The Role of Different Thesauri Terms and Captions in Automated Subject Classification

Koraljka Golub, PhD student
Knowledge Discovery and Digital Library Research Group,
http://www.it.lth.se/knowledgelib/
Department of Information Technology, Lund University, Sweden

Algorithm

• string-matching
  • looks if strings from the term list exist in the document to be classified
    • if the string is found, the class(es) associated with that string are assigned
    • one class can be designated by many terms, and each time the class is found, the corresponding weight ("1" in our example) is assigned to the class
  • scores for each class are summed up and classes with scores above a cut-off selected as the final ones for that document

Aim

• aim:
  • explore to what degree different types of terms in Ei influence classification performance
  • preferred terms, their synonyms, related, broader, narrower terms and captions
  • in combination with a stemmer and a stop-word list

• purpose:
  • imply which terms with which weights to use in the classification algorithm

Data collection

• a subset of Ei: class 9 (Engineering, General) and its sub-classes
  • 92 sub-classes at 5 hierarchical levels

• a subset of 35166 paper titles, abstracts and intellectually assigned classes from Compendex
  • at least one of the assigned classes belongs to class 9
  • on average, 2.2 classes per document

Number of terms per class

<table>
<thead>
<tr>
<th>TM Angle measurement</th>
<th>UF Angular measurement</th>
<th>BT Spatial variables measurement</th>
<th>RT Micrometers</th>
<th>MC 943.2 Mechanical Variables Measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td>electric @and electronic instruments +942.1, 1: mechanical variables measurements +942.2, 1: physical properties of gases @and liquids @and solids +931.2, 1: amperometric sensors +942.1, 1: sensors @and amperometric measurements +942.1, 1: angle measurement +942.2, 1: angular measurement +942.2, 1: mechanical variables measurement @and angles +943.2, 1: spatial variables measurement +943.2, 1: micrometers +943.2, 1: anisotropy +931.2, 1: magnetic anisotropy +931.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All 8099 932 92 1423 1691 4378 1739
Avg./class 88 10 1 15 18 48 19
## Results

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>BT</th>
<th>Ca</th>
<th>NT</th>
<th>PT</th>
<th>RT</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clas. doc. %</td>
<td>99.4</td>
<td>97.2</td>
<td>28.6</td>
<td>87.3</td>
<td>95.6</td>
<td>99.1</td>
<td>71.3</td>
</tr>
<tr>
<td>Avg. nbr. clas.</td>
<td>28.3</td>
<td>12.8</td>
<td>0.4</td>
<td>2.6</td>
<td>4.2</td>
<td>19.9</td>
<td>1.6</td>
</tr>
<tr>
<td>Macro. P</td>
<td>0.09</td>
<td>0.09</td>
<td>0.42</td>
<td>0.27</td>
<td>0.40</td>
<td>0.10</td>
<td>0.33</td>
</tr>
<tr>
<td>Macro. R</td>
<td>0.72</td>
<td>0.38</td>
<td>0.07</td>
<td>0.14</td>
<td>0.36</td>
<td>0.54</td>
<td>0.16</td>
</tr>
<tr>
<td>Micro. P</td>
<td>0.06</td>
<td>0.06</td>
<td>0.36</td>
<td>0.15</td>
<td>0.20</td>
<td>0.07</td>
<td>0.22</td>
</tr>
<tr>
<td>Micro. R</td>
<td>0.73</td>
<td>0.38</td>
<td>0.06</td>
<td>0.19</td>
<td>0.38</td>
<td>0.59</td>
<td>0.16</td>
</tr>
<tr>
<td>Macro. F1</td>
<td>0.13</td>
<td>0.10</td>
<td>0.08</td>
<td>0.11</td>
<td>0.27</td>
<td>0.13</td>
<td>0.15</td>
</tr>
<tr>
<td>Micro. F1</td>
<td>0.10</td>
<td>0.11</td>
<td>0.10</td>
<td>0.17</td>
<td>0.26</td>
<td>0.12</td>
<td>0.18</td>
</tr>
<tr>
<td>Avg. F1s</td>
<td>0.11</td>
<td>0.11</td>
<td>0.09</td>
<td>0.14</td>
<td>0.26</td>
<td>0.12</td>
<td>0.17</td>
</tr>
</tbody>
</table>

## Conclusions

- the majority of classes is found when using the All term list and stemming:
  - micro-averaged recall is 73%
  - here no weighting or cut-offs were applied, but will be experimented with in the future
- this study implies that all types of terms should be used for a term list in order to achieve best recall, but that higher weights could be given to preferred terms, captions and synonyms, as those yield highest precision

---

Thank you for your attention!