehouse (DW):

$$C_f : \{Vrep\} \text{ x } DW \rightarrow [0,1]$$

$$(Vrep, Vgi) \rightarrow C_f(Vrep, Vgi)$$

Formally Vrep = (V, E), Vgi = (Vi, Ei) where V (resp. Vi) is the set of nodes of the graph Vrep (resp. of the graph Vgi), E (resp Ei) is the set of arcs of the graph Vrep (resp. of the graph Vgi) and Vgcand={Vgi∈ DW / C_f(Vrep,Vgi)>=S_f } set of Vgi, candidates to comparison and S_f is a constant, determined in priori by experiments, indicating the threshold of filtering. We define the coefficient of filtering C_f (Vrep, Vgi), which depends on the matching φ_n research, as follows:

$$C_f(Vrep, Vgi) = \left[\frac{|\varphi_n(V)|}{|V|} + \frac{|\varphi_n(Vi)|}{|Vi|} \right] / 2 \qquad (5)$$

Where |X| indicates the cardinal of the set X and φ_n (V)={u'∈Vi / ∃ u∈ V ; φ_n (u)=u'}, φ_n (Vi) ={u∈V / ∃ u'∈Vi ; φ_n (u')=u}, φ_n an univocal matching between Vrep and Vgi (resp. between Vgi and Vrep) that preserves the arcs: φ_n :
Vrep → Vgi
u → v where v similar to u (ie u and v have the same labels or the similar labels).

C_f (Vrep, Vgi) allows evaluating the percentage of nodes common to Vrep and Vgi. In fact, the first part of C_f (Vrep, Vgi) determines the percentage of matched nodes of the structure Vrep, that is the degree of inclusion of Vrep in Vgi and the second part determines the percentage of matched nodes of the structure Vgi that is the degree of inclusion of Vgi in Vrep. The generic structures of the warehouse, for which C_f (Vsp, Vgi) is upper or equal to S_f are selected for the next steps of the comparison phase.

c) Reconciliation of Generic Views

This step consists in reconciling the generic views of the warehouse, candidate in the comparison, to the generic view Vrep representing the specific view of the new document. Some possible additions of nodes and arcs, missing in the generic views of the warehouse, can be envisaged.

d) Weighting

This step consists in attributing a weight to every relation. This value depends on the position of the relation in the graph. It thus allows translating the hierarchical context of arcs (Fig.1). We propose the weighting function from E (resp. of Ei) to] 0,1 [:

$$P_e : \begin{array}{l} E \mapsto \ ]0,1[\\ (u, v) \mapsto P_e(u, v) \end{array} \qquad P_e(u,v) = \begin{cases} 1 - \dfrac{ord(v)}{k} & \text{si } depth(v)=1 \\ P_e(x,u) - \dfrac{ord(v)}{k^{depth(v)}} & \text{otherwise, where node u extremity of the arc (x,u)} \end{cases} \tag{6}$$

K is a fixed parameter by the user (a power of 10) indicating the maximum number of son nodes (number of sons< k) for each node of the manipulated graphs and depth(v) is the profouness of v (depth of root node=0. For example (Fig.1), P_e(name,Id)=0.89.

The weighting function allows penalizing the arcs profoundness. The arc profoundness i (i> = 0) have more important than the arc profoundness i +1.

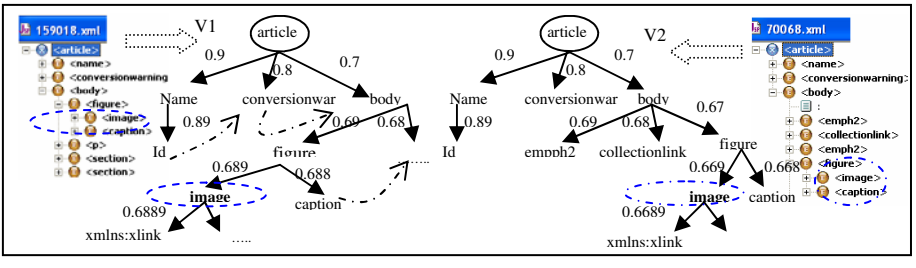

Fig. 1. Example representation, hierarchical relation, temporal relation and weighting

e) Calculation of Score Similarity

The similarity between the graphs Vrep and Vgi is based on the research for a better univocal matching, between both graphs: a matching which maximizes the similarity between Vrep and Vgi, while respecting the hierarchical order of components and relations matched. Our similarity measure takes into account at once ontological and hierarchical contexts of nodes and arcs matched. The decision depends on the threshold of similarity. More formally, if Sim (Vrep, Vgi) > =Ss then the graphs Vrep and Vgi are considered as similar. In this case the specific view represented by the graph Vrep will be the aggregated into the generic view Vgi, of the warehouse (the most similar). Otherwise, a new class is created; it is represented by the view Vrep.

Concerning the threshold of similarity Ss, several tests were realized to determine the optimal value. We noticed that the increase of the value of Ss leads to the creation of numerous classes. On the other hand, the decrease of this value implies the growth of the number of documents associated to each class that arouses heterogeneousness between documents of the same class. We noticed that the value of 0.8 (80 % of similarity) gave good results. To evaluate the similarity of two graphs, we propose:

$$Sim(G, G') = \begin{cases} 1 - \frac{(d_1 + d_2)}{2} & if \ d_1 * d_2 \neq 0 \\ 1 - (d_1 + d_2) & \text{otherwise} \end{cases}$$

$$\text{where } d_1 = \frac{1}{Nbchm_G} \sum_{vj \in V / vj \text{ leaf}} Dchm(v_j) \text{ and } d_2 = \frac{1}{Nbchm_{G'}} \sum_{v'j \in V' / v'j \text{ leaf}} Dchm(v'_j) \quad (7)$$

Where G=(V,E), G'=(V',E'), $Dchm(v_j) = \frac{\sum_{e_i \in Dchm(v_j)} D(e_i)}{\sum_{e_i} P_e(e_i)}$,

$$D(e_i) = \begin{cases} | P_e(e_i) - P_e(e'_i) | & \text{if } \exists e'_i \in E' / \varphi_e(e_i) = e'_i \\ P_e(e_i) & \text{otherwise} \end{cases}$$

φ_e matching between E and E'

For example, the similarity between the paths (Fig.1) v1:article\body\figure\image\ xmlns:xlink and v2:article\body\figure\image\ xmlns:xlink is equal to 0.97.

Unlike the measure (1), our measure allows to evaluate the degree of inclusion in both senses ($G \subseteq G'$ and $G' \subseteq G$). To preserve the semantics, this measure is based on a matching that preserves the arcs (hierarchical and spatiotemporal relations) and order of the matched components. This technique allows preserving the synchronization between objects matched of documents to compare (which is not the case for the measures (2), (3) and (4)).

In the context of comparison of the views of digital multimedia documents represented by graphs, our measure is more effective. Unlike the studied measures in section 2, our measure takes into account at once the ontological, hierarchical contexts and multimedia specificities, it is based on a matching which preserves the order of components and arcs between these components (to preserve semantics of the matched components).

4 Experimental Results

To validate our approach, we developed a prototype in Java. This prototype consists of two modules: the first allows parsing the document to extract the generic view. The second module allows evaluating the similarity of the generic view, that result of the first module, with the generic views of the documentary warehouse.

We used this tool to process XML documents 102 (1515 KB) extracted from the INEX 2007 corpus (http://wwwconnex.lip6.fr/~denoyer/wikipediaXML/). The 102 documents included in the documentary warehouse must be grouped into classes of similar documents (structurally). with a degree of similarity of 80%. We obtained the following clustering:

We notice that we have four classes for 102 documents. One class presents the peculiarity to represent only six documents. Otherwise, the other classes are rather homogeneous. In our comparison process, we imposed the constraint of order and that of the conservation of arcs; the graph matching (between the graphs G and G ') must respect the order of the matched components (1) and it must preserved the arcs (2).

Table 1. Characteristics of the corpus used

Number of documents	**102**
Average number of elements / Vsp	147,7
Average number of elements / Vrep	26,75
Average number of attributes / Vsp	194,3
Average number of attributes / Vrep	26,2
Average number of paths /Vrep	32,15
Average depthes /Vrep	6,5
Average capacity in KB / doc	14,85

Table 2. Number of aggregated views per class

GenericClass	Number of aggregated views
C1	39
C2	6
C3	12
C4	41

(1) : $\forall (u,v) \in E,\ ord(u) < ord(v) \Rightarrow ord(\varphi(u)) < ord(\varphi(v))$

(2) : $\forall (u,v),\ (u,v) \in E \Rightarrow (\varphi(u), \varphi(v))) \in E'$ *where* $G=(V,E)$ *and* $G'=(V',E')$

In the step 3.b) of filtering, the generic views Vgi of DW which we can't find a graph match (between Vrep and Vgi) verifying (1) and (2) are rejected. This is one of the reasons that have a class represents only six documents.

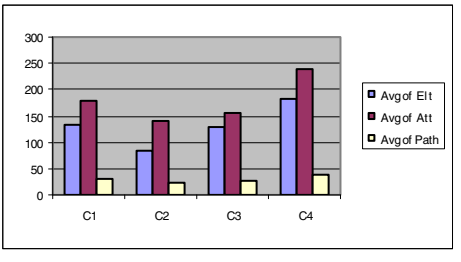

Fig. 2. The average number of the generic elements, attributes and paths per class

The first results of our experiments, on a corpus of 102 documents, are interesting and encouraging; they allow confirming the feasibility of our approach.

5 Conclusion and Outlook

In this paper, we are interested in the "text" and "image" media. We presented an approach of classification of digital multimedia documents based on the matching of the graphs representing documents. The choice of the graphs, as tool of description, is in adequacy with the multimedia XML documents. These tools allow a rich modeling of complex and structured objects. Our approach can be applied to other media types (audio, video).

We proposed: (1) a *filtering function* allowing selecting the generic views of the documentary warehouse, (the views candidates in the comparison). (2) a *weight function* of arcs (relations between components), taking into account the hierarchical

context (position of components matched of graphs to compare). (3) a *new structural similarity measure* of graphs, based on an univocal matching which preserves arcs (preserve semantics between the matched components) and take into account the multimedia character: temporal, spatial relations,...

We consider that our proposal is generic in that a number of studies, using trees (a tree is a particular case of graph) to describe documents, can be seen as particular cases of our approach. Our representation model is consistent with the media image (as a component of a multimedia document) independently of the annotation used. This model also allows navigation both between the structures of components of a document, and between different views of the same document.

Our future works will focus on improving our algorithms to optimize the comparison process and making our approach more efficient and we'll show the genericity of this approach.

Acknowledgment

With a gant of the "Action intégrée Maroco-Française" n° MA/10/233

References

1. Idarrou, A., Soulé-Dupuy, C., Mammass, D., Vallès-Parlangeau, N.: Appariement et similarité des graphes. In: JDTIC'09 Journée doctorales en Technologie de l'Information et de la Communication Juillet 16-18, Rabat Maroc (2009)
2. Idarrou, A.: Classification des documents multi-structurés: comparaison de structures Ateliers Jeunes Chercheurs. In: CIFED 2010, Sousse Tunis, pp 501–506 (2010)
3. Mezghiche, A.A., Souam, F.: Classification de structures arborescentes: cas de documents XML. In: COREA 2009, 6th French Information Retrieval Conference, Proceeding of LSIS-USTV 2009, Presqu'île de Giens, France, May 5-7, pp. 301–317 (2009) ISBN 2-9524747-1-0
4. Régin, J.C.: Développement d'outils algorithmiques pour l'intelligence Artificielle. Application à la chimie orgnique. PhD thesis, Université de Montpellier II (1995)
5. Djemal, K., Dupuy, C.S., Valles-Parlangeau, N.: Modélisation d'un Entrepôt de Documents Multi-structurés Dans: ÉDIT 2007: actes du colloque des doctorants de l'École Doctorale Informatique et Télécommunications, Toulouse, May 24-25 (2007)
6. 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 (2008)
7. Torjmen, M., Pinel-Sauvagnat, K.: Une étude que l'impact de la structure sur la recherche multimédia. In: Conférence frnacophone en Recherche d'Information et Application (CORIA 2009), Presqu'île de Giens-Var, 05/05/2009-07/05/2009, Ludovia, mai 2009, pp. 51–66 (2009)
8. Champin, P.-A., Solnon, C.: Measuring the similarity of labeles graphs. In: Ashley, K.D., Bridge, D.G. (eds.) ICCBR 2003. LNCS, vol. 2689, pp. 80–95. Springer, Heidelberg (2003)
9. Sorlin, S., Champin, P.-A., Solnon, C.: Mesurer la similarité de graphes étiquetés: Dans 9èmes Journées Nationales sur la résolution pratique de problèmes NP-Complets. In: JNPC 2003, pp. 325–339 (2003)

10. Sorlin, S., Sammoud, O., Solnon, C., Jolin, J. M.: Mesurer la similarité de graphes: Dans Extraction de Connaissance à partir d'Images (ECOI 2006). In: Vincent, N., Lomenie, N. (eds.) Atelier de Extraction et Gestion de Connaissances (EGC 2006), Lille, pp. 21–30 (2006)
11. Jouili, S., Tabone, S.: Applications des graphes en traitement d'images. In: International Conference on Relations, Orders and Graphs: Interaction with Computer Science-ROGICS'08, Mahdia, Tunisia, pp. 434–442 (2008)
12. Ambauen, R., Fischer, S., Bunke, H.: Graph edit distance with node splitting and merging, and its application to diatom idenfication. In: Hancock, E.R., Vento, M. (eds.) GbRPR 2003. LNCS, vol. 2726, pp. 95–106. Springer, Heidelberg (2003)
13. INEX 2007, Collection (same as old collection with Image IDs) (2007), http://www-connex.lip6.fr/~denoyer/wikipediaXML/