Report on the XML Mining Track at INEX 2005 and INEX 2006, Categorization and Clustering of XML Documents

Ludovic Denoyer, Patrick Gallinari, Anne-Marie Vercoustre

Report on the XML Mining Track at INEX 2005 and INEX 2006

Categorization and Clustering of XML Documents

Ludovic Denoyer[1], Patrick Gallinari[1], Anne-Marie Vercoustre[2]

[1] LIP6 - University of Paris 6
firstname.lastname@lip6.fr
[2] INRIA Rocquencourt
anne-marie.vercoustre@inria.fr

Abstract. This article is a report concerning the two years of the XML Mining track at INEX (2005 and 2006). We focus here on the classification and clustering of XML documents. We detail these two tasks and the corpus used for this challenge and then present a summary of the different methods proposed by the participants. We last compare the results obtained during the two years of the track.

1 Introduction

The XML Document Mining track[3] was launched for exploring two main ideas: first identifying key problems for mining semi-structured documents and new challenges of this emerging field and second studying and assessing the potential of machine learning techniques for dealing with generic Machine Learning (ML) tasks in the structured domain i.e. classification and clustering of semi structured documents.

This track has run for two editions: 2005 and 2006, and the present report summarizes the work done during these two years [4]. The track has been supported through INEX by the DELOS Network of excellence on Digital Libraries and by the PASCAL Network of excellence on Machine Learning.

Among the many open problems for handling structured data, we have focused in this track on two generic ML tasks: *classification* and *clustering*. The goal of the track was therefore to explore algorithmic, theoretical and practical issues regarding the classification and clustering of XML Documents. Note that one new task - Structure mapping[5] - has been proposed in the 2006 edition of the track but since only two submissions was made for this more complex task ([1],[2]), we will only discuss here the results obtained for classification and clustering. In the following, we first describe the mining problems addressed at

[3] http://xmlmining.lip6.fr
[4] the challenge will continue one year more in 2007
[5] Structure Mapping was defined as learning from examples how to map documents in one or several formats onto a predefined mediated schema

INEX in section 2, we then describe in section 3 the instances of the mining tasks at INEX 2005 and 2006. Finally in section 4.1 we summarize the different contributions of participants.

2 Categorization, Clustering of XML Documents

Dealing with XML document collections is a particularly challenging task for ML and IR. XML documents are defined by their logical structure and their content (hence the name semi-structured data). Both types of information should be addressed in order to effectively mine XML documents. Compared to other domains, where the structured data consists only of the "structure" part with no content this is much more complex and this had been addressed only for very specific problems in the ML community. Note that most existing ML methods can only deal with only one type of information (either structure or content). Structure document information is described through a labelled tree where labels do correspond to XML tags which may or may not carry semantic information. In the used document collections, content information is composed of text and images, but only the textual part was considered in the mining track of INEX. The textual content is usually contextually dependent of the logical structure (e.g. section content vs header or bibliography information). It should be stressed that XML documents usually come in large collections and that scalability is a fundamental issue for mining semi-structured data.

When dealing with semi-structured documents, according to the application context and on the prior information available on the collection, it may be relevant to consider the structure information alone or to consider both structure and content information. Two types of tasks were then defined corresponding to these two conditions: "Structure Only" (SO) and "Structure and Content" (SC).

Dealing with structure alone measures the ability to recover or classify structural classes corresponding to different document sources. The structure + content tasks are more challenging and encompass many different generic tasks in the document domain. In order to define more precisely the classification and clustering problems, let us consider the two following dimensions: structure (S) and thematic content (T). The former dimension characterizes the document structure generation and the latter its content class. Figure 1 illustrates a partition of a document collection composed of different thematic areas and structures. The goal of classification and clustering will be to identify different classes or groups of documents corresponding for example to different information sources (each source or "class" corresponds to a color on Figure 1). Classification will aim at discriminating between the different sources, clustering will try to recover some hidden source information. In this example, each source may be considered as a mixture of one or more themes and structures in different proportions. The different sources may overlap in different proportions leading to tasks with different complexity.

For a classification problem for example, the models will have to classify specific mixtures of content and structure information components corresponding to each source.

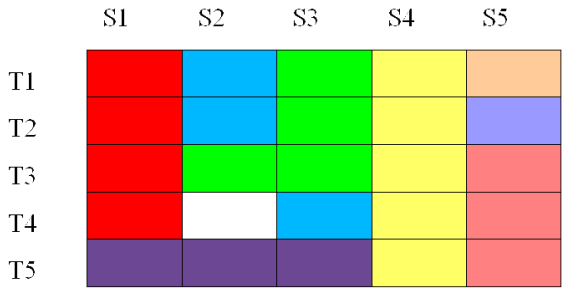

Fig. 1. The structural and thematic dimensions do respectively correspond to columns and rows in the figure. All documents in column S1 for example have structure S1 - or follow schema S1 - while all documents in row T1 have thematic T1. The different information sources are identified by colors. In this example, there are 9 distinct sources, each identified by a given color, among a maximum of 25 potential sources. Each source is defined here as a mixture of several structure and themes, for example, on the left bottom, documents share thematic content T5 and may have structure S1, S2 or S3.

Depending on the application in mind, one may distinguish different generic problems like:

- identify common structures across different content types or themes (structure oriented classification and clustering) - this is illustrated in Figure 2 where each class corresponds to a specific structure and deals with all the collection themes. The theme may appear in different proportions for each class
- identify different content types across structures (content oriented classification and clustering). Classes would then correspond to the horizontal lines in the figure.
- identify homogeneous classes for both structure and content information (mixed clustering and classification) - this corresponds to Figure 1 and is the more general task. It may come in many different variants.

From this enumeration, it should be clear that many different classification and clustering problems may occur in real situations and structural and content information will play different roles for these two generic tasks.

3 Tasks and Evaluation

We will now describe the different corpora and tasks proposed during the two years of the XML Document Mining track. During the first year of the track

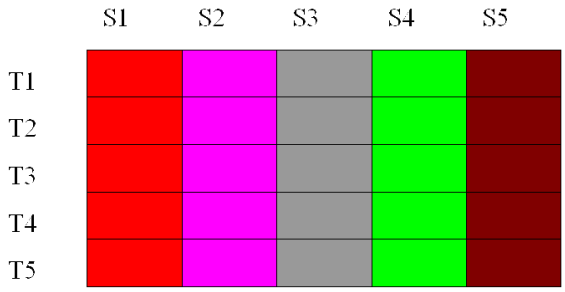

Fig. 2. Each source corresponds to a specific structure (5 sources here - corresponding to columns) - the different thematic (rows) will appear in different proportions in each source.

both the *structure only (SO)* classification/clustering tasks and *structure and content (SC)* tasks where considered, while the second year was more focused on the *structure and content* tasks.

3.1 Corpora

The different tasks proposed during the XML Mining Track were evaluated on three different XML corpora whose characteristics are provided in Table 1. The movie corpus is an artificial corpus based on the IMDB[6] Corpus while IEEE and Wikipedia are real world document corpora. In some cases the classes or clusters correspond to a natural partition of the corpus, while in other cases, several classes were generated artificially from an homogeneous corpus as detailed below. Three corpora have been used for the track:

1. The movie corpus is composed of about 9,500 XML documents that describe movies.
2. The IEEE corpus is composed of 12,000 scientific articles from IEEE journals in XML format. This has been the reference corpus for INEX from year 2002 to 2005 [3].
3. The XML Wikipedia corpus is composed of 150,000 english documents from Wikipedia, formatted in XML. This corpus is a subpart of the official corpus for INEX 2006 where each document belongs to exactly one category. A complete description of this corpus is available in [4] under the name of *classification corpus*.

3.2 Structure only task

Structure Only tasks correspond to classification or clustering using the structural description of XML documents alone, i.e. the XML ordered labelled tree

[6] http://www.imdb.org

Corpus	Nb Docs	Nb node labels	USed for SO	Used for SC
Movie	9,463	≈ 190	YES	NO
IEEE	12,108	≈ 150	YES	YES
Wikipedia	$\approx 150{,}000$	$\approx 10{,}000$	YES	YES

Table 1. Description of the different corpora used. Corpora were splitted using 50% of the documents for training and 50% for testing.

providing the relations between the document elements. The input space in this case corresponds to the tag alphabet, and to the relations between these tags. Note that the tag space is of limited size here, much less than the size of the dictionary for CS tasks, but can be quite high (up to 10 000 different tag names for Wikipedia).

Dealing with structure alone measures the ability to identify information sources in the context of XML data - e.g different Web servers, XML databases, ... Note that in many other domains (e.g. biology) data also come as ordered labelled trees so that investigating classification and clustering methods for such trees is a generic problem, with many applications outside the field of XML documents.

For the *structure only* task, we have proposed a set of five different classification/clustering subtasks. Four are based on the movie corpus, one on the IEEE corpus. For each subtask, the goal is to recover by classification or by clustering the different classes, i.e. the different structural sources in the corpus.

- **Movie N SO task**: This task is based on the *m-db-s-N*[7] corpora that are composed of XML documents describing movies expressed in eleven different schemas. These schemas were defined by hand and the documents from the corpus were mapped onto each schema using XSLT scripts. Number N quantifies the difficulty of the task: the higher N is, the more overlap there is among the 11 classes. We have defined 4 tasks identified by N = 0 to 3.
- **IEEE SO task**: This task consists in identifying two different families of structures coming from a same source. The *INEX SO* corpus used here corpus corresponds to the reference INEX IEEE corpus [3] where content information has been removed. Documents in this corpus come from two structural sources *Transactions on ...* journals and other IEEE journals. The two sources are composed of documents following the same DTD but with different writing rules. The goal here is to recover this class information.

3.3 Structure and content tasks

In these tasks, the goal is to identify categories defined corresponding to structural and thematic characteristics. Classes or clusters are then to be identified using both the structure and the content of the XML documents. We have used two different corpora here for defining two different tasks:

[7] http://xmlmining.lip6.fr

- **IEEE SC**: This task amounts at identifying 18 categories of the IEEE corpus that correspond to the 18 different IEEE journals present in the collection. These categories are both structural and thematic. As said before, there are two broad structural sources in this corpus (Transaction journals and other journals), while the same thematic can be covered in two different journals.
- **Wikipedia SC**: This task proposes to identify 60 categories in a corpus of about 150,000 XML documents that comes from the wikipedia XML corpus. In this corpus each document belongs to exactly one category and each category corresponds to a *portal* of wikipedia - see [4] for more informations. This collection allows us to evaluate the capacity of the different models to deal with large scale XML corpora. Note that the categories information that appear in the wiki document have been removed from the XML files.

3.4 Evaluation measures

Different measures have been used during the first year of the track and the second year. In all experiments, each document belongs to exactly one category or one cluster (see part 3.1).

Precision, Recall and F1 measure for classification In the field of classification, for a given category, *recall* is the ratio of the number of correctly classified documents in the category to the total number of documents from this category. *precision* is the ratio of the number of correctly classified retrieved in the category to the total number of documents assigned to this category. F_1 measure reflects both the precision and the recall (see Table 2).

$$Precision = \frac{C}{C+F}, Recall = \frac{C}{C+M}, F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (1)$$

Table 2. Precision, Recall and F_1. C is the number of correctly classifed documents, F is the number of falsely classified documents, M is the number of documents that are not correctly classified

Note that the recall, precision and F_1 presented in the result part are the mean of the recall, precision and F_1 over all the categories.

Purity for clustering For clustering, we can also define a *precision* and a *recall* (also called *purity*) for each of the clusters. These measures are computed as expressed in the previous section considering that a cluster is assigned to the category that corresponds to the majority of its documents. Note that these measures are only informative but do not really allow us to compare the quality of two methods. Moreover, these measures completely depend on the number of clusters found by the different models. The only way to really know if a clustering has a good quality is to look at the details of the clusters found. Note that the

track allowed participants to freely choose the number of clusters - but the real number of clusters was known for each corpus.

4 Models and Results

4.1 Models submitted

We present here the nine different methods used for clustering and classification of XML documents using *structure only* or *structure and content* information. Figure 3 shows a typology of the different works submitted to the track and the models used. We present here a small summary of the different methods.

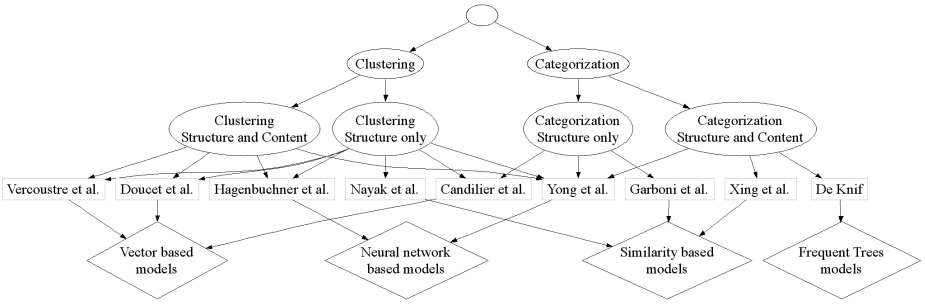

Fig. 3. The typology of the different methods proposed during the XML Mining track. This figures shows the different articles submitted to the track and the tasks concerned by the different models. We also present here the different ideas underlying each article. **Vector based models** are models that first transform XML documents to a vector (or a set of attributes-values) and then use classical vectorial models for clustering or classification. **Similarity based models** define a similarity measure over XML documents - or XML clusters - and then use this measure for clustering or classification. **Neural network based models** are models based on Self Organizing Map or Recursive Neural Networks. **Frequent Trees models** concern models that use the extension of frequent item sets to XML trees.

Vercoustre et al. - Clustering - 2005 The method presented [5] models each XML tree by a set of elements that basically correspond to the different subpaths of the XML tree. Each of these elements is characterized using different criterions: the length of the path considered, the root node label, the number of nodes in the path, etc. The models then projects each document into a vectorial space where each feature of the space corresponds to one of the possible element i.e each document is transformed into the frequential vector of its elements. The clustering problem is then considered as a classical vectorial clustering problem. This model is used for *structure only* clustering but can be used also with

structure and content if the author model the sub-path with some content information. The model is evaluated on the **Movie SO** and **IEEE SO** clustering tasks.

Garboni et al. - Classification - 2005 This model [6] was developped for the classification of *structure only* XML documents. It proposes a similarity measure between an XML document and an XML cluster. Each XML document is transformed into a sequence of its node labels. Then, each cluster is transformed into a set of sequences that are extracted from the set of all the sequences of the cluster using a classical sequential pattern extraction measure. The similarity measure between an XML document and a cluster of documents is thus computed as the longest common subsequence between the sequence of the document and the sequences that characterize the cluster. The model has been tested on the **Movie SO** classification task.

Candilier et al. - Classification and Clustering - 2005 [7] proposed to transform each XML tree into a set of attributes-values. The attributes-values sets are built using different relations between the nodes of the input tree (parent-child relations, next-sibling relations, set of distinct paths,...). Classification and clustering of these attributes-values sets are then made using the *Subspace clustering algorithm (SSC)* and the *C5* classification algorithm. The experiments are made on both the **Movie SO** and **IEEE SO** corpora.

Hagenbuchner et al. - Clustering - 2005 and 2006 The two papers [8] and [9] propose a method for the clustering of *Structure only* and *Structure and Content* XML documents. The method is based on an extension of *Self Organizing Map (SOM)* to *SOM-SD (SOM for Structured Data)* and *Contextual SOM-SD* that can take into account complex structures like labelled trees. This method was used on the **Movie SO**, **IEEE SO** and **IEEE SC** corpus.

Doucet et al. - Clustering - 2006 The article by Doucet et al. [10] introduces a method that transforms an XML document to a vector and then uses a K-means algorithm for the clustering. The transformation that projects a document to a vector takes into account both the structure and the content. The paper also proposes to integrate a *textitude* measure to the document description process that basically measures the ratio between the weight of the structural information and the weight of the content information. This method is used on the **IEEE SC**, **IEEE SO**, **Wikipedia SC** and **Wikipedia SO** corpora.

De Knijf - Categorization - 2006 The model proposed in [11] makes classification of XML document using Frequent attributes Trees. The algorithm is composed of 4 steps:

1. Each class is characterized by a set of Frequent Attributes Trees discovered using the *FAT-miner* algorithm

2. Emerging trees are selected for each category
3. Each document is then transformed into a vector where each component indicates if a particular emerging tree appear into the document
4. Last, a classical classification algorithm is used on these vectors (Binary decision tree)

This model is used on the **Wikipedia SC** corpus

Nayak et al. - Clustering - 2005 and 2006 The papers by Nayak[8] et al. ([12] and [13]) defines a similarity measure between an XML document and a cluster of XML documents. This similarity called CPSim (for Common Path Similarity) is computed during a matching step and is based on different criterions that take into account:

- the number of common nodes between the document and the documents of the cluster considered,
- the number of common nodes paths,
- the order of the nodes of the XML document,
- ...

This measure is then used by an incremental clustering algorithm called PCXSS. This model is applied on the **IEEE SO** and **Wikipedia SO** corpora.

Xing et al. - Categorization - 2006 In this paper ([14]), the auhors proposes to use both a tree edit distance and a Minimum Description Length criterion (MDL) for the classification of *content and structure* XML documents. The method models each class with a normalized regular hedge grammar (NRHG). This grammar is extracted from a set of document using the MDL principle. The distance between a document and a category is then computed by a tree edit distance between the XML tree and the grammar computed for the category. The model is tested on the **IEEE SC** corpus.

Yong et al. - Categorization and Clustering - 2006 This article [15] proposes to categorize XML documents using Graph Neural Networks (GNN). A GNN is an extension of the formalism of Recurrent Neural Network designed for graphs and trees. The intuitive idea behind GNN is that nodes in a graph represent objects or concepts and edges represent their relationships. A *state* vector is attached to each node which collects a representation of the object represented by the node, where the state is naturally specified using the information contained in the neighborhood of a node. The model is used for the **IEEE SO** and **IEEE SC** classification tasks.

[8] the two papers of 2005 and 2006 do not use exactly the same algorithm but are based on the same idea

4.2 Results

We present below the results obtained by the participants for the classification in Tables 3 and 4, and clustering in Tables 5 and 6. The different results are only indicative of the task difficulty and/ or of the potential of a method for a given task. The track was not designed as a benchmark for comparing finely different approaches and methods, but rather as a forum for investigating classification and clustering on structured documents. Participants were free to use any pre-processing for the data and to participate to any task. More results are provided in the full papers of each participant.

Method	Movie S (0 to 3)	IEEE S	Wikipedia S	IEEE SC	Wikipedia SC
Candilier et al.	96.8 % , 96.6 % , 94.7 % , 94.2 %	94.1 %	-	-	-
Garboni et al.	95 %, 94 %, 89 %, 80%	-	-	-	-

Table 3. Classification results using a recall measure during INEX 2005

Method	Movie S (0 to 3)	IEEE S	Wikipedia S	IEEE SC	Wikipedia SC
Yong et al.	- , - , - , -	48%	-	72 %	-
Xing et al.	- , - , - , -	-	-	60%	-
De Knijf	- , - , - , -	-	47%	-	-

Table 4. Classification results using a F1 measure during INEX 2006

Method	Movie S (0 to 3)	IEEE S	Wikipedia S	IEEE SC	Wikipedia SC
Vercoustre et al.	45% , 71 % , 66 % , 53 %	70%	-	-	-
Candilier et al.	78%,-,-,-	-	-	-	-
Hagenbuchner et al. 2005	97%, -,-,-	-	-	-	-
Nayak et al. 2005	60%,60%,59%,59%	65%	-	-	-

Table 5. Clustering results using a purity measure during INEX 2005

These results show some tendencies. Classification on the Movie S (N) datasets, each composed of 11 classes appears quite easy for all the values of parameter N. As said before, in order to create these corpora, artificial classes were generated from a Movie description dataset. The classes have different structures with different degrees of overlapping. Candilier et al. [7] also obtained excellent results for the classification of IEEE document structures into two classes. The content and structure classification task is more complex since the number of classes here is more important (18). For clustering, here too, some participants were able to

Method	Movie S (0 to 3)	IEEE S	Wikipedia S	IEEE SC	Wikipedia SC
Hagenbuchner et al. 2006	-, -,-,-	36 %	13%	-	-
Nayak et al. 2006	-,-,-,-	18%	12.5%	-	-
Doucet et al.	-,-,-,-,	13%	22%	34 %	42 %

Table 6. Clustering results using a purity measure during INEX 2005

recover the hidden classes in the Movie S (N) corpora sometimes with a high accuracy. The structure of the IEEE collections (transactions and non transactions) was also recovered up to 70 % by Vercoustre et al. [5]. For all the other tasks, performance was rather poor. Real sources like IEEE or Wikipedia collections are more challenging to mine than artificially built collections like Movie S. Note that in the literature, a majority of the experiments for evaluating structured data clustering methods have been performed on artificial data created via random tree generators. It is clear from the above experiments that they are not representative of real situations. The SO and SC tasks were investigated as separate tasks, and the respective influence of structure and content cannot be inferred from these results for the SC tasks. This should be investigated in the future.

5 Conclusion

We have presented here the different models and results obtained during the two years of XML Document Mining Track at INEX for both the classification and clustering tasks. The performances obtained show that the structure only task seems quite easy and simple models work very well on this task. This is why we have focused during the second year to the *structure and content* tasks. Concerning the SC tasks, the results obtained can certainly be improved with more sophisticated models. The structure and content tasks on both the IEEE and the Wikipedia corpus will continue next year during INEX 2007.

For INEX 2007, we will also propose the XML Mining track and we will focus on classification/clustering of content+structure XML documents on both the IEEE corpus and the Wikipedia Corpus. The experience of these two years of XML Mining track showed us that we have to define a more strict context and evaluation measures in order to really compare the different methods, and try to encourage participants to submit categorization results that are easier to analyze. For the next year, we will first preprocess all the data - it will allow participant to concentrate on the models - and propose a set of results obtained by classical flat categorization/clustering models in order to have some baseline models as comparison.

Acknowledgments

We are grateful to Remi Gilleron (INRIA, University of Lille), Marc Tommasi (INRIA, University of Lille), Marie Christine Rousset (LIG, Grenoble) and

Nathalie Pernelle (LRI, Orsay) for their help on the definition of the different tasks and the construction of the different corpora. We would like to thank all the participants for their efforts and hard work.

References

1. Maes, F., Denoyer, L., Gallinari, P.: XML structure mapping application to the pascal INEX 2006 XML document mining track. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2006)
2. Gilleron, R., Jousse, F., Tellier, I., Tommasi, M.: XML document transformation with conditional random fields. In: INEX 2006. (2007)
3. Fuhr, N., Gövert, N., Kazai, G., Lalmas, M., eds.: Proceedings of the First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX), Schloss Dagstuhl, Germany, December 9-11, 2002. In Fuhr, N., Gövert, N., Kazai, G., Lalmas, M., eds.: Workshop of the INitiative for the Evaluation of XML Retrieval. (2002)
4. Denoyer, L., Gallinari, P.: The Wikipedia XML Corpus. SIGIR Forum (2006)
5. Vercoustre, A.M., Fegas, M., Gul, S., Lechevallier, Y.: A flexible structured-based representation for XML document mining. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2005) 443–457
6. Garboni, C., Masseglia, F., Trousse, B.: Sequential pattern mining for structure-based XML document classification. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2005) 458–468
7. Candillier, L., Tellier, I., Torre, F.: Transforming XML trees for efficient classification and clustering. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2005) 469–480
8. Hagenbuchner, M., Sperduti, A., Tsoi, A.C., Trentini, F., Scarselli, F., Gori, M.: Clustering XML documents using self-organizing maps for structures. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2005) 481–496
9. Kc, M., Hagenbuchner, M., Tsoi, A., Scarselli, F., Gori, M., Sperduti, A.: XML document mining using contextual self-organizing maps for structures. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2006)
10. Doucet, A., Lehtonen, M.: Unsupervised classification of text-centric XML document collections. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2006)
11. Knijf, J.D.: Fat-cat: Frequent attributes tree based classification. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2006)
12. Tran, T., Nayak, R., Raymond, K.: Clustering XML documents by structural similarity with pcxss. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2006)
13. Nayak, R., Xu, S.: XML documents clustering by structures. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2005) 432–442
14. Xing, G., Xia, Z.: Classifying XML documents based on structure/content similarity. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2006)
15. Yong, S.L., Hagenbuchner, M., Tsoi, A., Scarselli, F., Gori, M.: XML document mining using graph neural network. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2006)