Demonstrating the Semantic GrowBag: Automatically Creating Topic Facets for FacetedDBLP

Jörg Diederich, Wolf-Tilo Balke, and Uwe Thaden
L3S Research Center
Leibniz Universität Hannover, Germany
{diederich|balke|thaden}@l3s.de

ABSTRACT

The FacetedDBLP demonstrator allows to search computer science publications starting from some keyword and shows the result set along with a set of facets, e.g., distinguishing publication years, authors, or conferences. Furthermore, it uses GrowBag graphs, i.e., automatically created categorization systems, to create a topic facet, with which a user can characterize the result set in terms of main research topics and filter it according to certain subtopics.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search Process; H.3.7 [**Digital Libraries**]: Collection

General Terms

Experimentation

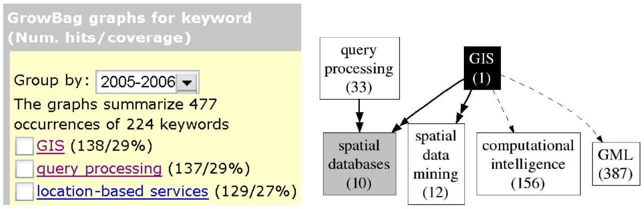
Keywords

Faceted search, digital library, categories, DBLP

1. THE FACETEDDBLP DEMONSTRATOR

Faceted search uses taxonomies to structure the information space that a user can explore, for example, the result set of a keyword search. In our FacetedDBLP demonstrator[1], a search engine for the DBLP[2] collection of computer science documents, in addition to usual facets like authors or publication year we also integrated a specific topic facet that allows to structure the result set of a keyword query according to the keywords associated with the documents. In contrast to pure keyword facets (a result set can easily contain thousands of keywords), topic facets are limited in size because they fold subsumed and related keywords into their associated main keyword. The demonstrator is based on our DBLP++ collection, comprising rich bibliographic information of about 870,000 documents from DBLP continuously updated, of which 95,000 are enhanced by abstracts and a set of about 513.000 keywords. Our topic facet is based on *GrowBag graphs*, community-specific and time-sensitive categorization systems [2, 1] based on the associations between DBLP publications and author keywords. They are created fully automatically based on 'implicit semantics' gained from co-occurrences of keywords with documents and can even be computed for specific periods of time. Even though they obviously cannot achieve the exact quality of manually-crafted categorization systems, they are still very helpful in structuring and assessing the result set of a keyword query.

Let us consider a short example query 'spatial databases' that finds 471 documents associated with 477 occurrences of 224 keywords, which is still too large for manual inspection. Our topic facet groups this result set into 50 GrowBag graphs (for the period 2005–2006), all of which contain a subset of the 224 keywords (cf. left part of Fig. 1, showing an excerpt of the topic facet). The matching GrowBag graphs in the topic facet are ordered by decreasing coverage, i.e., the 'GIS' graph (shown on the right hand side of Fig. 1, excluding for space reasons related keywords such as 'geo-collaboration' or 'crisis management') contains keywords which cover 138=29% of the 477 keyword occurrences.



GrowBag graphs for keyword
(Num. hits/coverage)

Group by: 2005-2006

The graphs summarize 477 occurrences of 224 keywords

□ GIS (138/29%)
□ query processing (137/29%)
□ location-based services (129/27%)

Figure 1: Topic facet / GrowBag graph for 'GIS'

In summary, topic facets can be used for two purposes: (1) to characterize the result set as a whole (The GrowBag graphs for 'GIS', 'query processing', and 'location-based services' cover around 28% of all keyword occurrences found in the result set, which are the 'main topics' from the result set. The remaining graphs starting from 'GPS' cover only 7% or less, hence they touch 'spatial databases' only little) and (2) to focus on a particular topic within the result set. For example, selecting the GrowBag graph for 'GIS' in the topic facet, reduces the result set to 102 documents.

2. REFERENCES

[1] J. Diederich and W.-T. Balke. The Semantic GrowBag Algorithm: Automatically Deriving Categorization Systems. Technical report, L3S Research Center, Leibniz Universität Hannover, Mar. 2007.

[2] J. Diederich, W.-T. Balke, and U. Thaden. The Semantic GrowBag Demonstrator for Automatically Organizing Topic Facets. In *Proc. of the SIGIR Workshop on Faceted Search*, Aug. 2006.

[1] http://dblp.l3s.de
[2] http://dblp.uni-trier.de

Copyright is held by the author/owner(s).
JCDL'07, June 17–22, 2007, Vancouver, British Columbia, Canada.
ACM 978-1-59593-644-8/07/0006.