# Text-Based Medical Case Retrieval Using MeSH Ontology

Mario Taschwer

Alpen-Adria-Universität Klagenfurt, Austria
mt@itec.aau.at

**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.

**Keywords:** medical case retrieval, MeSH ontology

## 1 Introduction

Medical case retrieval (MCR) is the problem of finding descriptions of diseases or patients' health records (document corpus) that are *relevant* for a given description of a patient's symptoms (query), as decided by medical experts. The ImageCLEF medical task 2013 contains an instance of this problem aiming at fully automatic retrieval, where the document corpus contains about 75,000 biomedical publications that should be queried for 36 symptom descriptions consisting of text and diagnostic images [1]. 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 [2].

The question whether and how text retrieval can be improved for medical case retrieval is therefore an interesting research problem, where retrieval performance is measured by *mean average precision* (MAP) or *binary preference* (bpref) [3]. One possibility to enhance text-based medical case retrieval is to utilize external knowledge about the biomedical domain as represented in various medical ontologies [4, 5]. PubMed[1] publications, where the document corpus of the ImageCLEF MCR task is drawn from, are annotated with Medical Subject Headings[2] (MeSH) terms, a controlled vocabulary organized in a tree structure.

Our approach to the ImageCLEF 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 and has been applied with varying success to the MCR problem [4, 6, 7]. Our approach differs in the way how relevant MeSH terms of a query are identified and which terms are selected for query expansion.

[1] http://www.ncbi.nlm.nih.gov/pmc/
[2] http://www.nlm.nih.gov/mesh/

```
*NEWRECORD
RECTYPE = D
MH = Abortion, Spontaneous
AQ = BL CF CI CL DH DI DT EC EH EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA
RH RI RT SU TH UR US VI
PRINT ENTRY = Miscarriage|T047|NON|EQV|UNK (19XX)|740329|abcdef
ENTRY = Abortion, Tubal|T047|NON|NRW|NLM (1980)|781218|abcdef
ENTRY = Spontaneous Abortion|T047|NON|EQV|NLM (2003)|020304|abcdef
ENTRY = Abortions, Spontaneous
ENTRY = Abortions, Tubal
ENTRY = Miscarriages
ENTRY = Spontaneous Abortions
ENTRY = Tubal Abortion
ENTRY = Tubal Abortions
MN = C13.703.039
FX = Aborted Fetus
FX = Abortifacient Agents
EC = veterinary:Abortion, Veterinary
MH_TH = NLM (1999)
ST = T047
AN = /chem ind permitted but do not confuse with ABORTION, INDUCED; check the tags
HUMANS & FEMALE & PREGNANCY
MS = Expulsion of the product of FERTILIZATION before completing the term of GESTATION
and without deliberate interference.
...
```

**Fig. 1.** Example MeSH record [4]. The primary MeSH term is given by the MH field, the ENTRY fields denote synonyms.

## 2 Retrieval Using MeSH Ontology

The MeSH ontology consists of *records* representing the nodes of a tree structure. A record describes a *primary MeSH term* and, among other information, a number of *synonyms* (Figure 1). A parent node in the tree represents a more general term than its child nodes. The child nodes of the root node (let us call them *top-level nodes*) are listed in Table 1. Following the approach of Diaz-Galiano et al. [4], we used only 3 top-level nodes for query expansion (nodes A, C, and E). The 3 selected subtrees contain 8,911 primary MeSH terms and 64,201 synonyms.

To identify relevant MeSH terms of a query, we chose a simple approach: primary MeSH terms and synonyms were stored in a word-based inverted index after removal of hyphens and punctuation characters, and after lower-case filtering. After applying the same preprocessing to query words, they were looked up in the index and marked in the corresponding MeSH terms (primary or synonym). Finally, the ratio of marked to all words of a MeSH term was used to produce a ranked list of primary MeSH terms. By applying a threshold to ratio values, the final list of relevant primary MeSH terms is obtained. For our experiments, we used a threshold of 0.8.

The original query was then expanded with the identified relevant primary MeSH terms and used for fulltext retrieval with one of two document indexes. The first index was generated from the documents' fulltext including titles, abstracts, and figure captions. The second index was created from the same docu-

**Table 1.** Top-level nodes of MeSH tree structure. Only the subtrees represented in bold face were used for query expansion.

| | |
|---|---|
| **Anatomy [A]** | Anthropology, Education, Sociology and Social Phenomena [I] |
| Organisms [B] | Technology, Industry, Agriculture [J] |
| **Diseases [C]** | Humanities [K] |
| Chemicals and Drugs [D] | Information Science [L] |
| **Analytical, Diagnostic, Therapeutic Techniques and Equipment [E]** | Named Groups [M] |
| Psychiatry and Psychology [F] | Health Care [N] |
| Phenomena and Processes [G] | Publication Characteristics [V] |
| Disciplines and Occupations [H] | Geographicals [Z] |

ment text extended with annotated MeSH terms, which had been retrieved from PubMed Central[3] in advance. Indexes were generated using Lucene[4] 3.6.2 with default token analyzer.

We also considered two variants of query expansion: (1) expansion by all synonyms of a relevant primary MeSH term (as in [4]), and (2) expansion by relevant primary MeSH terms only that are already contained in the original query. Variant (2) does not add any new words to the query, but increases the query term weight of added words (by Lucene's standard query processing).

## 3 Results

The retrieval results of our methods are given in Table 2. Only 4 runs were submitted to ImageCLEF 2013 (indicated by the submitted run ID), the others were added after submission. All submitted runs did not improve upon the baseline run of the task organizers (run ID HES-SO-VS_FULLTEXT_LUCENE, 0.1791 MAP), and the submitted query expansion runs stayed below our own fulltext baseline. Query expansion with all synonyms of relevant primary MeSH terms (see variant (1) Section 2) reduced average precision dramatically. However, the query term weighting approach (see variant (2) in Section 2) increased MAP by 1.4% to 0.1838 with respect to our fulltext baseline. All our runs, though, stayed substantially below the best textual MCR run submitted to ImageCLEF 2013 (run ID SNUMedinfo9, 0.2429 MAP).

[3] http://www.ncbi.nlm.nih.gov/pmc/
[4] http://lucene.apache.org/

**Table 2.** Medical case retrieval performance of our methods. Runs without submitted IDs have been added after submission. The relevance judgments of the 2012 dataset contained no relevant documents for 3 queries. These have been removed before evaluation, resulting in 23 queries. QE = query expansion, QTW = query term weighting.

| **Method** | **Run ID** | **2012 dataset** | | **2013 dataset** | |
|---|---|---|---|---|---|
| | Submitted run ID | MAP | bpref | MAP | bpref |
| **Fulltext index** | | | | | |
| Fulltext baseline | FULLTEXT<br>ITEC_FULLTEXT | 0.1856 | 0.1797 | 0.1689 | 0.1731 |
| QE with MeSH terms | MESHEXPAND1<br>ITEC_MESHEXPAND | 0.1823 | 0.1678 | 0.1581 | 0.1635 |
| QE with MeSH synonyms | MESHEXPAND2 | 0.0713 | 0.0829 | 0.0713 | 0.1247 |
| QTW of MeSH terms | MESHEXPAND3 | 0.1926 | 0.1727 | 0.1796 | 0.1882 |
| **Fulltext index with MeSH annotations** | | | | | |
| Fulltext baseline | FULLTEXTPLUS<br>ITEC_FULLPLUS | 0.1905 | **0.1801** | 0.1688 | 0.1720 |
| QE with MeSH terms | MESHEXPAND1PLUS<br>ITEC_FULLPLUSMESH | 0.1887 | 0.1653 | 0.1663 | 0.1634 |
| QE with MeSH synonyms | MESHEXPAND2PLUS | 0.0721 | 0.0823 | 0.0718 | 0.1258 |
| QTW of MeSH terms | MESHEXPAND3PLUS | **0.1991** | 0.1765 | **0.1838** | **0.1911** |

## 4 Conclusion and Further Work

To address the ImageCLEF medical case retrieval (MCR) task, we followed a text-based approach. We identified relevant primary MeSH terms in the query text and performed query expansion against two fulltext indexes (excluding and including MeSH annotations). The method did not improve mean average precision of retrieval performance. 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.

Strategies and parameters of the presented approach could be modified, e.g. identification of relevant MeSH terms, using other top-level nodes and relations in the MeSH ontology, and improving query term weights. Additionally, further work will apply other known query expansion methods using external knowledge, like pseudo-relevance feeback or MeSH term co-occurrence [5].

## References

1. de Herrera, A.G.S., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., Müller, H.: Overview of the ImageCLEF 2013 medical tasks. In: Working notes of CLEF 2013, Valencia, Spain (2013)
2. Müller, H., de Herrera, A.G.S., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., Eggel, I.: Overview of the ImageCLEF 2012 medical image retrieval and classi-

fication tasks. In Forner, P., Karlgren, J., Womser-Hacker, C., eds.: CLEF Online Working Notes. (2012)
3. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. 2nd edn. Addison-Wesley Publishing Company, USA (2011)
4. Díaz-Galiano, M.C., Martín-Valdivia, M., Ureña López, L.A.: Query expansion with a medical ontology to improve a multimodal information retrieval system. Comput. Biol. Med. **39**(4) (April 2009) 396–403
5. Bhogal, J., Macfarlane, A., Smith, P.: A review of ontology based query expansion. Inf. Process. Manag. **43**(4) (July 2007) 866–886
6. Mata, J., Crespo, M., Maña, M.J.: Using MeSH to expand queries in medical image retrieval. In: Proc. MICCAI, Medical Content-Based Retrieval for Clinical Decision Support. MCBR-CDS'11, Springer (2012) 36–46
7. Crespo, M., Mata, J., Maña, M.: Improving image retrieval effectiveness via query expansion using MeSH hierarchical structure. Journal of the American Medical Informatics Association [online] (September 2012) DOI 10.1136/amiajnl-2012-000943.