



Text-Based Medical Case Retrieval Using MeSH Ontology

Mario Taschwer

Institute of Information Technology (ITEC), Alpen-Adria Universität Klagenfurt, Austria

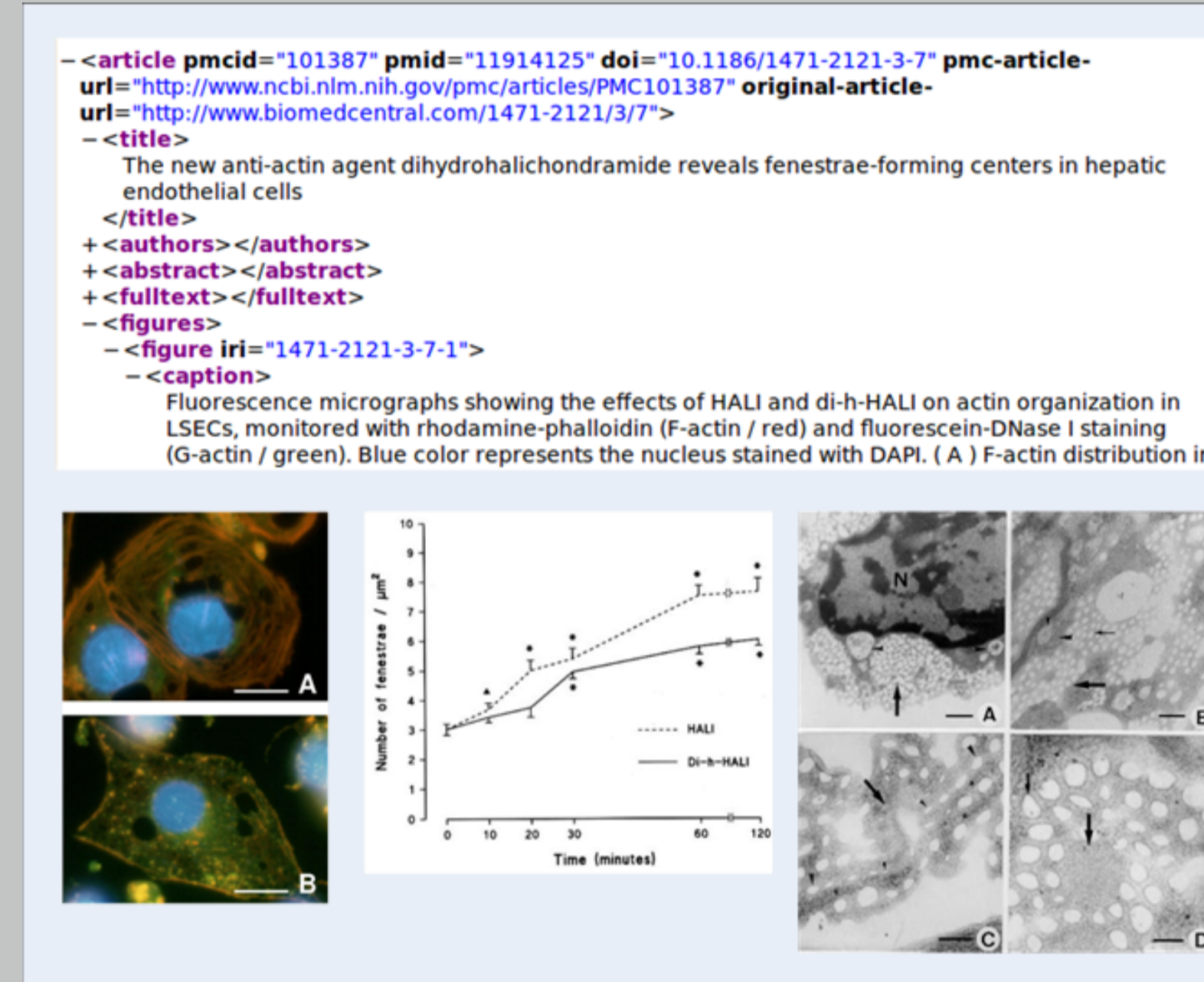


Abstract

Our approach to the ImageCLEF medical case retrieval task consists of text-only retrieval combined with utilizing the Medical Subject Headings (MeSH) ontology. MeSH terms extracted from the query are used for query expansion or query term weighting. MeSH annotations of documents available from PubMed Central are added to the corpus. Retrieval results improve slightly upon full-text retrieval.

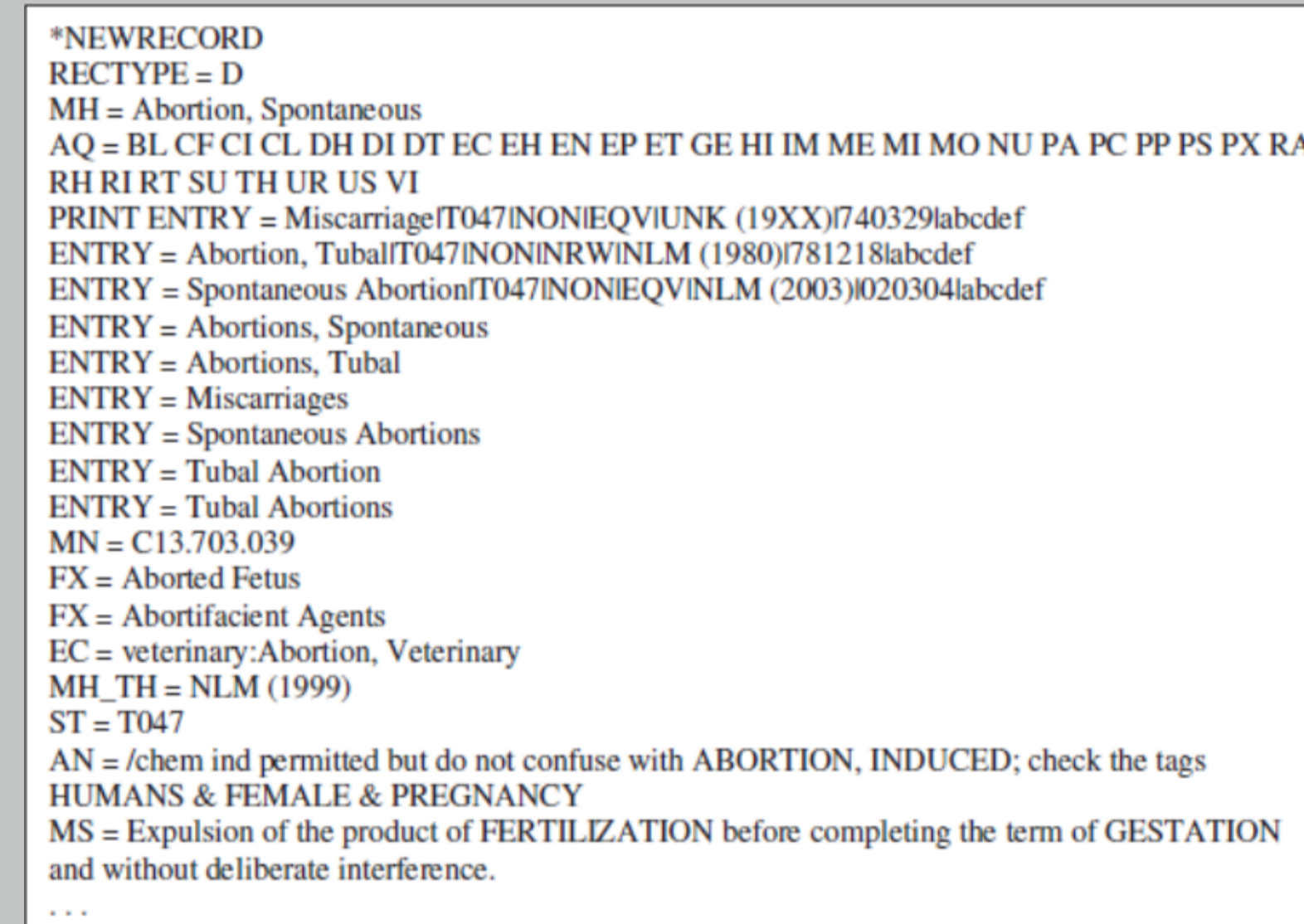
Problem: Medical Case Retrieval (MCR)

- ▶ Given a description of patient symptoms (query), find descriptions of diseases or patients' health records (document corpus) that are relevant as decided by medical experts.
- ▶ The ImageCLEF 2013 medical task contains an instance of this problem: 75,000 biomedical publications (example shown in figure) are to be queried by 35 symptom descriptions consisting of text and diagnostic images.
- ▶ Previous editions of this task showed that text-only retrieval performed roughly four times better than visual-only retrieval, and combinations of text and visual retrieval could not improve over text-only retrieval.
- ▶ **How can text retrieval be improved for MCR?**



Approach: Utilizing Medical Subject Headings (MeSH) Annotations

- ▶ MeSH is a controlled vocabulary commonly used to annotate biomedical publications. The annotations can be used as an external knowledge source to enhance text retrieval.
- ▶ Our approach to the MCR task is to expand a given query by relevant MeSH terms in order to improve the average precision of fulltext retrieval results.
- ▶ This idea is not new, but our approach differs in the way how relevant MeSH terms of a query are identified and which terms are selected for query expansion.
- ▶ *Figure:* Example MeSH record containing primary term (MH) and synonyms (ENTRY).



Details: Retrieval Using MeSH Ontology

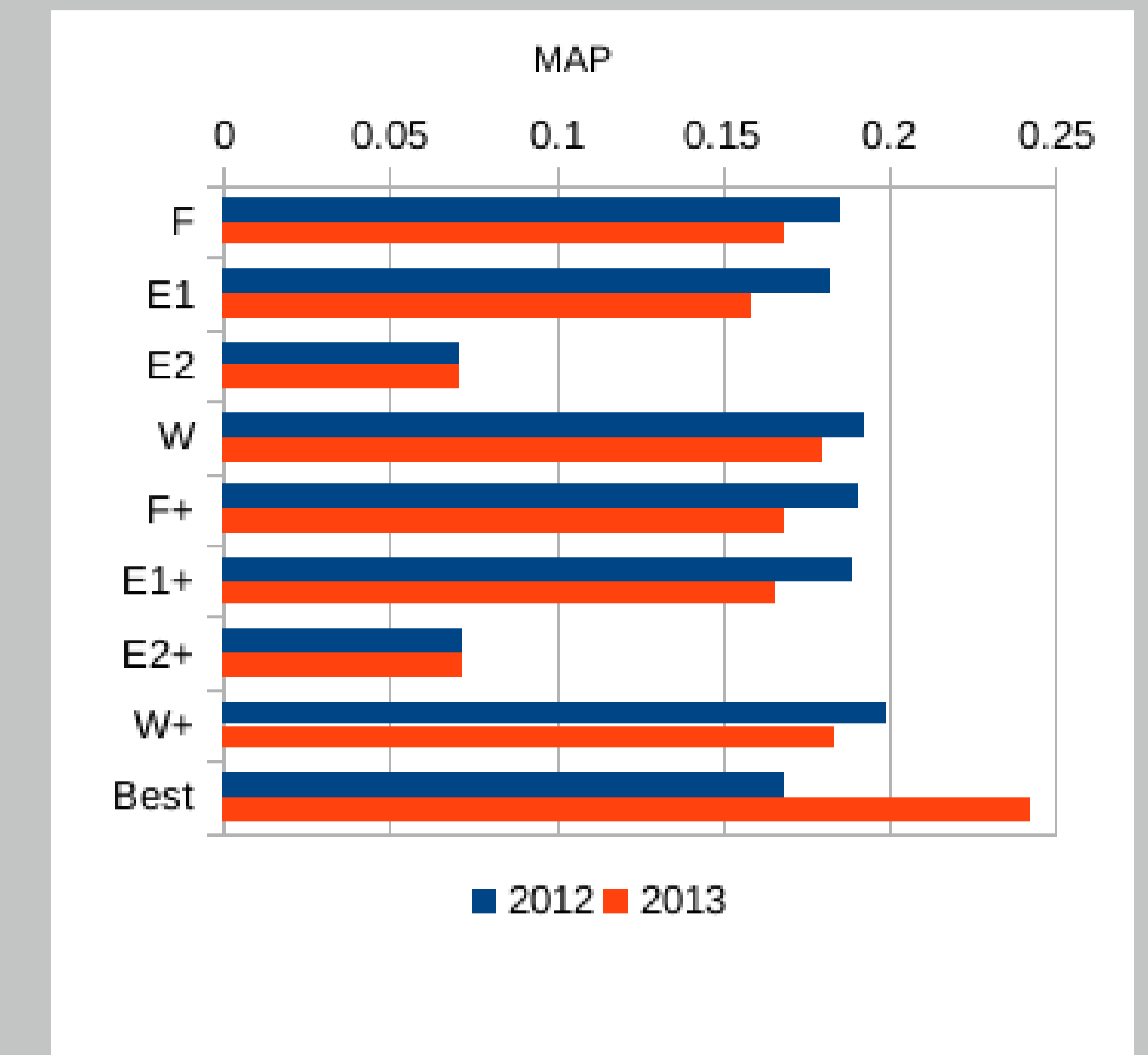
- ▶ We use only 3 out of 16 subtrees of the MeSH ontology comprising 8,911 primary MeSH terms and 64,201 synonyms:
 - ▷ Anatomy [A]
 - ▷ Diseases [C]
 - ▷ Analytical, Diagnostic, Therapeutic Techniques and Equipment [E]
- ▶ A MeSH term is considered relevant for a query if at least 80% of its constituent words occur in the query.
- ▶ Query expansion is performed with all primary MeSH terms where one of its synonyms is relevant for the query.
- ▶ Used Lucene 3.6.2 with default token analyzer to produce two indexes of the document corpus, one with and one without MeSH annotations.
- ▶ Variants of query expansion considered:
 - ▷ Expansion with all synonyms of relevant MeSH terms.
 - ▷ Expansion with primary MeSH terms that are completely contained in the query. This amounts to query term re-weighting.

MCR Results on the ImageCLEF Dataset

Method	ID	2012 dataset		2013 dataset	
		MAP	bpref	MAP	bpref
Fulltext index					
Fulltext baseline	F	0.1856	0.1797	0.1689	0.1731
QE with primary MeSH terms	E1	0.1823	0.1678	0.1581	0.1635
QE with MeSH synonyms	E2	0.0713	0.0829	0.0713	0.1247
QTW of primary MeSH terms	W	0.1926	0.1727	0.1796	0.1882
Fulltext index with MeSH annotations					
Fulltext baseline	F+	0.1905	0.1801	0.1688	0.1720
QE with primary MeSH terms	E1+	0.1887	0.1653	0.1663	0.1634
QE with MeSH synonyms	E2+	0.0721	0.0823	0.0718	0.1258
QTW of primary MeSH terms	W+	0.1991	0.1765	0.1838	0.1911
<i>Best result achieved by other ImageCLEF participants</i>	Best	0.1690 ¹	0.1499 ¹	0.2429	0.2417

QE = query expansion, QTW = query term weighting

¹ Results not corrected; the relevance judgments of the 2012 dataset did not contain any relevant documents for 3 queries.



- ▶ Query expansion with all synonyms of relevant MeSH terms reduced average precision dramatically.
- ▶ Query expansion with primary MeSH terms did not improve average precision w.r.t. fulltext retrieval.
- ▶ However, the query term weighting approach increased MAP by 1.4% to 0.1838 compared to our fulltext baseline. It also improves slightly upon the baseline run of the ImageCLEF 2013 medical task organizers (HES-SO-VS FULLTEXT LUCENE, 0.1791 MAP).
- ▶ All our runs, though, stayed substantially below the best textual MCR run submitted to ImageCLEF 2013 (SNUMedinfo9, 0.2429 MAP).

Conclusion and Further Work

- ▶ Our approach to query expansion using relevant MeSH terms did not improve MCR performance upon fulltext retrieval.
- ▶ However, a variant of query expansion amounting to query term weighting of primary MeSH terms already present in the query could slightly improve retrieval performance.
- ▶ Further work could modify strategies and parameters of the presented approach:
 - ▷ improve identification of relevant MeSH terms,
 - ▷ use other top-level nodes and relations in the MeSH ontology,
 - ▷ further improve query term weights.
- ▶ Moreover, other known query expansion methods could be applied to utilize the MeSH ontology:
 - ▷ pseudo-relevance feedback
 - ▷ MeSH term co-occurrence

Further Information and Contact

- ▶ Working note paper: www-itec.uni-klu.ac.at/~mt/2013/09/imageclef-2013-participation/
- ▶ PhD proposal: www-itec.uni-klu.ac.at/~mt/2013/03/phd-proposal/
- ▶ Contact: mt_at_itec.aau.at

