XML Document Mining Challenge

Ludovic DENOYER
LIP6 – University of Paris 6
Description of the Challenge

- DELOS (INEX)/PASCAL Challenge
- Bridging the gap between Information Retrieval and Machine Learning

- 2 Steps:
  - 1st Part: June 2005 to November 2005
  - 2nd Part: December 2005 to April 2006 (?)

What do we need to do?

- Need to find models
  - Extending existing models
  - Creating new models

- Tasks
  - Revisit classical tasks
    1. What is categorization of structured documents?
      1. Categorization of whole documents?
      2. Categorization of parts of document (multi-thematic case)?
      3. Categorization of the document in different structure families?
  - Find new interesting task
    - Structure mapping
Goal of the challenge

- Categorization and Clustering of XML
- Creation of baseline corpora
- Evaluation and Comparison of different methods for mining XML corpora handling
  - Structure information only
  - Both, Content and Structure Information
Asumptions of the challenge

1. The structure information is relevant for the categorization and clustering tasks and can be used to detect different « structural families » of documents (different types of documents)

2. The structure can be helpfull in order to detect the topic of a document
Example

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Description of the corpora (for the first part of the challenge)

- 8 different corpora
  - Structure Only:
    - Moviedb (m_db_s) 1 to 4: Collection of heterogeneous document about movies
      - Goal: Discover 11 different sources
    - INEX (inex_s): Collection of XML scientific articles
      - Goal: Discover 2 different sources (articles from « Transactions on ... » journals and other articles)
  - Structure And Content
    - Moviedb (m_db_cs_) 1 and 2
      - Goal: Discover 11 different categories of movies (comedy, action,...)
    - INEX (inex_cs)
      - Goal: Discover 6 different topics of the articles
Methods proposed for the Challenge

- About 15 interested teams
  - 6+1 participants

- 6+1 articles
- 6 structure only methods:
  - 3 for categorization and 4 for clustering

- 1 baseline method which takes into account both structure and content
  - Categorization and clustering
Models Description

- 3 « Families »
  - XML -> Vectors (2 articles)
    - (Clustering and Classification)
  - XML -> DTD (2 articles)
    - (Classification)

- Other Methods
  - SOM CSOM
    - (Clustering)
  - Decision Tree
Evaluation

- Purity and Entropy for clustering
- Recall for categorization

- Comparison between categorization and clustering

- Clustering : Problem of Number of Clusters !!
MovieDB Structure Results

Clustering Micro Purity

- Candillier
- Hagenbuchner
- Nayak
- Vercoustre
- Baseline NB
- Baseline Parent

Graph showing clustering micro purity for different databases.

Databases: m_db_s_0, m_db_s_1, m_db_s_2, m_db_s_3
MovieDB Structure Results

![Bar Chart](image)
INEX Structure Results

![Clustering Results Graph]

- Micro Purity
- Macro Purity

- Nayak
- Vercoustre
- Baseline NB
- Baseline Parent
INEX Structure Results

![Bar Chart](chart.png)

- **Nayak**
- **Vercoustre**
- **Baseline NB**
- **Baseline Parent**
- **Candiller Classification**
- **Xing Classification**

**Y-Axis:** Micro Purity and Macro Purity

**X-Axis:** Various Baselines and Classification Methods
Structure+Content Results

Clustering

Baseline NB
Baseline PA
<table>
<thead>
<tr>
<th>Method</th>
<th>F1 micro</th>
<th>F1 macro</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.59</td>
<td>0.605</td>
</tr>
<tr>
<td>Structure model</td>
<td>0.619</td>
<td>0.622</td>
</tr>
<tr>
<td>SVM TF-IDF</td>
<td>0.534</td>
<td>0.564</td>
</tr>
<tr>
<td>Fisher kernel</td>
<td><strong>0.661</strong></td>
<td><strong>0.668</strong></td>
</tr>
<tr>
<td>Discriminant learning</td>
<td>0.575</td>
<td>0.600</td>
</tr>
</tbody>
</table>
Conclusion

- Using our corpora/tasks:
  - Structure classification
    - Easy
      - Artificial structural families in MovieDB
      - Simple task Inex (2 categories)
  - Structure clustering
  - Content and structure
    - Structure may help detect topics for classification and clustering
- Need feedback from participants
  - Define new benchmarks
- Need real world applications
- A Technical Report to be written
2nd Part of the challenge

- New corpora
- New tasks: Structure mapping ?
- Attract new participants from ML, DB communities
  - The challenge is opened to participants who are not french!
Thanks to

- Mostrare (University of Liile 3 – France)
- Orsay (University of Paris 11 – France)
- Guillaume WISNIEWSKI – LIP6

Who actively participated to the definition of the different tasks and developed some of the datasets