Filtering Long Lists of Web Objects Using Automatically Generated Facets

Michal Holub

Mária Bieliková

Current state

- digital libraries, e-shops, adverts
- many objects, few filtering options
- filters from simple attributes, not from the attributes' values
- filters do not reflect the actual data
- filters are not personalized

Goals

- **automatically** create filters based on the actual data
- filters from textual values of attributes
- group filters logically to facets
- adapt them to user's needs

Filters construction

- process textual information about objects
- count document frequencies of terms
- determine part-of-speech
- select terms contained in 10% - 50% of objects
- cluster filters

Data

- digital libraries - ACM, DBLP, CEUR-WS
- 17 000 proceedings
- 1 000 journals
- 1 million papers
- 2 million authors

objects → **frequencies** → **terms** → Part-of-Speech → **filters** → **facets**

YAGO, DBpedia → **facts**

Results - publication titles

term	occurrences
proceedings	95 %
workshop	79 %
international	39 %
web	19 %
semantic	15 %
conference	11 %
systems	10 %
information	8 %

Results - paper titles

term	occurrences
based	11 %
data	10 %
using	8 %
semantic	8 %
web	8 %
ontology	5 %
systems	5 %
model	4 %
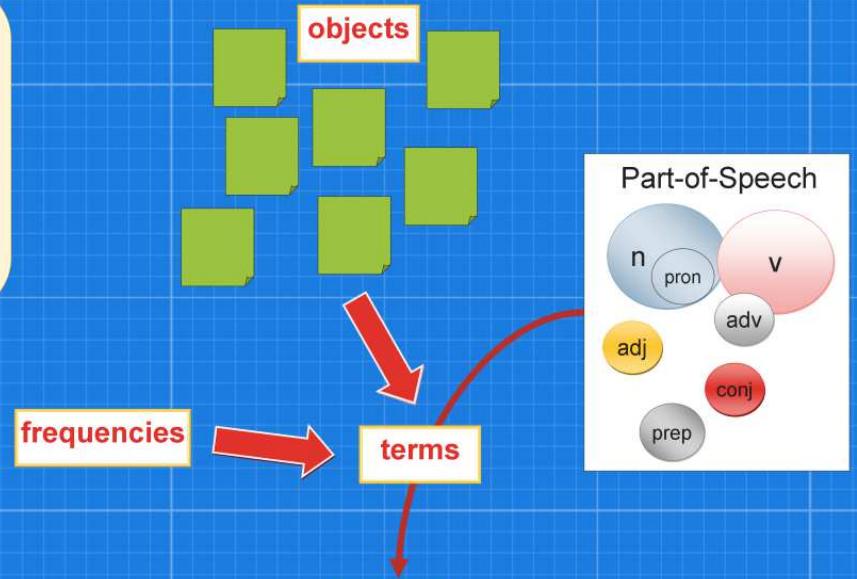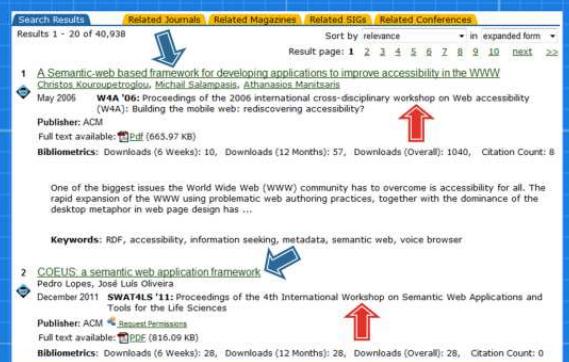
Contributions

- automatic filter generation
- filters reflect the underlying data
- processing textual values of attributes
- clustering of filters

holub@fiit.stuba.sk

Faculty of Informatics and Information Technologies
Slovak University of Technology, Bratislava

PeWe@FIIT
personalized web group