# Report on the XML Mining Track at INEX 2007 Categorization and Clustering of XML Documents

Ludovic Denoyer and Patrick Gallinari
LIP6 - University of Paris 6

**Abstract**

This report concerns the last edition of the XML Mining Track at INEX 2007. A preceding report has been already published concerning the two preceding editions of the track. We present here the new corpus used for this third phase and briefly describe the models and the results obtained by the different participants.

## 1 Introduction

The XML Document Mining track[1] 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 three editions during INEX 2005, 2006 and 2007 and the fourth phase is currently being launched. The two first editions have been summarized in an other report ([1]) and we focus here on the 2007 edition. 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, the track focuses on two generic ML tasks: supervised *classification* and unsupervised *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[2] - has been proposed in the 2006 edition of the track. In the following, we first describe the mining problems addressed at INEX in section and describe the new corpus used in 2007. We then describe the different models submitted by the participants and analyze the results. Finally in section 6 we briefly describe the future directions that will be explored during the 2008 edition of the track.

## 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). In a ML point of view, one has to develop techniques able to handle and to learn using both structural information and content information. Note that most existing ML methods can only deal with only one type of information (either structure or content). Moreover the developed models have to be able to handle a large amount of data because categorization and clustering of XML documents are often made on large XML corpora. In an IR point of view, the models have to identify structural and content similarities between documents in order to organize these documents in clusters (of classes) that can be *thematic* - like in classical categorization and clustering of flat documents - or *structural* -

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

[2] Structure Mapping was defined as learning from examples how to map documents in one or several formats onto a predefined mediated schema

|  | Total | Train | Test |
|---|---|---|---|
| Number of documents | 96,611 | 48,306 | 48,305 |
| Number of internal nodes | ≈ 9 M | 4,505,141 | 4,487,819 |
| Number of distinct words | 446,916<br>(depending on the preprocessing) |  |  |
| Number of words | 33,944,462 | 17,261,996 | 16,682,466 |
| Mean length of the documents | 351.4 | 357.3 | 345.5 |
| Number of distinct tags | ≈ 5,800 |  |  |
| Size of the corpus | ≈ 720 Mbytes | ≈ 360 Mbytes | ≈ 360 Mbytes |

Table 1: Statistics on the corpus

for example, the structure of XML documents can be used to detect certain types of document like *biography* for example.

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, the content alone or both the structure and content information. In the preceding tracks, we have considered *Structure only* tasks and *Structure and Content* tasks. Due to the complexity of the preceding tracks - too many tasks - we have decided in 2007 to focus on a *Structure and Content* task. More precisely, the 2007 track was composed of:

- a *single label classification* task where the goal was to find the single category of each document in a supervised manner

- a *single label clustering* task where the goal was to associate each document to a single cluster

# 3 Corpus

We have used a single corpus for both classification and clustering. The corpus is composed of about 96,000 documents extracted from the *Wikipedia XML Corpus* [2] and split in two parts:

- The training part composed of 50% of the documents,

- The testing part composed of the 50% remaining documents.

The corpus is downloadable from the XML Mining website[3]. The documents are organized in 21 categories that correspond to different *Wikipedia Portals* and basically correspond to thematic categories. We have tried to build a corpus that will be both large enough to correspond to real applications and small enough to able participants with new systems to participate without losing time to develop complex softwares.

The tables 1 and 2 give some statistics concerning the complete corpus.

Note that the categories are not well balanced: some categories are large - *Portal:Law* is composed of about 25% of the documents - while some other are very small - *Portal:Music* is composed of about 0.5 % of the documents. This is interesting because some models tend to learn only on large categories and the corpus will help to measure the capacity of the models to deal with small classes. Moreover, the corpus has been built in order to propose some ambiguous categories like for example *Portal:Pornography* and *Portal:Sexuality* or *Portal:Chistianity* and *Portal:Spirituality*.

[3] *http://xmlmining.lip6.fr*

| Id | Category | Size |
|---|---|---|
| 2112299 | Portal:Law | 24213 |
| 1597184 | Portal:Literature | 16929 |
| 1484914 | Portal:Sports and games | 14595 |
| 1480358 | Portal:Art | 7624 |
| 1886386 | Portal:Physics | 5149 |
| 3091788 | Portal:Christianity | 4671 |
| 2773006 | Portal:Chemistry | 4567 |
| 1685758 | Portal:History | 3246 |
| 3091127 | Portal:Spirituality | 2704 |
| 2914908 | Portal:Sexuality | 2402 |
| 2879927 | Portal:War | 2217 |
| 1507239 | Portal:Aviation | 1217 |
| 2328885 | Portal:Formula One | 1188 |
| 1486363 | Portal:Astronomy | 1105 |
| 1895383 | Portal:Trains | 953 |
| 2314377 | Portal:University | 605 |
| 2635947 | Portal:Comics | 600 |
| 2257163 | Portal:Pornography | 458 |
| 2263642 | Portal:Writing | 412 |
| 474166 | Portal:Music | 401 |

Table 2: Description of the different categories.

### 3.1 Tasks and Evaluation measures

The track was composed of one supervised categorization task and one unsupervised clustering task. Each submission has been blinded evaluated by the organizers on the testing corpus.

#### 3.1.1 Categorization:

For categorization, we have asked the participants to submit one category for each of the documents of the testing set. We have then evaluated how much the categories found by the participants correspond to the real categories of the documents. For each category, we have computed a *recall* that corresponds to the percentage of documents of the category that have been correctly classified. The global classification performances have been measured using a *micro average recall* and a *macro average recall*. The *macro average recall* corresponds to the mean of the recall over all the categories. The *micro average recall* corresponds to the mean of the recall of the categories weighted by the size of the categories. The difference between the *macro recall* and *micro recall* gives an idea of the performances of a model among small classes.

#### 3.1.2 Clustering:

For the clustering task, the participants have submitted a cluster index for each of the documents of the testing set. We have then evaluated if the obtained clustering corresponds to the real categories of the documents. For each submitted cluster, we have computed a *purity* measure that is a recall of the cluster considering that the cluster belongs to the category of the majority of its documents. We have also used a *micro average purity* and a *macro average purity* in order to summarize the performances of the different models over all the documents and all

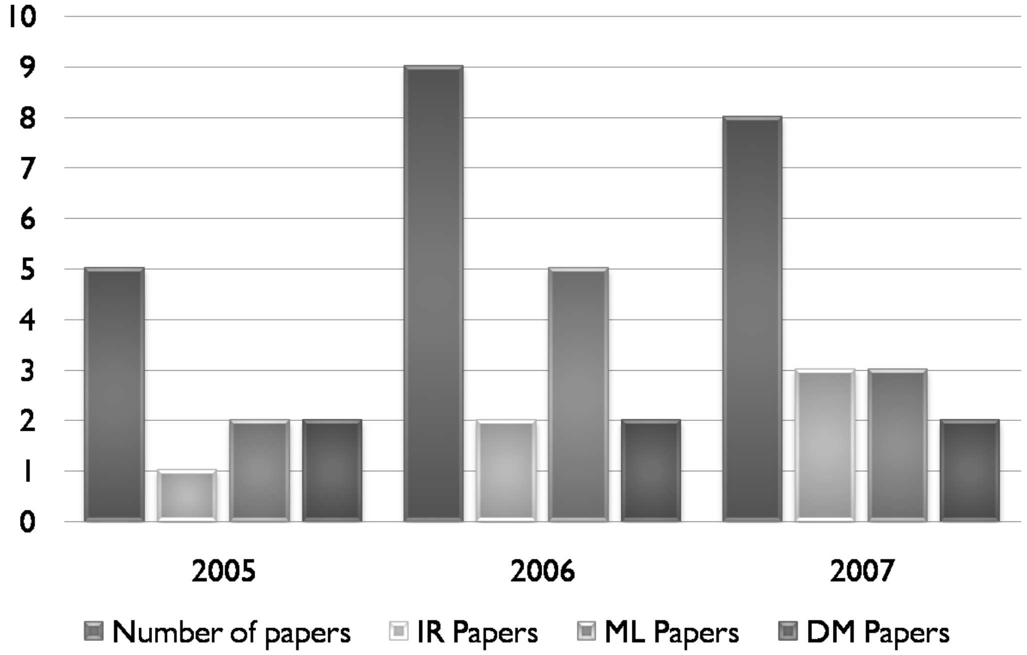

Figure 1: Distribution of the different participants in different research communities by number of submitted papers.

the clusters. Note that the evaluation of clustering is very difficult and it is still an open problem particularly with semi-structured document where clusters can correspond to structural clusters or to thematic clusters. The measure proposed here just gives an idea of how much a model is able to find the 21 categories in an unsupervised way.

## 4 Participants and Submissions

The participants of the XML Mining track mostly come from three different research communities: Machine Learning, Information Retrieval and Data Mining (DM). The figure 1 summarizes the distribution of these three different communities during the three XML Mining tracks[4] and shows that the track clearly bridge the gap between traditionnal research communities.

[4] the classification of the participants in different research communities is only an information given by the authors of the article and only reflect their point of view.

Eight different research teams have participated to the 2007 edition of the track: three for categorization and five for clustering. Note that ML researchers are mostly involved in the Categorization task and DM researchers are mostly involved in the clustering task. We first briefly summarize the different techniques used by the participants this year[5]. The methods used during the two first phases of the track are avaiable in the preceding report [1].

## 4.1 Categorization models

### 4.1.1 Campos et al: [3]

The paper proproses to use probabilistic methods for flat text categorization considering that some structural features of each XML document have been translated to plain text. Different methods for adding the structural information have been tested and two probabilistic models have been tested - Naive Bayes and OR Gate.

### 4.1.2 Murugeshan et al: [4]

The authors propose a vectorial method for the computation of documents similarities. This method is close to TF-IDF and based on the *Negative Category Document Frequency* - called NCD which goal is to reduce the weight of a term according to its distribution over negative categories.

### 4.1.3 Yang et al: [5]

The article proposes the *Structured Link Vector Model* - SLVM - in order to represent the XML documents as vectors. These vectors are then used with a Support Vector Machine for classification. Basically, the SLVM representation is based on the computation of TF-IDF values for each word in each node of the XML documents in order to capture both a content and structural information.

## 4.2 Clustering models

### 4.2.1 Hagenbuchner et al: [6]

The paper proposes a method for the clustering XML documents based on an extension of *Self Organizing Map (SOM)* to *GraphSOM (SOM for for Graph)* - GraphSOM is an extension of SOM-SD used by the same authors during the XML Mining track at INEX 2006.

### 4.2.2 Tran et al. [7]

proposes an algorithm which relies on an incremental and a pairwise clustering approach. In this method the structure of the XML documents is represented as a collection of paths and the content is represented using a latent semantic kernel.

### 4.2.3 Yao et al: [8]

The article proposes to use agglomerative clustering algorithms with XML documents represented with root-based text path descriptors combined with rare patterns.

### 4.2.4 Kutty et al: [9]

The article uses an incremental clustering method for XML clustering. The documents are represented using frequent subtrees discovery.

[5] The full papers are available on the INEX 2007 proceedings

| Article | Micro-average recall | Macro-average recall |
|---|---|---|
| Yang et al. | 87.2 % | 83.9 % |
| Campos et al. | 78.9 % | 76.0 % |
| Murugeshan et al. | 78.0 % | 75.7 % |

Table 3: Classification Results

| Article | Micro-average purity | Macro-average purity | Number of clusters |
|---|---|---|---|
| Yao et al. | 44.4 % | 44.7 % | 5 |
| Yao et al. | 53.4 % | 60.2 % | 10 |
| Hagenbuchner et al. | 25.1 % | 26.6 % | 10 |
| Tran et al | 38.9 % | 40.4 % | 10 |
| Kutty et al. | 25.0 % | 24.9 % | 10 |
| Yao et al. | 53.6 | 59.1 | 15 |
| Yao et al. | 57.9 | 67.2 | 21 |
| Tran et al. | 37.6 | 39.9 | 21 |
| Hagenbuchner et al. | 26.4 | 26.9 | 21 |
| Kutty et al. | 25.0 | 25.0 | 21 |
| Yao et al. | 59.5 | 66.3 | 30 |
| Tran et al. | 38.9 | 40.4 | 30 |

Table 4: Clustering Results

# 5 Official results

We present here the results obtained by the different participants for the track.

## 5.1 Categorization

The categorization results - table 3 - show first that the proposed models are quite good on the classification task. They all have a macro-average score that is lower than their micro-average showing that the small categories are not well classified in comparison to the large ones. As in flat-text categorization, the best method is based on the use of a Support Vector Machine with a structural representation of the documents.

## 5.2 Clustering

For the clustering, we have obtained some results for different numbers of clusters. The results are presented in table 4. The method by Yao et al. is clearly better than the other ones proposed method on this corpus showing that this method is able to find thematic clusters. The clustering evaluation is made here using the thematic categories and this evaluation is not very good. We would obtain more robust results by looking deeply into the clusters but it is too-much time consuming for this track. We will try to have a better clustering evaluation next year.

# 6 Conclusion

We have presented here the different models and results obtained during the XML Document Mining Track at INEX 2007. There results cannot be directly compared to those obtain during 2006 and 2005 but the three years of XML Mining Track give now a good overview of clustering and classification methods for XML documents. The XML Mining website[6] currently refer all the articles and proposed methods during the three years and is now a nice repository of XML Mining methods.

This track will continue one year more in 2008.

## Acknowledgments

We would like to thank all the participants for their efforts and hard work.

## References

[1] Denoyer, L., Gallinari, P.: Report on the xml mining track at inex 2005 and inex 2006: categorization and clustering of xml documents. SIGIR Forum **41**(1) (2007) 79–90

[2] Denoyer, L., Gallinari, P.: The Wikipedia XML Corpus. SIGIR Forum (2006)

[3] de Campos, L.M., Fernandez-Luna, J.M., Huete, J.F., Romero, A.E.: Probabilistic methods for structured document classification at inex'07. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2007)

[4] Murugeshan, M.S., Krishnamurthy, L., Mukherjee, S.: Lakshmi krishnamurthy and saswati mukherjee. an ncd based approach for wikipedia categorization task. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2007)

[5] Yang, J., Zhang, F.: Xml document classification using extended vsm. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2007)

[6] Hagenbuchner, M., Tsoi, A.C., Sperduti, A., Kc, M.: Efficient clustering of structured documents using graph self-organizing maps. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2007)

[7] Tran, T., Nayak, R., Bruza, P.: Document clustering using incremental and pairwise approaches. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2007)

[8] Yao, J., Zerida, N.: Rare patterns to improve path-based clusteringof wikipedia articles. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2007)

[9] Kutty, S., Tran, T., Nayak, R., Li, Y.: Clustering xml documents using closed frequent subtrees- a structure-only based approach. In: Workshop of the INitiative for the Evaluation of XML Retrieval. (2007)

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