On the Use of MeSH Headings to Improve Retieval Effectiveness

Stephen Blott
Cathal Gurrin
Gareth J. F.Jones
Alan F. Smeaton
Thomas Sodring

School of Computing
Dublin City University
Dublin, Ireland

{sblott,cgurrin,gjones,asmeaton,tsodring}@computing.dcu.ie

Abstract

Geneticists today spend as much time cataloguing and analysing digital data as they do in wet labs. One of the key problems they face, is that of identifying genes and publications that are relevant to their work or research. This paper describes an approach to retrieval that incorporates structural MeSH heading data into the search process. In particular, we describe a technique using a pseudo-relevance feedback process based on MeSH terms to improve retrieval effectiveness. Experimental results suggest improvements in mean average precision on the order of three percent.

1 Introduction

Molecular biologists study the biochemical function, chemical structure and evolutionary history of genes and proteins from all types of organisms, from human beings to fruit flies and yeast [8, 3]. While molecular biologists still spend much of their time in wet labs, they nowadays often spend equally as much time in front of computers. Information has become a critical research tool, and several large genomic databases have been created to facilitate the exchange of information within the community. These databases are repositories not just for genetic information, such as genes and gene sequences, but also for papers and reports relating to the sequencing and discovery of that genetic information, and the associated bibliographic data and citation indexes.

Among the larger examples of genomic databases are the nucleotide sequence database operated jointly by GenBank [4] at the National Centre for Biological Information in the US, the DNA Data Bank of Japan [1], and EMBL [2], the European Molecular Biology Laboratory. These databases have become huge. The GenBank nucleotide database, for instance, contains nucleotide sequences

from more than 130,000 different organisms. As of August 2002, GenBank contained approximately 22,617,000,000 bases in 18,197,000 sequence records. Moreover, the GenBank database is growing as rapidly now as it ever has.

Life scientists spend prolonged periods of time using these databases. They may begin searching among research literature, and then search for related genes and gene sequences within GenBank. Bibliographic data associated with the related gene sequences may then lead the scientist back to other literature, and so on. Citation indexes between articles in the research literature can also be used to discover relevant related articles. Such a search session may involve navigation back and forth between genetic databases and document databases. Anybody who has searched the web is familiar with the frustrating experience of pursuing links, backtracking, and returning to the search engine to reformulate the query and begin again.

Genomic databases are not isolated collections of data. Rather, they are interrelated databases of bibliographic data, genetic data, and the links between these. The hypothesis underlying the work described here, is that the links between these databases can be used to improve retrieval effectiveness. Specifically, our approach is analogous to the way links are used on the world-wide web, as popularised in the well-known Google search engine, using graph topology information to locate potentially highly relevant nodes in the overall graph. In the case of Google, the "graph" consists of web pages and hypertext links between them. Highly-ranked web pages are those which have a high degree of similarity with the user's query, have many other web pages of relevance pointing to them, and also point themselves to many other web pages which are similar to the query.

In this paper, we explore retrieval techniques exploiting the MeSH terms in genomic bibliographics databases. MeSH, standing for "Medical Subject Headings", is a controlled vocabulary thesaurus managed by the US National Library of Medicine. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity [5]. Quoting [5]:

> *MeSH descriptors are arranged in both an alphabetic and a hierarchical structure. At the most general level of the hierarchical structure MeSH terms are very broad headings such as "Anatomy" or "Mental Disorders". At more narrow levels are found more specific headings such as "Ankle" and "Conduct Disorder". There are 21,973 descriptors in MeSH. In addition to these headings, there are 132,123 headings called Supplementary Concept Records within a separate chemical thesaurus. There are also thousands of cross-references that assist in finding the most appropriate MeSH Heading, for example, "Vitamin C see Ascorbic Acid". These entries include 23,512 printed see references and 102,346 other entry points.*

MeSH terms are manually assigned to documents within genomic collections. In this paper, we investigate the hypothesis that documents discussing related topics will have similar MeSH terms associated with them, and that this similarity can be used to improve the overall retrieval effectiveness above a system treating individual documents as discrete entities.

Our experimental evaluation is in terms of the Genomic Track of the 2003 Text Retrieval Conference (TREC). In particular, our approach was to base our experimentation on runs of standard retrieval methods, and then use the MeSH categorisation of documents as the basis for a pseudo-

relevance feedback process to improve upon the initial ranking.

2 Medical Subject Headings (MeSH) and Medline

This section provides the background material necessary to understand the approach described in the rest of the paper.

2.1 Medical Subject Headings (MeSH)

MeSH terms are used to categorise medical publications much as the Dewey-Decimal system is used in general libraries. MeSH is a vocabulary of terms, from the very general to the very specific, in terms of which medical literature is classified. Examples of general and specific terms include:

C01	Bacterial Infections and Mycoses
C01.252	Bacterial Infections
C01.252.400.155.569.200	Erythema Chronicum Migrans
C01.252.400.155.569.600	Lyme Neuroborreliosis

Based on the code on the left in the example, above, the terms are categorised hierarchically, with the length of the code corresponding loosely with its specificity. In addition, the same term is often repeated at different places within that hierarchy. For example:

C01.252.400.825.480	Lyme Disease
C01.252.847.193.569	Lyme Disease

Intuitively, terms whose codes share a long common prefix are specific and similar. For example, the three terms below are similar to one another:

C01.252.847.840.744.725	Syphilis, Congenital
C01.252.847.840.744.800	Syphilis, Cutaneous
C01.252.847.840.744.871	Syphilis, Latent

each being related to disease caused by members of the syphilis bacterial family, and each sharing the prefix "C01.252.847.840.744". Other terms which do not share such long common prefixes might be considered less similar, for example:

C01.252.400.210.210.250	Conjunctivitis, Inclusion

The MeSH terms used in this work was "2003 MeSH", which includes 21,837 terms organised into a hierarchy of 39,829 distinct nodes.

The basis of the work described here is to exploit the similarity of MeSH the terms used to categorise biomedical publications to improve retrieval effectiveness from biomedical bibliographic databases.

2.2 The Medline Database

Medline [6] is the US National Library of Medicine (NLM) database of indexed journal citations and abstracts now covering nearly 4,500 journals published in the United States and more than 70 other countries. Medline includes references to articles indexed from 1966 to the present, with new citations added weekly. All citations are assigned MeSH Terms and Publication Types from NLM's controlled vocabulary. MEDLINE citations and abstracts are available as the primary component of NLM's PubMed database [6], which is searchable via the Internet.

2.3 The TREC Genomics Track (TrecGen), 2003

In 2003, the Text REtrieval Conference (TREC) organised by the US National Institute for Standards and Technology (NIST) ran a track on genomic retrieval (TrecGen). Under TrecGen, participants were provided with an extract of over 500,000 records from the Medline database, and a set of sample topics. Participants were then set the task of locating documents from the Medline database that are relevant to the sample topics. This paper reports on a set of experiments carried out within the TrecGen framework designed to fulfil this task.

3 Using MeSH Terms for Retrieval

Since the focus of our work is to seek to make use of the MeSH terms to improve retrieval, we took as the starting point for the approach described the baseline set of results generated using SMART-based Okapi approach provided by Jacques Savoy of the University of Neuchatel, Switzerland. The basic approach in our work was to use a notion of similarity between documents measured using MeSH terms to generate a new score for each document. This score is then combined with the original score generated by the Okapi system. The document list is then re-ranked with the intention of creating a new ranking that places more relevant documents higher in the ranked list.

We explored two approaches to this problem, which we refer to as Method 1 and Method 2, respectively. These methods are described in detail below. However, we begin with a discussion of the issues and techniques that are common to both methods.

3.1 Basic use of MeSH Data

The Medline database consists of over 500,000 citations, each including title, authors and abstract. Given a topic, we assume an initial ranking of these Medline citations. The approach described here is then to re-rank the top N documents based on the similarity of the top N documents to the most highly ranked documents. The aim of the work was to augment the original ranked list with a document similarity score measured using MeSH terms.

Within the Medline database, each citation is annotated with several MeSH terms, with the average being around 16 terms per document. As noted in Section 2.1, MeSH terms are associated with codes, and those codes are organised in a hierarchy. In a first processing step, the mesh terms were replaced

with the corresponding codes. Terms with multiple codes were replaced with all of the corresponding codes.

Each document is labelled with only a small number of MeSH terms. We might think of these as analogous to keywords assigned to other documents. Since the number of MeSH terms associated with a document is generally low and MeSH terms are often very specific, it will often be the case that there are few MeSH terms in common between documents. Because of this, standard information retrieval matching will often find very few matches between documents and the resulting match scores will bear more relation to random matches of MeSH between documents than a reliable measure of document topic similarity. Thus we need an approach to give us a quantitative measure of the "closeness" of terms that do not match exactly.

The hierarchy of MeSH codes can be considered to be a tree. The structure of the tree is used here to give us a measure of the distance between individual MeSH terms. This distance measure was computed as follows.

The collection frequency was calculated for each node in the tree; that is, the number of documents in the collection annotated with the corresponding MeSH term. These collection frequencies were used to calculate weighted collection frequencies where the weighted collection frequency of any node in the tree is the number documents that contain a MeSH term corresponding to that node or to any descendant of the node. This is motivated by the notion that the descendants are members of the class described by their parent. Intuitively, this corresponds to the idea of pushing collection frequencies up the tree. The weighted collection frequency of the root[1] is just the count of all documents in the collection, since all MeSH terms are descendants of the root.

A simplified version of this calculation is shown in Figure 1. The left-hand tree shows some assumed collection frequency values, and the right-hand tree shows the corresponding weighted collection frequencies.

In practice we adopted a slightly more complex counting procedure. In general Medline citations are annotated with several MeSH terms. In calculating the weighted collection frequency for a node, it is possible that a single document could contribute to the collection frequency of more than one descendant. In this case, the weighted collection must be calculated to reflect this. Thus, the weighted collection frequency of a node is less than the sum of its own collection frequency and its children's weighted collection frequencies if the same document appears more than once among the children.

The similarity of MeSH terms was calculated in terms of these weighted collection frequencies. In particular, the similarity between two MeSH terms was taken to be the the weighted document frequency of their lowest shared ancestor divided by the total number of documents in the collection. For example, the similarity of nodes A and B in Figure 1 is $\frac{111}{125}$, whereas that of nodes C and D is $\frac{14}{125}$ (with smaller values indicating greater similarity).

[1]The MeSH classification system is actually a forest rather than a single tree. We obtained a single tree by artificially creating a new root that is a parent of all the actual roots.

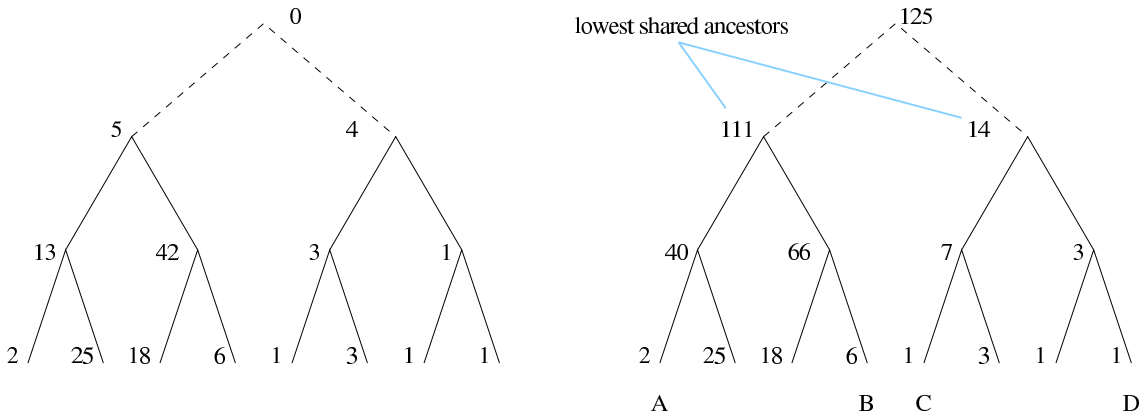

Figure 1: Calculation of weighted collection frequencies from collection frequencies.

3.2 Method 1

The TrecGen search topics do not include MeSH terms. Thus in order to exploit the terms contained in the documents themselves our first method sought to form MeSH term search queries using a pseudo-relevance feedback approach. For each of the 50 topics we took the top 100 PMIDs (document identifiers) in the lists provided by Jacques Savoy and formed a list of corresponding MeSH terms (with duplicates removed). A fixed number of these top ranked documents were then assumed to be relevant for each topic and the MeSH terms contained in these documents ranked using Robertson's Offer Weight [7]. The top ranked terms were then used as a search query which was scored against the top ranked documents.

The similarity between each MeSH term in the query and each document was computed using the distance tree. To compute the matching score between the query and each document, for each query term the highest scoring matching term was found in the document. The overall matching score between the query and document was taken as the highest scoring individual term match between the query and the document, with the highest scoring value the deciding factor. This score was then combined with the original matching score for the document computed by Savoy in a weighted sum and the document list re-ranked.

A number of runs were carried out with the training topic set to optimise the parameters of the system. After experimentation the following values were selected for the submitted test run. The top 80 ranked terms generated from assuming that the top 100 documents were relevant were used as the search query. It was only found to be beneficial re-rank the top 20 documents from Savoy's list. This effectively means that this method has no effect on precision at cutoff values below 20.

This then lead to a new score for each of the documents in the top 20 which were combined with the remaining 980 documents to produce a final ranking. The score combination was carried out by first normalising both Savoy's Okapi score and our MeSH derived matching score with respect to the highest score in each list and then multiplying Savoy's score by 9 and adding to our score.

The result from our official Method 1 submission is shown in Table 3. Table 4 shows the percentage changes for this method compared to Savoy's baseline run. It can be seen that Method 1 produces

small improvements in average precision and precision at various cut off levels.

The use of the maximum term-term matching score between the query and document is not the standard approach taken in information retrieval where we would normally expect to use the sum of the matching scores for all the query terms. However, the use of the MeSH tree and the distance matching score between the terms in the query and the document is not a standard approach in information retrieval. In addition to the official submission we carried out exploratory runs using the sum of the query term matching scores. The results of these runs produced significant reductions in retrieval effectiveness and thus we did not use this approach in the final submission.

3.2.1 Further Experiments

Subsequent to our official submission we considered this issue further. That we found best performance on the development topics by assuming that the 100 ranked documents are relevant with 80 selected expansion terms for the search query is not a typical pseudo-relevance feedback result for Robertson's query expansion method. We would more typically expect to assume 10 documents relevant and form a query by selecting about the 20 top ranked terms. Given that we know that most of the assumed relevant documents in 100 will be non-relevant and that many of the 80 selected terms will not be related to relevant documents, it is clear that many of the terms in the query will be "noise" among the useful terms. In standard information retrieval where an exact match is required between the query and document terms many of these noise query would fail to find a match in the documents or would do so in small numbers and not impact significantly on the ranking of the overall query document matching scores. By computing a non-zero matching score between all term-term pairs the summation overall query terms will introduce significant noise in the overall score and lead to potentially meaningless ranking of documents. By contrast in this situation using the maximum term-term matching score as the query document matching score will in general eliminate the noise since the best match is likely to be between two strongly related (possibly identical) nodes in the MeSH tree.

In order to attempt to overcome this problem we conducted further experiments computing the query-document matching score the opposite way round. We know that each document has a (usually) small number of carefully selected highly significant MeSH terms associated with them. Thus we decided to match each document against the query. In this case we only include term-term associations once for each document term and we expect that most of the noise terms in the query will not often be the highest scoring matching term for the document terms. Thus the term-term matching scores should now on average be more meaningful.

3.3 Method 2

This section now describes, Method 2, a second, more ad hoc approach to using MeSH similarity to improve retrieval effectiveness. Under Method 2, each document was compared pairwise with each of the top 10 ranked documents to calculate This mesh score was calculated as illustrated in Figure 2. A pairwise score is calculated for document paired with each of the top 10 ranked documents. The mesh score for each document is the maximum of the pairwise scores. The pairwise score was calculated as

```
for each document i
  mesh-score[i] = 0
  for each top 10 document j, where j <> j
    score = 0
    for each MeSH term Mi in i
      for each MeSH term Mj in j
        if similarity(Mi, Mj) > score then
          score = similarity(Mi, Mj)
    if score > mesh-score then
      mesh-score[i] = score
  end for j
  mesh-score[i] = old-score[i] + CONSTANT * mesh-score[i]
end for i
```

Figure 2: Method 2 Algorithm

	Okapi run (SMART)	Method 1	Method 2
Mean Average Precision	0.1635	0.1669	0.1667
Precision at 5 docs	0.1480	0.1560	0.1680
Precision at 10 docs	0.1280	0.1360	0.1360
Precision at 15 docs	0.1133	0.1120	0.1187
Precision at 20 docs	0.1000	0.1040	0.0980
Precision at 30 docs	0.0827	0.0827	0.0058

Figure 3: Experimental Results

the maximum mesh similarity between any pair of MeSH terms between the documents.

The CONSTANT referred to on the second last line of the algorithm is simply a weighting factor (or a fudge factor). Decreasing this factor decreases the influence of the mesh scores on the resulting ranking, whereas increasing it, increases the impact of the mesh scores on the resulting ranking.

4 Experimental Evaluation

The experimental results are summarised in Figures 3 through 6. Figure 3 illustrates the mean average precision and precision at N for the baseline Okapi run, and the two methods described here. Percentage improvements are summarised in Figure 4.

Figures 5 and 6 are the equivalent results for the extended methods described in Section 3.2.1. It can be seen that there is improvement in the average precision for selection of the maximum term-term matching score between the document and the query (the same method as used in the official submission), but that the result for summing across all the terms appearing in each document is now better than the maximum matching score approach. Results for cutoff precision however still appear to be favour the maximum matching score approach.

	Okapi run (SMART)	Method 1	Method 2
Mean Average Precision	0.1635	2.08%	1.96%
Precision at 5 docs	0.1480	5.41%	13.51%
Precision at 10 docs	0.1280	6.25%	6.25%
Precision at 15 docs	0.1133	-1.15%	4.77%
Precision at 20 docs	0.1000	4.00%	-2.0%
Precision at 30 docs	0.0827	0.00%	-92.99%

Figure 4: Percentage Improvement

	Okapi run (SMART)	Max Score	Summed Scores
Mean Average Precision	0.1635	0.1680	0.1689
Precision at 5 docs	0.1480	0.1720	0.1560
Precision at 10 docs	0.1280	0.1240	0.1240
Precision at 15 docs	0.1133	0.1040	0.1013
Precision at 20 docs	0.1000	0.1010	0.0960
Precision at 30 docs	0.0827	0.0820	0.0820

Figure 5: Experimental Results

5 Conclusion

Our results for the TrecGen task demonstrate that incorporating relationships between the MeSH term labels of documents retrieved using a standard information retrieval can lead to small improvements in both average precision and cutoff precision.

The experiments described in this paper are of an exploratory nature and further work needs to be concentrated in several areas. In particular the method used to compute term-term similarity in the MeSH is only one of many possibilities and these need to explored and analysed in detail. Further than this it might be possible to use the structure of the MeSH tree more effectively than computing a term-term similarity metric.

Acknowledgements. We would like to thank Jacques Savoy of the University of Neuchatel, Switzerland, for his generosity in making available the Okapi runs that were the basis for the experiments described here. We would also like to thank Susan McDonnell and Donal O'Shea of the School of Biotechnology at DCU for their help and advice on the biological background to this work. In addition, this work was funded in part by the GenIRL (Genomic Information Retrieval using Links) Project, sponsored under Enterprise Ireland's Basic Research Grants Scheme.

References

[1] DNA Data Bank of Japan.
`http://www.ddbj.nig.ac.jp/`.

	Okapi run (SMART)	Max Score	Summed Scores
Mean Average Precision	0.1635	2.75%	3.30%
Precision at 5 docs	0.1480	16.22%	5.41%
Precision at 10 docs	0.1280	-3.13%	-3.13%
Precision at 15 docs	0.1133	-8.21%	-10.59%
Precision at 20 docs	0.1000	1.00%	-4.00%
Precision at 30 docs	0.0827	-0.85%	-0.85%

Figure 6: Percentage Improvement

[2] European Molecular Biology Laboratory.
`http://www.ebi.ac.uk/Databases/`.

[3] L. Hunter. *Artificial Intelligence and Molecular Biology*. Out of print, now available on the web, 1993.
`http://www.aaai.org//Library/Books/Hunter/hunter.html`, see particularly chapter 1.

[4] National Center for Biological Information.
`http://www.ncbi.nlm.nih.gov/`.

[5] N. L. of Medicine. Medical subject headings. `http://www.nlm.nih.gov/mesh/`.

[6] U. N. L. of Medicine. `http://www.ncbi.nlm.nih.gov:80/entrez/query/static/overview.html`.

[7] S. E. Robertson and K. S. Jones. Simple proven approaches to text retrieval. Technical Report Technical Report TR356, Cambridge University Computer Laboratory, 1997.

[8] J. Shrager. Just enough molecular biology for compter scientists. Notes available on web, 2002.
`http://aracyc.stanford.edu/~jshrager/jeff/mbcs/`.