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/knowlib/
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

Ei thesaurus

- *both* a thesaurus *and* a classification scheme

http://engine-e.lub.lu.se/

TM Amperometric sensors
UF Sensors--Amperometric measurements
MC 942.1
...
TM Angle measurement
UF Angular measurement
UF Mechanical variables measurement--Angles
BT Spatial variables measurement
RT Micrometers
MC 943.2
...
TM Anisotropy
NT Magnetic anisotropy
MC 931.2

931.2 Physical Properties of Gases, Liquids and Solids
...
942.1 Electric and Electronic Instruments
...
943.2 Mechanical Variables Measurements

1: electric @and electronic instruments =942.1,
1: mechanical variables measurements =943.2,
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 =943.2,
1: angular measurement =943.2,
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 =913.2

Aim and purpose

- 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

TM Angle measurement
UF Angular measurement
BT Spatial variables measurement
RT Micrometers
MC 943.2
...
TM Anisotropy
NT Magnetic anisotropy
MC 931.2

942.1 Electric and Electronic Instruments
...
943.2 Mechanical Variables Measurements

	All	BT	Ca	NT	PT	RT	ST
Total	8099	932	92	1423	1691	4378	1739
Avg./class	88	10	1	15	18	48	19

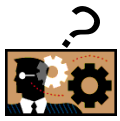
Results

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

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!



&