Semtinel: Interactive Supervision of Automatic Indexing

Kai Eckert, Heiner Stuckenschmidt
Computer Science Institute
University of Mannheim
{kai, heiner}@informatik.uni-mannheim.de

Magnus Pfeffer
University Library
University of Mannheim
pfeffer@bib.uni-mannheim.de

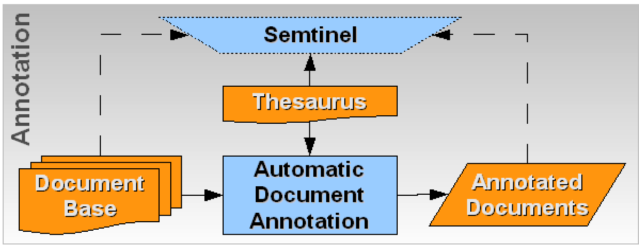
Categories and Subject Descriptors:
H.3.1 [**Content Analysis and Indexing**]: Thesauruses; H.5.2 [**User Interfaces**]: Graphical User Interfaces

General Terms: Algorithms, Measurement

Keywords: Thesaurus-Based Retrieval, Visualization

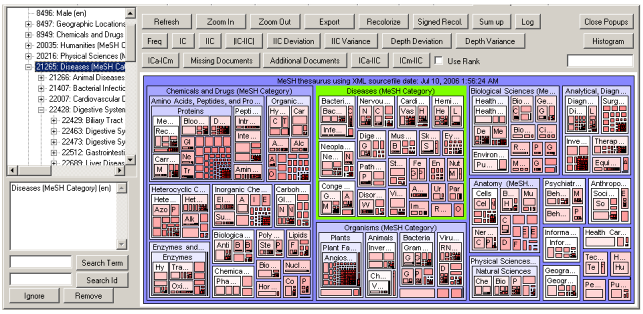
The use of thesaurus-based indexing is a common approach for improving the result of document retrieval. With the growing amount of documents available, manual indexing is no longer a feasible option and statistical methods for automated document indexing are becoming an attractive alternative. But especially in areas where manual indexing could be complemented or replaced by automatic systems, the correctness and completeness of the resulting annotations is very important.

We argue that the quality of automatic indexing not only depends on the involved indexing system, but also on the quality of the thesaurus in regard to its ability to adequately cover the contents to be indexed. A manual verification of all automatically assigned annotations is obviously not a solution and it is questionable, if the verification of random samples would be sufficient to ensure an overall annotation quality that is comparable to manual annotations. We propose the integration of a revision performed by a human expert and supported by Semtinel into the process of automatic document indexing, as shown in the following diagram.



To enable the expert to keep up with the speed of an automatic indexing system, Semtinel provides a highly optimized toolbox with various statistical analysis methods, as well as the possibility to get an in-depth view on concrete annotation results. Beside our current prototype implementation, new sophisticated analysis methods are to be developed that support this approach.

One such method that is currently under development is the *IC Difference Analysis*, a thesaurus evaluation that is based on a combination of statistical measures and appropriate visualization techniques. It supports the detection of potential problems in a thesaurus that affect the automatic indexing and is presented in [1].



The basic idea of this analysis is the identification of problematic concepts by the comparison of their actual information content to the expected information content. The expected information content is based on the concept's position in the thesaurus hierarchy or determinded by a reference set of manually assigned annotations. The results of this analysis are visualized using the treemap algorithm. This visualization makes it easy for the human expert to browse the whole thesaurus and depict problematic concepts (Screenshot above).

Our current prototype supports three different applications of the IC Difference Analysis, various statistics for each concept, direct access to the documents that are annotated by a given concept, highlighting of text positions in the documents that are potentially responsible for the automatic assignment and access to all concepts that were additionally assigned.

With these features, we currently can support the following revisions that might be necessary:

- Deletion and/or merging of rarely used terms
- splitting, extension or restriction of extensively used terms
- review of the thesaurus structure to avoid extensive subclassing
- identification of problematic concepts for the automatic indexer, i.e. concepts that are erroneously assigned due to misleading occurrences in the documents with improper sense

The implementation is tested on different real life thesauri and document sets, including the MeSH thesaurus with Pubmed abstracts as document base. More information and a flash demonstration can be found at http://www.kaiec.org.

References

[1] K. Eckert, H. Stuckenschmidt, and M. Pfeffer. Interactive thesaurus assessment for automatic document annotation. In *Proceedings of The Fourth International Conference on Knowledge Capture (K-CAP 2007), Whistler, Canada*, 2007.

JCDL '08 Pittsburgh, PA, USA