# Biomedical Retrieval: How Can a Thesaurus Help?

Leonie IJzereef[1], Jaap Kamps[1,2], and Maarten de Rijke[1]

[1] Informatics Institute, University of Amsterdam
[2] Archives and Information Studies, Faculty of Humanities, University of Amsterdam

**Abstract.** Searching specialized collections, such as biomedical literature, typically requires intimate knowledge of a specialized terminology. Hence, it can be a disappointing experience: not knowing the right terms to use and being unaware of synonyms or variations in terminology might result in low recall scores. We study the role of a thesaurus in the biomedical information retrieval process. We start by giving a description of vocabulary mismatch problems between natural language queries and relevant documents in biomedical literature search; we provide a detailed case study and observe the impact of vocabulary mismatch problems on retrieval effectiveness. Additionally, we analyze the associated MeSH thesaurus terms used to index the documents in the collection. Based on our observations, we propose a method for exploiting the MeSH thesaurus to improve retrieval effectiveness and, more specifically, to increase recall. We carry out a series of thesaurus-based retrieval experiments that show substantial performance improvements. We conclude with a detailed analysis of the retrieval results.

## 1 Introduction

In the rapidly growing domain of biomedicine, large numbers of papers are published every day. The resulting information overload makes it hard for scientists to stay up-to-date on the latest findings. Therefore, researchers resort to online databases to identify only that part of the literature that is relevant to their own research focus.

To be able to effectively use a bibliographic search engine, a good and detailed understanding of the topic is necessary to choose the right query terms that retrieve all and only the relevant literature. If a scientist lacks domain knowledge when looking for literature on a specific topic, the retrieval process can be a real challenge: not knowing the right terms to use and being unaware of synonyms or variations in terminology might result in low recall scores.

The classic approach to overcome the mismatch between natural language queries and documents relevant to the user's information need is to use controlled vocabularies. Since the early 1960s, controlled vocabularies such as thesauri have been used to improve the retrieval process [13]. A controlled vocabulary dictates what are the preferred terms to use; selected terms are assigned to each publication by a human indexer, and since search requests are also formulated using

the controlled terminology, there is no vocabulary mismatch. This method is often called *manual indexing*, to contrast it with *automatic indexing* that uses (selected terms in) the free-text of publications as indexing terms. The task of a searcher boils down to locating the appropriate controlled terms, a task that turns out to be highly non-trivial in practice [18, 10]. Perhaps professional search intermediaries or experienced users are well equipped to select the right search term, but the effectiveness of average end-users varies greatly.

The effectiveness of controlled vocabularies for information retrieval has been extensively studied in the literature, dating back to the seminal work at Cranfield [3, 4]. Intuitively, it seems obvious that thesauri can overcome vocabulary mismatch problems; however, previous experimental studies have shown that it's highly non-trivial [21]. Below, we discuss some of the relevant research; an encyclopedic overview is beyond the scope of this paper, however. All in all, the literature gives, at best, mixed results on the effectiveness of controlled vocabularies for information retrieval.

Our aim is to better grasp how a thesaurus can help improve the retrieval effectiveness of ordinary, natural language queries. Our strategy is the following. First, we focus on the potential vocabulary gap in biomedical literature retrieval: we provide a detailed study of vocabulary mismatch problems between natural language queries and relevant documents and show its impact on retrieval performance. In addition, we analyze the thesaurus terms manually assigned to the documents in the collection. Based on our observations, we carry out retrieval experiments and discuss how a thesaurus can be used to improve retrieval effectiveness.

The main contributions of this paper are two-fold:

- A detailed analysis of retrieval queries and relevant documents showing that vocabulary mismatch problems have a negative impact on retrieval effectiveness. This analysis together with the analysis of the assigned thesaurus terms suggests that the semantic knowledge provided by a thesaurus can be useful for biomedical retrieval in two ways: its lexical information can be used as a controlled vocabulary to overcome problems with synonymy and lexical variance and its relational knowledge is potentially useful for identifying relevant related terms.
- We demonstrate the use of thesaurus terms assigned to documents for blind and relevance feedback and provide an analysis of the results. We find that using thesaurus-based feedback can improve both precision and recall. However, for the relevance feedback methods to be successful some effort on the part of the user is needed. Nevertheless, for a scientist interest in high recall values, e.g., looking for all relevant literature on a topic, this investment may be worthwhile.

The remainder of this paper is organized as follows. In Section 2 we describe the thesaurus and evaluation data we use. In Section 3 we provide a detailed case studies of the queries, relevant documents and assigned thesaurus terms of four actual retrieval topics. Section 4 presents the results and an in-depth analysis

of some thesaurus-based retrieval experiments. Finally, in Section 5, we draw conclusions and present directions for further research.

### 1.1 Related work

For the open domain, it has been shown that it's hard to use controlled vocabularies due to ambiguity of query words [23]. For more restricted domains, such as biomedicine, there is renewed interest in using controlled vocabularies and semantic knowledge sources due to expanding domains and increasing information needs. In the field of biomedicine, more than 100 different controlled vocabularies (including thesauri and ontologies) are available [16]. Moreover, these vocabularies are already being used for cataloging, classifying, and indexing literature.

Srinivasan [20] compares query expansion based on a statistical thesaurus with expansion via retrieval feedback. She concludes that combining both term selection methods gives the best results, but that the improvement is relatively small in comparison with standard free-text based blind feedback methods.

The term selection method used by French et al. [6] is comparable to the method of Srinivasan: for every word/phrase a list of associated thesaurus terms is computed based on co-occurrences in a training set. However, query augmentation is done by selecting those terms of the list that have been assigned to the greatest number of documents relevant to the query. This gold standard experiment showed that adding one or more suggested terms to the query can potentially improve retrieval effectiveness significantly. Nevertheless, the automatic term selection procedure still has to be defined.

Kostial and Paralic [17] describe a thesaurus-based document boosting procedure (using MeSH and a medical document collection). They combine a basic retrieval procedure with a simple formula based on overlap between thesaurus terms assigned to the query and the documents. The results are promising, but they also circumvent the term selection procedure by assuming that terms relevant to the query are known.

There are also more recent, and more positive, results. Kraaij et al. [12] use thesaurus based relevance feedback for their TREC Genomics 2004 ad hoc task [22] experiments. After a first basic retrieval run, the MeSH thesaurus headings of the top 3 documents are used for a second MeSH retrieval run. They show that a combination of the results of both runs outperforms the basic run, but that the added value of the MeSH run is not convincing. Shallow analysis showed that it only seems to improve precision.

Using a bibliographic database, Savoy [19] evaluates and compares the retrieval effectiveness of various free-text and (human controlled) controlled vocabulary search models. He concludes that the best mean average precision is obtained when both free-text and controlled vocabulary retrieval are combined.

Another feedback technique is described by Kamps [11]: he suggests reranking of the set of initially retrieved documents based on controlled vocabulary terms assigned to documents. He reports a significantly improved retrieval effectiveness based on evaluation on two different domain-specific bibliographic collections, above and beyond the use of standard Rocchio blind feedback.

## 2 Thesaurus and Data Collection

For our detailed case study and experiments we use the National Library of Medicine's MeSH [15]. This choice is based on the features of our data collection: the MEDLINE [14] bibliographic database we use contains citations that are indexed with controlled vocabulary terms from the MeSH thesaurus. Before giving a more detailed description of our data collection, we recall the main features of the MeSH thesaurus.

### 2.1 The MeSH thesaurus

The MeSH thesaurus is used by the National Library of Medicine for indexing biomedical journals and cataloging books, documents and audiovisuals. The core of the MeSH thesaurus is a hierarchical structure that consists of sets of terms naming descriptors. At the top level we find 15 general category headings, such as *Diseases* and *Chemicals and Drugs*. At deeper levels we find more specific headings such as *Brain infarction* (sixth level of *Diseases* branch) or *Dissociative Anesthetics* (ninth level of *Chemicals and Drugs*).

The hierarchy is an eleven-level tree structure that contains over 22,500 headings. Besides the hierarchical structure there are many cross-references that map headings to each other. The main cross-reference fields that can be included in a descriptor's record are the following.

**Scope Note** Provides additional information about the MeSH heading, which can include related MeSH terms.

**See also** Contains related terms that may be of interest.

**Previous Indexing** Contains the MeSH term used before the current descriptor became available.

Together with a descriptor, one or more qualifiers (83 in total) can be used to specify a particular aspect of the descriptor. For example: the qualifier *complications* can be used with diseases to indicate conditions that co-exist or follow.

In addition to the hierarchical structure, there is a separate database with over 139,000 Supplementary Concept Records that consists of chemicals mainly. These supplementary headings are mapped to one or more headings in the main MeSH tree.

It is well known that any thesaurus of the size of MeSH has problems with completeness and consistency [1, 2]. Therefore, it would be useful to analyze MeSH to determine its strengths and weaknesses, and their influence on retrieval. However, for our experiments we take the MeSH thesaurus at 'face value.'

### 2.2 TREC Genomics data collection

To be able to study vocabulary mismatch problems between queries and documents and to look into the potential role of a thesaurus, we use the TREC 2004 Genomics Track ad hoc task [22] data collection. This collection consists of a

selection of 10 years (1994–2003) of MEDLINE citations containing over 4.5M abstracts, 50 retrieval topics and accessory gold-standard data.

Every document in the collection is manually indexed with one or more MeSH headings (from the main tree) and additional qualifiers. For our document collection, this results in 2.6 million unique descriptor-qualifier(s) combinations. In our study, we only take the descriptors into consideration and therefore we ignore the qualifiers. Furthermore, we excluded some frequent but in this context not content-bearing headings such as *Support, Non-U.S. Government* and *Comparative Study*, treeless headings such as *Male* and *Female* and headings with low discriminating values such as *Human* and *Animals*. This leaves us with a total of 21,930 unique headings assigned to the documents in our collection. There is some variation in the number of MeSH headings assigned to each document. Figure 1(Left) shows the distribution of the number of headings assigned to documents. For every document, one or more headings can be marked as main topic of the document. Every heading is placed at one or more nodes in

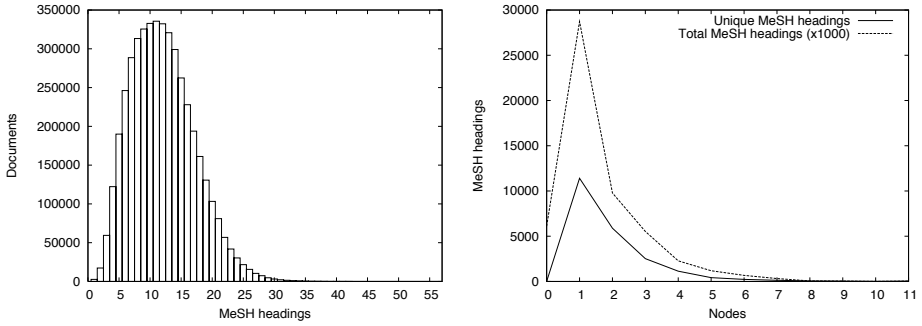


**Fig. 1.** (Left): Number of MeSH headings assigned to the documents. (Right): Nodes per MeSH heading.

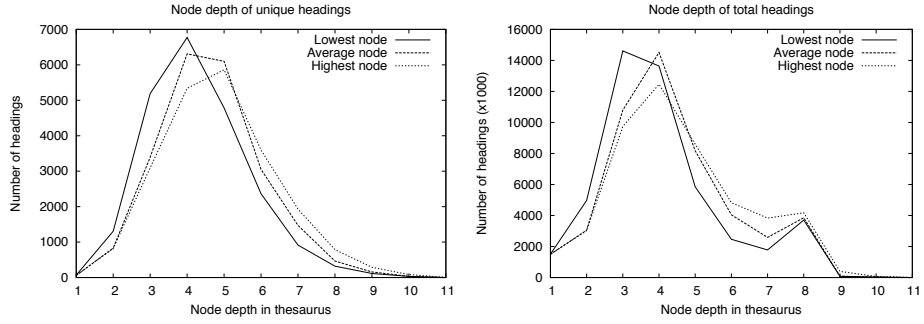


**Fig. 2.** Node depth of MeSH headings. (Left): for every unique heading (types). (Right): for all headings in collection (tokens).

the thesaurus. Figure 1(Right) shows the distribution of unique headings, as well as the distribution of heading occurrences, over the number of nodes at which they are placed in the thesaurus. The depth of the nodes in the thesaurus is related to the specificity of the heading; the deeper the heading is placed in the thesaurus, the more specific it is. Since a heading may be placed on multiple nodes, we can define the depth of a heading by the minimal, maximal, or average depth of its placements in the thesaurus. Figure 2 shows (Left) the distribution of unique MeSH headings and (Right) all heading occurrences over their depth in the thesaurus.

### 2.3 Evaluation topics

We selected four topics for our study of vocabulary gaps between the query (the topic's title) and the document collection. The selection is based on the outcome of a retrieval run (with the topic title as the query) with a baseline vector space-based retrieval system. Many other, possibly better performing approaches could have been chosen here [8]. For the TREC 2004 Genomics track ad hoc task systems using stemming and feedback methods turned out to be the most effective. Systems attempting to map controlled vocabulary terms did not fare as well. Given that we are focusing on the role of the thesaurus in the retrieval process, we decided to use a rather basic retrieval system. With a mean average precision (MAP) score of 0.1716 over all 50 topics, our retrieval score is somewhat lower than the mean MAP of the TREC Genomics 2004 ad hoc task participants. However, recall that we only use the short topic statement of the title field, whereas most other participants use the given additional information about the information need too.

We selected one well performing topic (Topic 9 requesting "mutY"), one average performing topic (Topic 21 asking for "Role of p63 and p73 in relation to DNA damage") and two poorly performing topics (Topic 1 and 14 targeting "Ferroportin-1 in humans" and "Expression or Regulation of TGFB in HNSCC cancers," respectively). The four selected topics can be found in Table 1; we include the MAP and recall at 1,000 documents for the title only-based run.

## 3 Case Studies

In this section, we compare the vocabulary used in natural language queries, the textual content of relevant documents and the MeSH headings assigned to these documents.

### 3.1 Queries and relevant documents

For every topic we take the topic title as our query. We realize that these queries might not perfectly reflect the information need, but since these titles have been formulated by real biologists, we assume that they closely approximate genuine

search queries. In this section, we compare the queries and the textual content of the relevant documents.

As can be seen in Table 1, the topic with the shortest title (topic 1) achieves the highest mean average precision of the four. In the relevant documents for this query we find that the query word *mutY* occurs in almost every relevant document. As that there are only 168 documents in the corpus that contain *mutY*, this single word query gives very good results. However, if a less frequent synonym such as *hMYH* had been chosen as keyword, scores would have decreased dramatically. An example of this can be seen in topic 3: The main keyword *ferroportin-1* has many synonyms (e.g. *IREG1* and *SLC11A3*) and these can all be found in the relevant documents. By using only *ferroportin-1* together with the very frequently occurring term *human*, only a single relevant document is retrieved. A similar problem occurs in topic 4. Here, two acronyms are used: *TGFB* for *Transforming Growth Factor beta* and *HNSCC* for *Head and Neck*

**Table 1.** Selected topics with mean average precision and number of retrieved relevant documents (max. 1000 docs retrieved). For the topics 1, ..., 4 listed below, the original TREC topic IDs are 9, 21, 1, and 14, respectively.

| **Topic** | | **MAP**<br>**rel_ret** |
|---|---|---|
| 1. Title: | mutY | |
| Need: | Find articles about the function of mutY in humans | 0.8676 |
| Context: | mutY is particularly challenging, because it is also known as hMYH. This is further complicated by the fact that myoglobin genes are also typically located in search results. | 113/115 |
| 2. Title: | Role of p63 and p73 in relation to DNA damage | |
| Need: | Do p63 and p73 cause cell cycle arrest or apoptosis related to DNA damage? | 0.1910<br>40/80 |
| Context: | DNA damage may cause cell cycle arrest or apoptosis. p63 and p73 may play a role in mediating these sequelae of DNA damage. | |
| 3. Title: | Ferroportin-1 in humans | |
| Need: | Find articles about Ferroportin-1, an iron transporter, in humans. | 0.0000<br>1/79 |
| Context: | Ferroportin1 (also known as SLC40A1; Ferroportin 1; FPN1; HFE4; IREG1; Iron regulated gene 1; Iron-regulated transporter 1; MTP1; SLC11A3; and Solute carrier family 11 (proton-coupled divalent metal ion transporters), member 3) may play a role in iron transport. | |
| 4. Title: | Expression or Regulation of TGFB in HNSCC cancers | |
| Need: | Documents regarding TGFB expression or regulation in HNSCC cancers | 0.0000<br>0/21 |
| Context: | The laboratory wants to identify components of the TGFB signaling pathway in HNSCC, and determine new targets to study HNSCC. | |

*Squamous Cell Carcinoma*. Since both terms occur only as spelled out terms, this effects the retrieval score dramatically.

However, choosing 'wrong' synonyms or acronyms is not the only reason for poor retrieval scores. For all four topics we see that besides the keywords (or their synonyms/acronyms) one or more semantically related words occur frequently in the relevant documents. As we can see in the description of topic 3, *iron transport* plays a central role in this topic. If we look at the most frequently occurring words/phrases in the relevant documents, we find many closely related terms that are not directly expressed in the query. Some examples: *iron overload*, *iron metabolism*, *transferrin*, *iron deficiency*, *ferritin*, *iron homeostasis*. This is also the case for topic 2, where we see various terms that are closely related to the query, such as *p53*, *cell death*, *tumors*, *transcription* and *transactivation*. Now, it comes as no surprise that we find synonyms and semantically related terms of the keywords in the text of the relevant documents. The key question is how to identify these important terms and how to use them to improve retrieval performance. We hope that detailed analyses such as done in this paper will give us answers to these questions.

### 3.2 Relevant documents and thesaurus terms

All documents in our collection are indexed with MeSH headings, hence we can use this meta-data in our retrieval process. To gain more insight in the role these thesaurus terms and the thesaurus itself could play in the retrieval process, we studied the relation between textual content of the documents and MeSH headings. Before describing this, we show how the MeSH headings assigned to the documents are related to each other.

For all four topics, we created frequency lists for MeSH headings assigned to the relevant documents. Next, we took all MeSH headings that occur in at least 10% of the relevant documents for a topic. This resulted in 19 to 27 MeSH headings per topic. For every heading on the list, we identified its relations with other headings based on the thesaurus.

Among all relations present in the thesaurus, we focus only on *direct* relations between two headings. These come in two kinds: hierarchical parent-child relations and cross-referential relations. Cross-references are relations to headings mentioned in the *Scope*, *Previous Indexing* or *See also* field of a descriptor's record.

For three of the four topics we find many direct relations (approximately 20) between the frequent MeSH headings for that topic. These direct relations are both hierarchical and cross-reference relations. The cross-reference relations are most often not bidirectional: a heading such as *DNA* is often referred to in the *Scope Note* or *Previous Indexing* field of other headings, but only has five *See Also* references itself. If we look at the information need of topic 2, for example, we find that it can be divided into three aspects: Certain proteins, DNA damage and the relation between these two. These three aspects can be seen in the relations between the MeSH headings. One group of seven related headings is focused on the type of proteins and genes involved (e.g., *DNA-binding*

*proteins* and *Tumor suppressor genes*). Another small group of headings contains relations between DNA damage related headings (e.g., *Mutation* and *Apoptosis*). The last group of nine related headings is related to interactions and processes (e.g., *Gene expression regulation*, *Trans-activation* and *Genetic transcription*).

Topic 4, however, shows a different pattern than the other three topics: although most of its 19 main MeSH terms seem to be related, only a few direct relations can be found based on the thesaurus. For example, *Transforming Growth Factor Beta* and *TGFB receptors* do not have a direct relation in the thesaurus. The same holds for *Head and Neck Neoplasms* and *Squamous Cell Carcinoma*. This can either mean that the thesaurus is not really consistent when it comes to cross-references or these relations are relatively 'new' or quite uncommon, and hence they do not appear in the thesaurus.

All four topics express a quite general information need that does not ask for very specific characteristics likely to be found in only one or a few articles. For these general topics, both the text of the relevant documents and as well as their assigned MeSH-headings have a sufficient level of specificity. This is confirmed when looking at the topics: if we compare frequently occurring words/phrases in the text with frequently occurring MeSH headings, we find a clear relation between text and MeSH headings for all four topics. Almost all frequently occurring nouns and compounds in the text are lexical variants or synonyms of one or more frequent MeSH headings. For example, *IREG1*, *ferroportin-1*, and *SLC11A3* all refer to the MeSH heading *metal transporting protein 1*.

For nouns or phrases that are instances of a heading that is on the Supplementary Concept Headings list, which are not used for indexing documents, we see something interesting. In most cases we find one of the frequently assigned headings in the *Heading Mapped to* or *Previous Indexing* field of the Supplementary Concept heading: for *MutY*, we find the two most frequent headings for topic 1, *DNA Glycosylases* and *N-Glycosyl Hydrolases*, in the two mentioned fields of the Supplementary Concept record of heading *MutY adenine glycosylase*.

The last issue to discuss here is the role of the MeSH headings marked as main topic of a document. For topic 1 there are two MeSH headings (*DNA Glycosylases* and *DNA Repair*) that occur in respectively 87 and 43 of the 115 documents as main topic. For the other three topics the main focus is less clear. If we look at the headings marked as main topic with respect to the relations with the other frequently occurring headings, we see that they are not necessarily headings that have many relations or headings that are at the 'center' of a group of related headings.

### 3.3 Queries and thesaurus terms

Besides examining the relation between the text and the MeSH headings, we can study the relations between the queries and the relevant documents. As said before, we use the title of the topics as our search query, assuming that most scientists will start their search process by entering just a few keywords. When we manually identify the MeSH headings that are most closely related to the keywords of our queries, we find that many of these headings are part of the

Supplementary Concept Headings list. Note that the documents have only the preferred MeSH terms (i.e., descriptors) assigned to them. Non-preferred terms such as synonyms (i.e., non-descriptors) can be found in the Supplementary Concepts list. Again we find that in most cases the *Heading Mapped to* and the *Previous Indexing* field refer to MeSH headings frequently used to index the relevant documents. All other content-bearing keywords used in the four queries are instances of MeSH headings that occur frequently in the relevant documents.

### 3.4 Summary of our observations

We conclude this section by summarizing our main observations:

- Low or average retrieval scores are likely to be caused by vocabulary mismatch problems between the query and the relevant documents. This vocabulary gap is often caused by using low frequent synonyms or related terms as keywords.
- MeSH headings that are frequently assigned to the relevant documents of a topic are likely to be directly related to each other in the thesaurus; these relations can either be hierarchical or cross-referential.
- MeSH headings tend to have the same specificity as the frequently occurring words/phrases in the titles and abstracts of the relevant documents.
- Query keywords and frequently occurring words/phrases in the title and abstract of relevant documents can often be mapped to headings on the Supplementary Concept Headings list. The *Heading Mapped to* and the *Previous Indexing* fields on the record of these headings often refer to MeSH headings that occur frequently in the relevant documents.
- MeSH headings frequently marked as main topic that are assigned to the relevant documents do not necessarily play a central role in the information need of the topic.

These observations suggest that the semantic knowledge provided by a thesaurus can be useful for biomedical retrieval in two ways: its lexical information can be used as a controlled vocabulary to overcome problems with synonymy and lexical variance, and its relational knowledge is potentially useful for identifying relevant related terms.

## 4 Retrieval Experiments

In the previous section we provided a detailed comparison of queries, textual content and MeSH headings assigned to the relevant documents of our four TREC Genomics topics. In this section we take a closer look at the potential of a thesaurus for biomedical retrieval; to this end, we carry out a number of retrieval experiments. Besides comparing precision and recall scores averaged over all 50 TREC Genomics 2004 topics, we zoom in on the four selected topics to gain further insight in the retrieval features that cause retrieval (in)effectiveness.

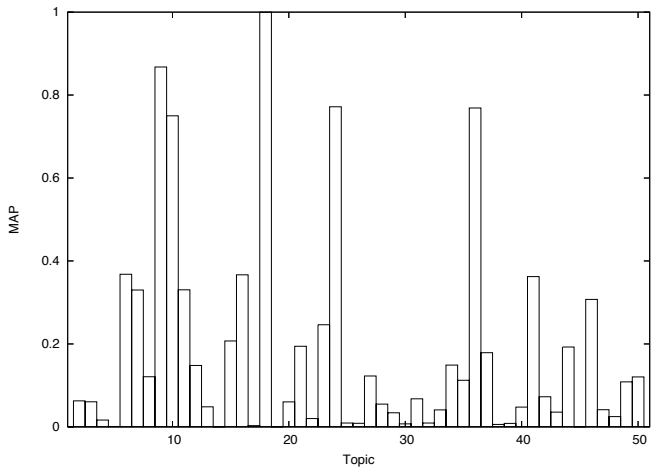


**Fig. 3.** Average precision per topic for the baseline run.

### 4.1 Baseline results

Our baseline run is based on the TREC Genomics 2004 ad hoc task. We use the topic titles as queries and the title, abstract and MeSH-heading fields as retrieval fields. Documents and queries are not stemmed, but stop-words are removed. Our vector space-based retrieval system achieves a mean average precision (MAP) of 0.1716 measured over a maximum of 1000 retrieved documents per topic. Of the total of 8268 relevant documents, 2762 were retrieved.

Figure 3 shows the average precision per topic for our baseline run. As can be seen, there is a huge variety in average precision per topic. This same variety can be found in the average scores per topic of the TREC Genomics participants [8].

### 4.2 Thesaurus-based experiments

In our experiments, we will try to reverse engineer the role a thesaurus can play to improve retrieval using a natural language query. Recall from our analysis above that the relevant documents typically have closely related MeSH terms assigned to them. Hence, these MeSH terms provide useful retrieval cues. So if we select MeSH terms frequently assigned to relevant documents, we can use them to improve retrieval effectiveness.

In reality the set of relevant documents is unknown. So how should we select the relevant MeSH terms? We could ask the user to select the relevant documents in the set of initially retrieved documents. That is, we could use relevance feedback to obtain a set of relevant documents and, again, select the most frequently assigned MeSH terms. Finally, since relevance feedback still requires interaction

with a user, we could simply assume that the first few, initially retrieved documents are relevant. That is, we could use pseudo-relevance feedback to obtain a set of pseudo-relevant documents and, again, select the most frequently assigned MeSH terms.

**Methods** We carried out several feedback experiments. Based on the output of the baseline run, we used the following documents as input for the feedback algorithm:

1. First 10 retrieved documents (this amounts to blind feedback)
2. All relevant documents within the first 10 documents
3. Relevant documents within the first 100 documents, with a maximum of 10
4. 10 random chosen relevant documents

In a real life retrieval situation, information about relevance can be provided by real users: they can be asked to judge initial retrieval results on their relevance, and based on this selection a new retrieval run can be done. Since we do not have access to real users to give this feedback, we simulate them by using the gold-standard data of the TREC Genomics ad hoc task for feedback.

Our first retrieval method does not involve this feedback and works with completely blind feedback. For the second and third method the first 10 and 100 retrieved documents were compared with the gold-standard data. The last method is completely artificial: to be able to get an idea of the potential of thesaurus-based relevance feedback, we choose ten relevant documents from the gold-standard collection.

For all document sets selected for feedback, we created frequency lists for MeSH headings assigned to the documents. Experiments showed that selecting the 35 headings that are most frequently assigned to the selection was optimal for feedback purposes.[3]

The lists of selected MeSH headings were used as queries for a new retrieval run on the MeSH heading fields of the collection. Combining this run with the baseline run resulted in a new ranked list of retrieved documents. Experiments showed that using the CombMNZ method [5] for combining both runs with a relative weight of 0.9 on the baseline run gives the best results. Evaluation of the results is based on the TREC Genomics 2004 ad hoc task gold-standard data.

**Results** Table 2 shows the results of our feedback experiments. All four feedback runs show a significant improvement in MAP compared to the baseline run.[4] The most important reason for this improvement is the large increase in recall: for all four feedback runs, over a 100 more relevant documents are retrieved.

[3] MeSH headings occurring very frequently (more than 100,000 times) or infrequently (less than 5 times) in the total collection were not taken into consideration.

[4] For all runs, significance was proved with over 98% confidence. To determine statistical significance we used the bootstrapping method, a non-parametric inference test that has previously been applied to retrieval evaluation by, e.g., Wilkinson [24].

**Table 2.** Retrieval results based on a maximum of 1000 retrieved documents

| Run | MAP | Precision@30 | Recall@1000 |
|---|---|---|---|
| 0 Baseline run | 0.1716 | 0.3220 | 2762/8268 |
| 1 Baseline + blind feedback (top 10) | 0.1801 | 0.3127 | 2896/8268 |
| 2 Baseline + relevance feedback (top 10) | 0.1876 | 0.3267 | 2888/8268 |
| 3 Baseline + relevance feedback (top 100) | 0.1996 | 0.3667 | 2933/8268 |
| 4 Baseline + 10 relevant docs | 0.2011 | 0.3453 | 2971/8268 |

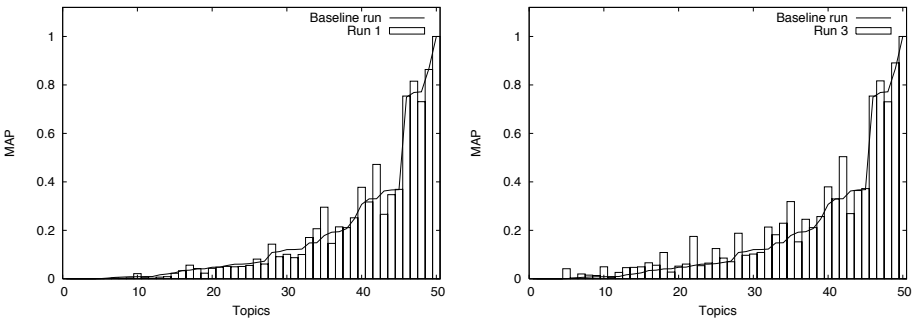


**Fig. 4.** (Left): Average precision scores for the baseline run vs. run 1. (Right): Average precision scores for the baseline run vs. run 3. In both plots, topics are ordered by increasing average precision score, not by topic ID.

Although runs 2 and 3 were set to a maximum of 10 feedback documents, the average number of documents was respectively 3.86 and 7. For run 2, no relevant feedback documents were found for 10 topics.

In run 3 the top 100 retrieved documents could be used for feedback; however, for 19 of the 50 topics the 10 relevant documents were found in the top 30. For 6 topics, no relevant documents were found. The degree of improvement seems to be strongly related to the quality of the feedback documents: the more relevant documents are used, the better the MAP score is.

We compared the average precision score per topic for the baseline run, run 1 and run 3 (Figure 4(Left) and (Right), respectively). For both non-baseline runs, we see a big variety in the change of average precision scores. Although the MAP score increases when blind feedback is used (run 1), the graph shows that for a majority of the topics average precision does not improve (Figure 4(Left)). In run 3, when relevant documents are used for feedback, MAP increases for most topics. Especially topics with low initial MAP scores benefit. This is likely due to the fact that for most of these topics, most documents used for feedback in run 1 were irrelevant, whereas in run 3 only relevant documents, if available, are used.

In general we can conclude that thesaurus-based feedback improves mean average precision and especially recall. In real retrieval scenarios, asking a user

for feedback is needed to identify relevant documents. While this may seem a time consuming job, for a scientist searching for all available literature on a certain topic, i.e., interested in boosting recall, this might be a good investment.

**Case studies** For two of the four selected topics, topic 3 and 4, no relevant documents can be found in the top 100 results of our baseline run. As a consequence, their retrieval scores do not improve by the first 3 feedback methods. For poorly performing topics, feedback is only useful if a user takes the time to find some relevant documents to use in the feedback procedure. Nevertheless, we hope to be able to define a method for the automatic assignment of MeSH headings to queries (such as [6] or [7]) in future research.

A comparison of the effectiveness of the feedback methods for the other two topics can be found in Table 3. For topic 2, recall is improved for all feedback runs. However, for this topic blind feedback (run 1) works better than the other two relevance feedback runs. Topic 1, which already has a very good retrieval score, is not hurt by adding feedback to the retrieval process. To conclude our feedback analysis, we take a closer look at the results of topic 2 ("Role of p63 and p73 in relation to DNA Damage"). We see that for this topic, feedback does improve MAP and recall, but that the different feedback approaches do not show big differences in scores.

Let us take a look at MeSH headings used for the feedback runs. When we compare the 35 feedback headings for every run with the 10 most frequently occurring MeSH headings of the gold-standard relevant documents, we find that there is an overlap of at least 6 (see Table 4). Only run 4, whose feedback headings are created based on gold-standard data only, shows less overlap. However, this has no effect on the retrieval results, since MAP and recall stay relatively stable for all feedback runs.

When comparing the 35 feedback headings with the list of MeSH headings that occur in at least 10% of all relevant documents (24 headings in total), we find an overlap of 11 for run 1, 15 for run 2, 11 for run 3, and 7 for run 4. For all four runs, most other MeSH headings on the feedback lists are closely related to the 24 frequent headings of the relevant documents. For many of these headings a direct relation, either hierarchical or cross-referential, can be found in the thesaurus.

Hence, this suggests that using a larger number of relevant MeSH headings for feedback does not necessarily improve retrieval. Yet all improvements are

**Table 3.** Retrieval scores for topic 1 and 2.

| **Topic** | Measure | **Baseline run** | **Run 1** | **Run 2** | **Run 3** | **Run 4** |
|---|---|---|---|---|---|---|
| 1 | MAP | 0.8676 | 0.8638 | 0.8908 | 0.8910 | 0.8989 |
| | Recall | 113/115 | 113/115 | 113/115 | 113/115 | 115/115 |
| 2 | MAP | 0.1944 | 0.2145 | 0.2209 | 0.2255 | 0.2302 |
| | Recall | 40/80 | 45/80 | 44/80 | 44/80 | 44/80 |

**Table 4.** Occurrence of top 10 MeSH headings of relevant documents in feedback heading lists.

| MeSH headings | Run 1 | Run 2 | Run 3 | Run 4 |
|---|---|---|---|---|
| DNA-Binding Proteins | X | X | X | X |
| Nuclear Proteins | X | X | X | X |
| Apoptosis | X | X | X | |
| Protein p53 | X | X | X | X |
| DNA Damage | X | X | X | |
| Phosphoproteins | X | X | X | X |
| Trans-Activators | X | X | X | |
| Cultured Tumor Cells | | X | | |
| p53 genes | X | X | X | X |
| Tumor Suppressor Genes | X | X | X | X |

obtained by feedback lists containing at least six MeSH headings frequently occurring in the relevant documents. In future research, we will look deeper into our feedback mechanisms and feedback results to see what the optimal settings for thesaurus-based feedback are.

## 5 Conclusion and Discussion

We studied the role of a thesaurus in biomedical retrieval. The relative effectiveness of controlled and natural languages is one of the longest standing debates in information retrieval, dating back to the original Cranfield experiments [4]. In particular, the use of controlled vocabularies to better articulate natural language queries, usually through some form of query expansion, has received a great deal of attention. For example, for automatic query expansion with thesaurus terms, Srinivasan [20] reports moderate improvement, but the improvement is overshadowed by the improvement due to standard text-based blind feedback. Based on the manual assignment of controlled terms to natural language queries, Hersh et al. [9] report a drop in retrieval effectiveness for a wide range of query expansion methods. A recurring pattern in the literature is that expanding natural language queries with controlled terms pays off for some fraction of the queries, but is detrimental for a larger fraction of the queries.

In light of the inconclusive evidence in the literature, we opted for a somewhat different approach to the question of how to select controlled terms to be added to a natural language query. Traditionally, the selection is based on the topic statement and the goal is to select those terms that are topically relevant for the information need. Our hypothesis is that this selection process should also be based on the role that the controlled terms play in the retrieval process, i.e., whether they are good retrieval cues for the search engine. Hence, to better grasp how a thesaurus can help improve the retrieval process, we performed a detailed analysis of a number of queries. In particular, we tried to analyse vocabulary mismatch problems, the related thesaurus terms, and their influence on

retrieval. Our detailed analysis of four retrieval topics showed that vocabulary mismatch problems between queries and relevant documents have a negative impact on retrieval effectiveness. That is, there is a range of queries for which the natural language statement fails to be effective. In our thesaurus-based experiments, we found that using thesaurus terms for blind and relevance feedback can improve precision as well as recall. In general, the improvements increase with the amount of true relevance feedback provided to the system. However, the fully automatic runs using only pseudo-relevance feedback also led to improved retrieval effectiveness. The inherent shortcoming of feedback-based techniques, as highlighted by our success/failure analysis, is the failure to improve topics for which no relevant document is initially retrieved. To minimize the detrimental effect of query expansion on the fraction of queries for which the natural language query is effective, we use a combination of runs based on the original and on the expanded query.

Our results may contribute to a better understanding of the role of controlled vocabularies in information retrieval [21]. Our study is still limited and we plan to extend it in a number of ways. First, a similar analysis should be performed for a larger set of queries. Second, we plan to experiment with other methods of selecting thesaurus terms based on initially retrieved documents. Third, we want to study different ways of incorporating controlled terms in the retrieval model, and the relation to models of text-based blind feedback. Fourth, we plan to analyse the intrinsic properties of the MeSH thesaurus, including its completeness, coherence, and consistency, and test the robustness of our approaches against imperfect resources. A better understanding of the effectiveness of thesaurus terms as retrieval cues is crucial for the selection of controlled terms. This may also influence our view of the goal of thesaurus-based expansion in the first place. If the natural language queries provide excellent retrieval cues for a large fraction of the queries, we can only hope to improve when the original query fails. That is, we could envision offering thesaurus-based query expansion as a query refinement option: in case a user is unsatisfied with the set of documents returned, she may choose to use the expanded query.

*Acknowledgments* Leonie IJzereef's work was carried out in the context of the Virtual Laboratory for e-Science project (`www.vl-e.nl`). This project is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and is part of the ICT innovation program of the Ministry of Economic Affairs (EZ). Jaap Kamps was supported by a grant from the Netherlands Organization for Scientific Research (NWO) under project numbers 612.066.302 and 640.001.501. Maarten de Rijke was supported by grants from NWO, under project numbers 017.001.190, 220-80-001, 264-70-050, 365-20-005, 612.000.106, 612.000.207, 612.069.006, and 612.066.302.

## References

[1] W. Ceusters, B. Smith, and L. Goldberg. A terminological and ontological analysis of the NCI thesaurus. *Methods of Information in Medicine*, 2005, in press.

[2] W. Ceusters, B. Smith, A. Kuman, and C. Dhaen. Mistakes in medical ontologies: Where do they come from and how can they be detected? In *Ontologies in Medicine: Proceedings of the Workshop on Medical Ontologies*. IOS Press, Amsterdam, 2003.

[3] C. W. Cleverdon. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, College of Aeronautics, Cranfield UK, 1962.

[4] C. W. Cleverdon. The Cranfield tests on index language devices. *Aslib*, 19:173–192, 1967.

[5] E. A. Fox and J. A. Shaw. Combination of multiple searches. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 243–252. National Institute for Standards and Technology. NIST Special Publication 500-215, 1994.

[6] J. C. French, A. L. Powell, F. Gey, and N. Perelman. Exploiting a controlled vocabulary to improve collection selection and retrieval effectiveness. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 199–206, New York, NY, USA, 2001. ACM Press.

[7] N. Grabar, P. Zweigenbaum, L. Soualmia, and S. Darmoni. Matching controlled vocabulary words. In G. Surjan, R. Engelbrecht, and P. McNair, editors, *Proceedings of MIE 2003, Eighteenth International Congress of the European Federation for Medical Informatics*. IOS Press Publisher, 2003.

[8] W. Hersh, R. T. Bhuptiraju, L. Ross, P. Johnson, A. Cohen, and D. Kraemer. Trec 2004 genomics track overview. In *The Thirteenth Text Retrieval Conference: TREC 2004*, Gaithersburg, MD, 2004. National Institute of Standards and Technology.

[9] W. Hersh, S. Price, and L. Donohoe. Assessing thesaurus-based query expansion using the UMLS metathesaurus. In *Proc. of the 2000 American Medical Informatics Association (AMIA) Symposium*, pages 344–348, 2000.

[10] M. Iivonen. Consistency in the selection of search concepts and search terms. *Information Processing and Management*, 31:173–190, 1995.

[11] J. Kamps. Improving retrieval effectiveness by reranking documents based on controlled vocabulary. In S. McDonald and J. Tait, editors, *Advances in Information Retrieval: 26th European Conference on IR Research (ECIR 2004)*, volume 2997 of *Lecture Notes in Computer Science*, pages 283–295. Springer-Verlag, Heidelberg, 2004.

[12] W. Kraaij, M. Weeber, S. Raaijmakers, and R. Jelier. MeSH based feedback, concept recognition and stacked classification for curation tasks. In *Proceedings of TREC 2004*. NIST, 2005.

[13] F. Lancaster. *Vocabulary Control for Information Retrieval*. Information Resources Press, Arlington, Virginia, second edition, 1986.

[14] National Library of Medicine. Medical Literature Analysis and Retrieval System Online (MEDLINE). `http://www.nlm.nih.gov/pubs/factsheets/medline.html`, May 2005.

[15] National Library of Medicine. Medical Subject Headings (MeSH). `http://www.nlm.nih.gov/mesh/`, May 2005.

[16] National Library of Medicine. Unified Medical Language System (UMLS). `http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html`, May 2005.

[17] J. Paralic and I. Kostial. Ontology-based information retrieval. In *Proceedings of the 14th Int. Conference on Information and Intelligent Systems - iis2003*, pages 23–28, 2003.

[18] T. Saracevic and P. B. Kantor. A study of information seeking and retrieving. III. searchers, searches, overlap. *Journal of the American Society for Information Science and Technology*, 39:197–216, 1988.

[19] J. Savoy. Bibliographic database access using free-text and controlled vocabulary: an evaluation. *Information Processing and Management*, 41:873–890, 2005.

[20] P. Srinivasan. Query expansion and MEDLINE. *Information Processing and Management*, 32(4):431–443, 1996.

[21] E. Svenonius. Unanswered questions in the design of controlled vocabularies. *Journals of the American Society for Information Science*, 37:331–340, 1986.

[22] TREC Genomics Track. TREC Genomics Track. `http://ir.ohsu.edu/genomics/`, May 2005.

[23] E. M. Voorhees. Using WordNet to disambiguate word senses for text retrieval. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 171–180, New York, NY, USA, 1993. ACM Press.

[24] J. Wilbur. Non-parametric significance tests of retrieval performance comparisons. *Journal of Information Science*, 20:270–284, 1994.