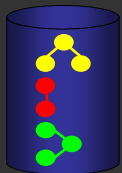


# Ontology-based Retrieval

Edgar Meij

ILPS, Informatics Institute  
University of Amsterdam

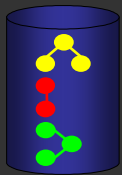


vl·e

**AID 18-10-2005**

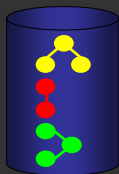
# Focus

- Retrieval using background knowledge
- Query (re)formulation
  - ✓ Interactive
  - ✓ Non-interactive



# Does that work?

- Evaluation done with TREC Genomics
  - ✓ 4.5 million MedLine abstracts (9 Gb)
  - ✓ 50 Questions (topics)
    - defined by biomedical researchers*
  - ✓ Metrics, e.g. precision/recall
  - ✓ Gold standard



# TREC Genomics

- Evaluation done with TREC Genomics
  - ✓ 50 Questions (topics)

...

<111> Provide information about the role of the gene PRNP in the disease Mad Cow Disease.

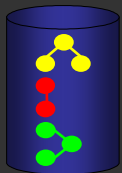
...

<113> Provide information about the role of the gene MMS2 in the disease Cancer.

...

<127> Provide information on the role of the gene alpha7 nicotinic receptor subunit gene in the process of ethanol metabolism.

...



# TREC Genomics

- Evaluation done with TREC Genomics
  - ✓ 4.5 million MedLine abstracts (9 Gb)

PMID- 10605436

TI - Concerning the localization of steroids in centrioles and basal bodies by immunofluorescence.

AB - Specific steroid antibodies, by the immunofluorescence technique, regularly reveal fluorescent centrioles and cili-bearing basal bodies in target and nontarget cells. Although the precise identity of the immunoreactive steroid substance has not yet been established, it seem noteworthy that exogenous steroids can be vitally concentrated by centrioles, perhaps by exchange with steroids already present at this level. This unexpected localization suggest that steroids may affect cell growth and differentiation in some way different from the two-step receptor mechanism.

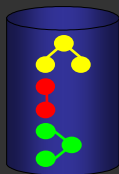
MH - Animals

MH - Lymphocytes/\*cytology

MH - Centrioles/\*ultrastructure

MH - Male

...

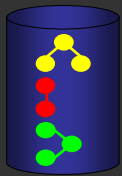


vl-e

# Research questions

## Thesaurus-based query expansion

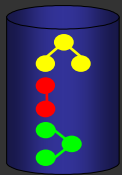
1. Can automatically extracted synonyms help?
2. Can the structure of the MeSH thesaurus help?
3. Can the contents of the MeSH thesaurus help?



# 1. Gene name expansion

✓ Automatically extracted from MedLine corpus

*.. binds hepatocyte nuclear factor 4 (HNF4) and COUP/TF-related proteins. . .*



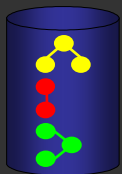
# 1. Gene name expansion

- ✓ Automatically extracted from MedLine corpus

<111> Provide information about the role of the gene PRNP in the disease Mad Cow Disease.



<111>+(PRNP "protein gene" "prp gene" "prion protein gene" ) Mad Cow Disease

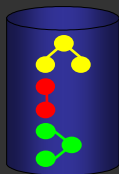


## 2. Structure of MeSH

- ✓ Made an index of MeSH with Lucene
- ✓ Try to identify the MeSH terms that are most related to a topic by querying this index with the query terms
- ✓ These MeSH terms are used to formulate a new query

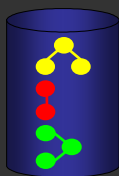
```
<111> MH:D020262 MH:D003561
```

```
MH:D008942 MH:D000971 MH:D015605
```



# 3. Contents of MeSH

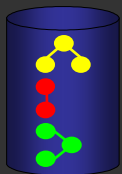
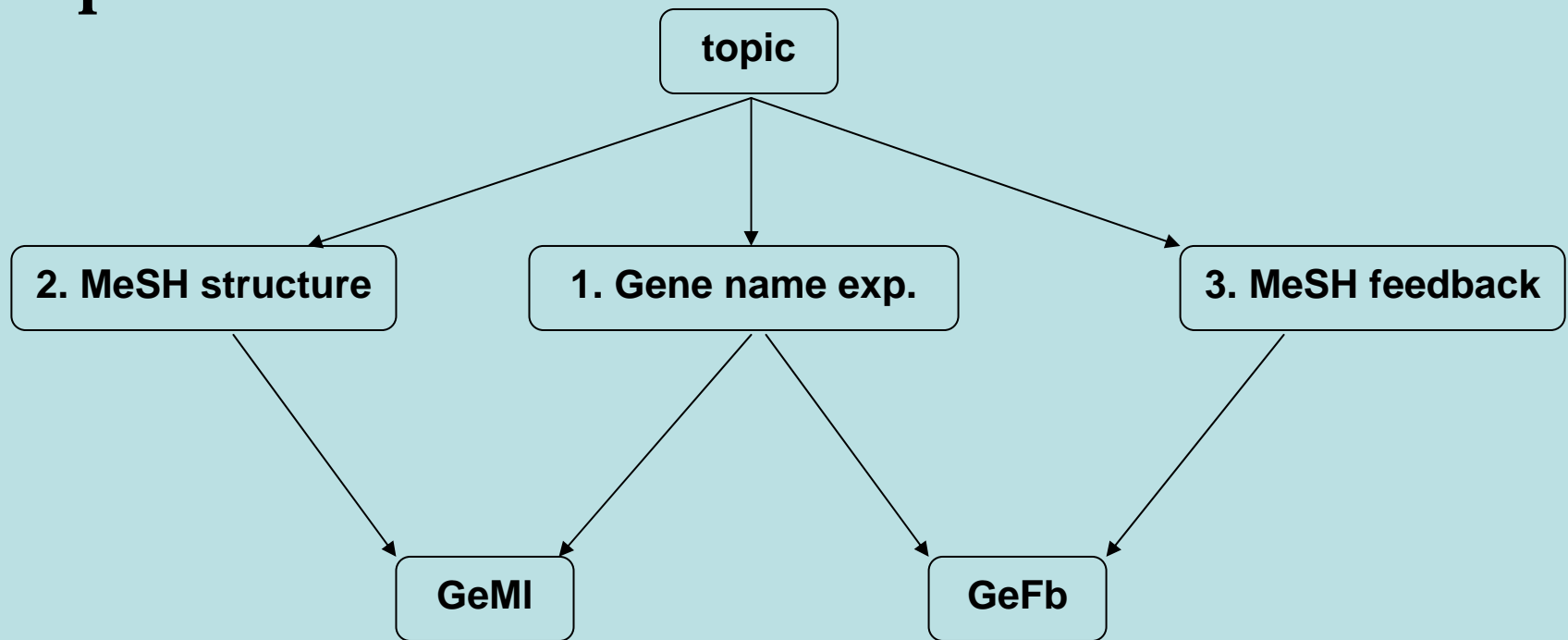
- blind feedback on MeSH terms
  - ✓ Determine MeSH terms of top-ranking documents
  - ✓ These MeSH terms are used to formulate a new query



```
<111> MH:D020262 MH:D003561 MH:D016179  
MH:D003907 MH:D016643 MH:D015608  
MH:D008942 MH:D000971 MH:D015605
```

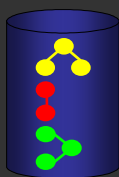
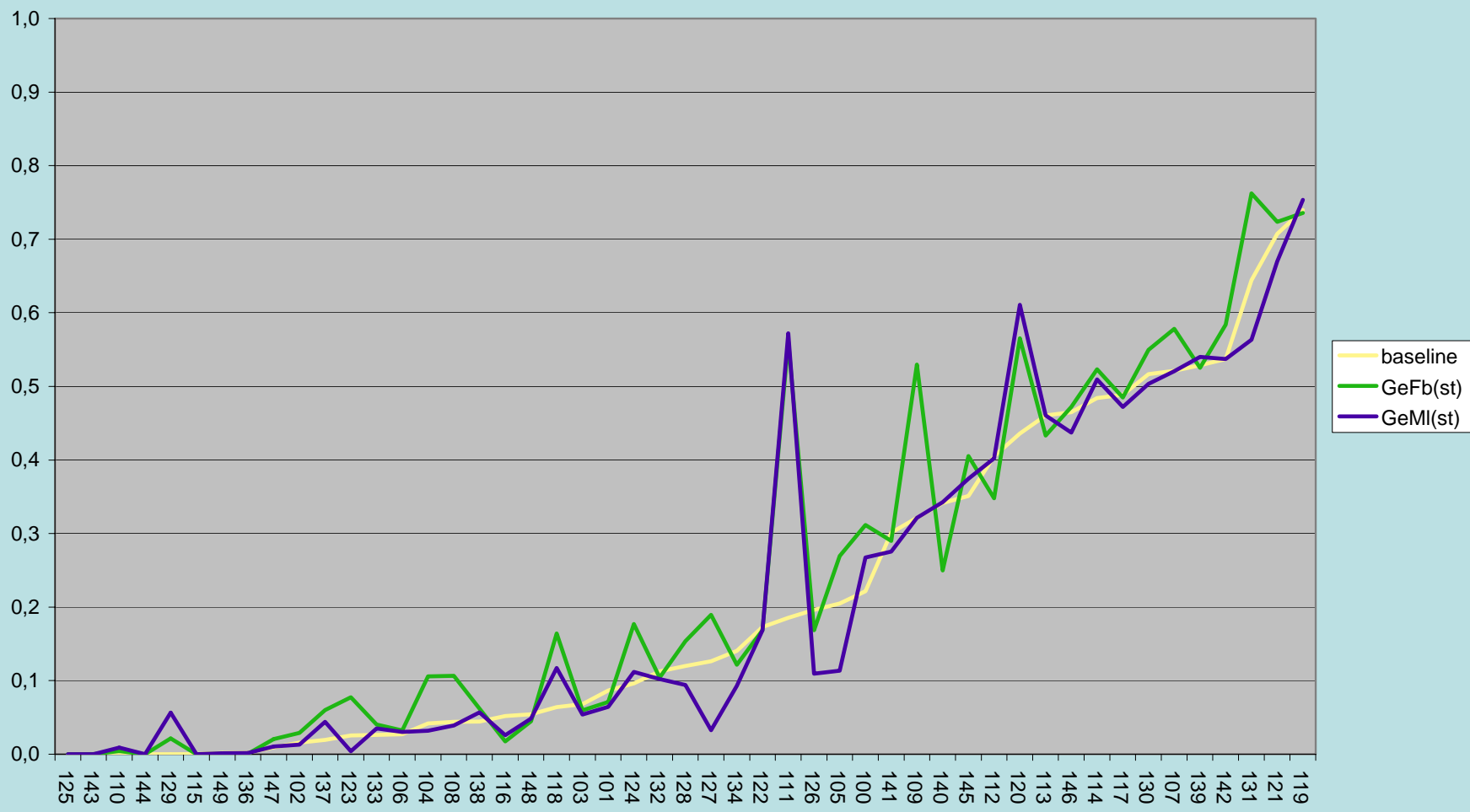
# Training data

- Weighted combination of approaches proved best



# Mean average precision

*Average precision of single topic is mean of the precision scores after each relevant doc retrieved*



vl-e

# Some numbers

- Mean Average Precision

- ✓ Baseline: 0.21

- ✓ GeMl: 0.22

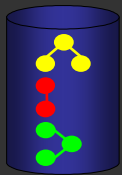
- ✓ GeFb: 0.24

- Mean recall

- ✓ Baseline: 66.6%

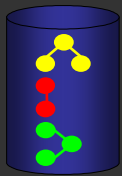
- ✓ GeMl: 66.6%

- ✓ GeFb: 70%



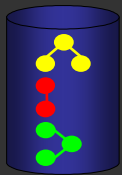
# Concluding

- **Current approach improves:**
  - ✓ Mean recall
  - ✓ Mean average precision
- **But not for all queries/topics**
  - ✓ Currently investigating...

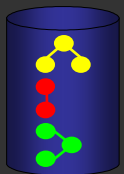


# Future work

- Thesaurus-based retrieval too narrow
- New focus: Semantics-based retrieval
- Semantics from:
  - ✓ Structured documents, e.g.
    - XML/OWL
  - ✓ Semi-structured documents, e.g.
    - Case-based retrieval
  - ✓ ...?



# Thank You!



vl·e

[emeij@science.uva.nl](mailto:emeij@science.uva.nl)