Balancing Manual and Automatic Indexing for Retrieval of Paper Abstracts*

Kwangcheol Shin[1], Sang-Yong Han[1], and Alexander Gelbukh[1,2]

[1] Computer Science and Engineering Department,
Chung-Ang University, 156-756, Seoul, Korea
kcshin@archi.cse.cau.ac.kr, hansy@cau.ac.kr
[2] Center for Computing Research,
National Polytechnic Institute, 07738, DF, Mexico
gelbukh@gelbukh.com, www.Gelbukh.com

Abstract. MEDLINE is a widely used very large database of abstracts of research papers in medical domain. Abstracts in it are manually supplied with keywords from a controlled vocabulary called MeSH. The MeSH keywords assigned to a specific document are subdivided into MeSH major headings, which express the main topic of the document, and MeSH minor headings, which express additional information about the document's topic. The search engine supplied with MEDLINE uses Boolean retrieval model with only MeSH keywords used for indexing. We show that (1) vector space retrieval model with the full text of the abstracts indexed gives much better results; (2) assigning greater weights to the MeSH keywords than to the terms appearing in the text of the abstracts gives slightly better results, and (3) assigning slightly greater weight to major MeSH terms than to minor MeSH terms further improves the results.

1 Introduction

MEDLINE is a premier bibliography database of National Library of Medicine (NLM; `www.nlm.gov`). It covers the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, the preclinical sciences, and some other areas of the life sciences. MEDLINE contains bibliographic citations and author abstracts from over 4,600 journals published in the United States and in 70 other countries. It has approximately 12 million records dating back to 1966 [8].

Medical Subject Headings (MeSH) is the authority list of controlled vocabulary terms used for subject analysis of biomedical literature at NLM [6]. It provides an extensive list of medical terminology having a well-formed hierarchical structure. It includes major categories such as anatomy/body systems, organisms, diseases, chemicals and drugs, and medical equipment.

* Work supported by the ITRI of the Chung-Ang University. The third author is currently on Sabbatical leave at Chung-Ang University. Corresponding author: S.-Y. Han.

K. Shin, S.-Y. Han, A. Gelbukh. Balancing Manual and Automatic Indexing for Retrieval of Paper Abstracts. Text, Speech and Dialogue (TSD-2004). Lecture Notes in Artificial Intelligence, N 3206, Springer-Verlag, 2004, pp. 203–210.

Expert annotators of the National Library of Medicine databases, based on indexed content of documents, assign subject headings to each MEDLINE document for the users to be able to effectively retrieve the information that explains the same concept with different terminology. Manual annotation with MeSH terms is a distinctive feature of MEDLINE [8].

MeSH keywords assigned to each individual document are subdivided into *MeSH Major headings* and *MeSH Minor headings*. MeSH Major headings are used to describe the primary content of the document, while MeSH Minor headings are used to describe its secondary content. On average, 5 to 15 subject headings are assigned per document, 3 to 4 of them being major headings [6].

MEDLINE is supplied with its own search engine. To use it, users give their keywords as a query to the system. The system automatically converts the query into Boolean form and retrieves the data from the MeSH field and the author information fields. No relevance ranking is provided; the retrieved documents are returned in no particular order.

We show that applying a vector space model-based search engine cite3 to full-text MEDLINE data gives much better results. Then, we show that assigning greater weights to the MeSH terms than to the words from the full text of the document slightly improves the quality of the results, which is further improved with assigning greater weights to MeSH major headings than to MeSH minor headings. In this way, we obtain slightly better ranking of the search result than with the traditional vector space model which used in the SMART system [10] and much better results than with the Boolean model used in the search engine provided with MEDLINE.

On the other hand, our experiments show that the improvement obtained with modulating the term weights is less than one could expect.

This paper is organized as follows. Section 2 gives a short introduction to the vector space model. Section 3 describes the proposed technique to modulate the MeSH terms weights. Section 4 presents the experimental results. Finally, Section 5 provides some discussion and conclusions.

2 Vector Space Model

The vector space model has the advantage over the Boolean model in that it provides relevance ranking of the documents: unlike the Boolean model which can only distinguish relevant documents from irrelevant ones, the vector space model can indicate that some documents are very relevant, others less relevant, etc.

In the vector space model [10] the documents are represented as vectors with the coordinates usually proportional to the number of occurrences (*term frequency*) of individual content words in the text. Let the document collection consist of d documents containing in total n different words except the stopwords (functional words, too frequent words, and too rare words). Then the vector space model for this collection is represented by the (sparse) $d \times n$-dimensional matrix $w = \|w_{ij}\|$ [3, 4]. Here w_{ij} is the weight of the i-th term in j-th document, usually

```
MJ  BONE-DISEASES-DEVELOPMENTAL: co. CYSTIC-FIBROSIS: co. DWARFISM: co.
MN  CASE-REPORT. CHILD. FEMALE. HUMAN. SYNDROME.
AB  Taussig et al reported a case of a 6-year-old boy with the Russell
    variant of the Silver-Russell syndrome concomitant with cystic
    fibrosis. We would like to describe another patient who...
```

Fig. 1. A sample of MEDLINE data

calculated as the *tf-idf* (*term frequency-inverse document frequency*) value:

$$\textit{tf-idf}_{ij} = \frac{f_{ij}}{\max f_{kj}} \log \frac{n_i}{n}. \tag{1}$$

where f_{ij} is the frequency of the term i in the document j and n_i is the number of the documents where the i-th term occurs.

The similarity between two documents is measured using the cosine measure widely used in information retrieval—the cosine of the angle between the two vectors x_i and x_j:

$$s(x_i, x_j) = \frac{x_i^T x_j}{\|x_i\| \, \|x_j\|} = \cos\left(\theta(x_i, x_j)\right),$$

where θ is the angle between the two vectors. To simplify calculations in practice, the vectors are usually normalized so that their norm $\|x\|$ be 1. This measure ranges between 0 (the two documents have no word in common) and 1 (the similarity between two copies of same document).

This measure is easy to understand and its calculation for sparse vectors is very simple [4]. Specifically, the cosine measure between the user query and a document is used to quantitatively estimate the relevance of the given document for the given query.

3 Modulating MeSH Term Weights

MEDLINE documents contain MeSH keywords as shown in Figure 1. The `MJ` field lists the major MeSH terms manually assigned to this document, `MN` field the minor MeSH terms, and `AB` the full text of the document, namely, the abstract of a research paper.

The MeSH terms are known to be more important than other words in the text of the document. Thus, we expected that increasing their weights in each document vector we would obtain better results. Indeed, a MeSH keyword assigned by the reviewer "stands for" several words in the document body that "vote" for this more general keyword. For example, for the text "*... the patient is allergic to ... the patient shows reaction to ... causes itch in patients ...*" the annotator would add a MeSH term ALLERGY. Though this term appears only once in the document description, it "stands for" three matching terms in the text

```
QU  What are the effects of calcium on the physical properties of
    mucus from CF patients?
RD  139 1222 151 2211 166 0001 311 0001 370 1010 392 0001 439 0001
    440 0011 441 2122 454 0100 461 1121 502 0002 503 1000 505 0001
```

Fig. 2. An example of a CF query with answers

body—namely, *allergic*, *reaction*, and *itch*. Our hypothesis was that increasing its weigh would more accurately describe the real frequency of the corresponding concept in the document and thus lead to better retrieval accuracy.

In the same way, we supposed that assigning slightly greater weight to the more important MeSH major headings than to the less important MeSH minor headings would reflect the intuition of the human annotator on the intensity of the corresponding topics.

As the experimental results reported in Section 4 show, these hypotheses proved to be true, but the improvement was much less than one could expect.

We used the following procedure. First, we assigned the weights as described in the previous section (with the length of all vectors normalized to 1). Then we use the following formula to increase the weight of MeSH terms:

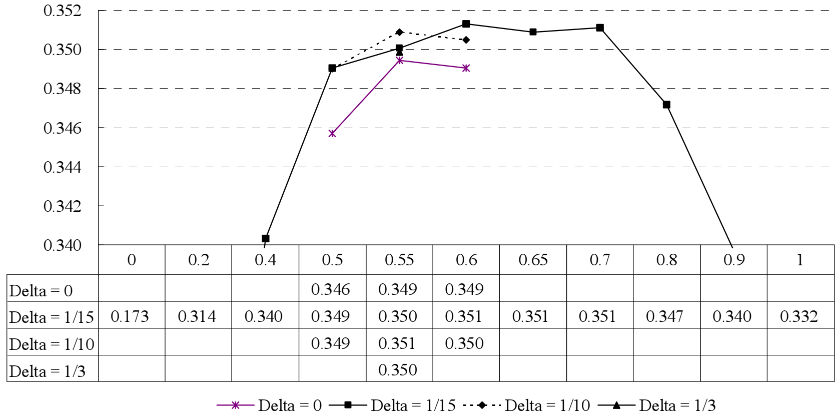
$$w_{ij} \leftarrow \begin{cases} (1+\delta)\rho\, w_{ij} & \text{if term } i \text{ is a MeSH Major term in document } j \\ (1-\delta)\rho\, w_{ij} & \text{if term } i \text{ is a MeSH Minor term in document } j \\ (1-\rho) w_{ij} & \text{if term } i \text{ is not a MeSH term in document } j, \end{cases} \tag{2}$$

where ρ is a parameter regulating the sensitivity of the formula to the MeSH terms, and δ is parameter regulating the sensitivity to the difference between the major and minor MeSH headings. With this:

- When $\rho = 1$, the texts of the abstracts are ignored, and only the MeSH keywords are taken into account in indexing, as it is currently done in the search engine supplied with MEDLINE.
- When $\rho = 0$, the MeSH terms are ignored, and only the full texts of the abstracts are taken into account in indexing, as in standard search engines such as the SMART system [10].
- When $\rho = 0.5$, both the MeSH terms and the full texts of the abstracts are taken into account in indexing, without any distinction. This is equivalent to ignoring the field labels in Figure 1.
- Finally, with other values of ρ, more attention is paid to either MeSH headings or the full text of the abstracts.

4 Experimental Results

We experimented with the well-known Cystic Fibrosis (CF) reference collection, which is a subset of MEDLINE. It has 1,239 medical data records supplied with



	0	0.2	0.4	0.5	0.55	0.6	0.65	0.7	0.8	0.9	1
Delta = 0				0.346	0.349	0.349					
Delta = 1/15	0.173	0.314	0.340	0.349	0.350	0.351	0.351	0.351	0.347	0.340	0.332
Delta = 1/10				0.349	0.351	0.350					
Delta = 1/3					0.350						

Fig. 3. Experimental results with different parameter ρ and δ, with stemming

100 queries with relevant documents provided for each query. A sample query is shown in Figure 2. The 3-digit numbers are the numbers of the documents known to be relevant for this query. The four-digit groups reflect the relevance scores ranging from 0 to 2 assigned to the given document by four experts who manually evaluated the relevance of each document with respect to each query.

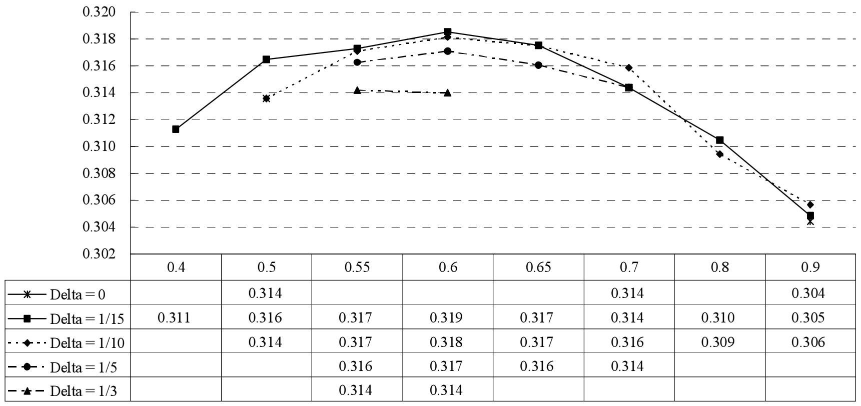
We used the MC program [2] to convert documents from the CF collection into vectors. Stopwords and the terms appearing in less than 0.2% or more than 15% of the documents were excluded. With this, the CF collection had 3,925 terms remaining. Then the *tf-idf* value was calculated for each of the document according to (1) and then the obtained 1,239 vectors were normalized.

Then the weights of the MeSH terms in each document vector and each query vector were modulated by (2), and the vectors were re-normalized. Then we ordered the documents by the cosine measure relative to each query, and measured the average quality.

We measured the quality of the results for an individual query in terms of the R-precision [3], which is the precision of the retrieved set formed by the R highest-ranked documents, where R is the number of relevant documents in the collection according to the human experts' judgments. We considered a document relevant if at least one of the four experts who evaluated the CF collection marked it as relevant.

Figure 3 shows the experimental results with different parameter ρ and δ from (2). One can see that the values of ρ in the range between 0.5 and 0.7 give the best results. These results can be interpreted as follows (cf. the end of the previous section):

- The manually assigned MeSH terms are important for indexing, since indexing only full texts of the abstracts ($\rho = 0$) without the manually assigned MeSH headings greatly deteriorates the results.



	0.4	0.5	0.55	0.6	0.65	0.7	0.8	0.9
Delta = 0		0.314				0.314		0.304
Delta = 1/15	0.311	0.316	0.317	0.319	0.317	0.314	0.310	0.305
Delta = 1/10		0.314	0.317	0.318	0.317	0.316	0.309	0.306
Delta = 1/5			0.316	0.317	0.316	0.314		
Delta = 1/3			0.314	0.314				

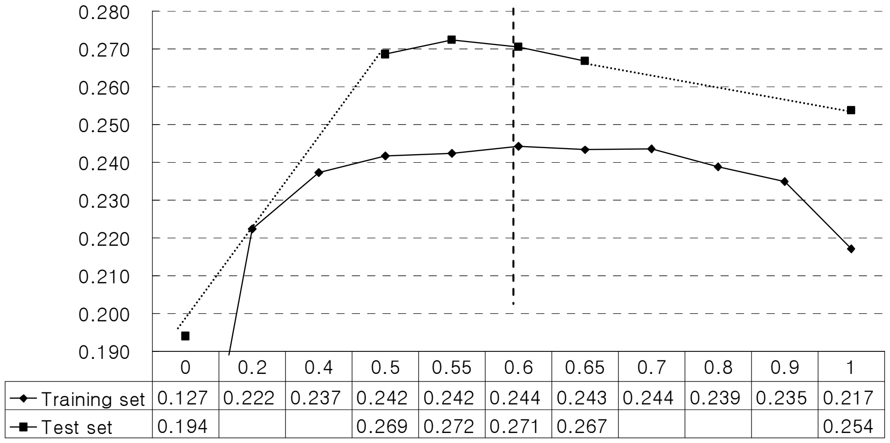
Fig. 4. Experimental results with different parameter ρ and δ, without stemming

- The full texts of the abstracts are not very important. Indexing only MeSH terms ($\rho = 1$) gives acceptable results while reduces memory requirements and increases the speed of the system. This is the option currently used by the search engine supplied with MEDLINE (though it uses the Boolean search model).
- Still, the full texts of the abstracts are of some importance. Taking into account both MeSH headings and the full texts of the abstracts, ignoring the field codes in Figure 1 ($\rho = 0.5$), gives better results than only abstracts or only MeSH headings.
- As expected, the optimal results are achieved with slightly greater weights of the MeSH terms than those of the words from the texts of the abstracts ($\rho = 0.6$).
- What is more, giving slightly greater weights to MeSH major headings than to MeSH minor headings ($\delta = 1/15$) gives better results than equal weights for all MeSH keywords ($\delta = 0$).

The high importance of the MeSH terms observed in our experiments might result from the specific form of the queries in the CF collection. Most of the significant words in these queries are MeSH terms, only one of the 100 queries not containing any MeSH terms. No surprise that when we ignore all MeSH terms, there are too few significant words left in the queries. It is not clear whether the effect of the MeSH terms would be the same if the users formulating the queries were not aware of the MeSH vocabulary.

Figure 3 presents the results with stemming [9], with which the traditional vector space model shows better results than on non-stemmed data. Experiments without stemming confirm the same conclusions, see Figure 4.

To verify that our results do not suffer from overlearning, we divided the set of queries into two portions, 90 and 10 queries, and also divided the document collection into two equal parts. Then we applied the 90% of queries to a half of



	0	0.2	0.4	0.5	0.55	0.6	0.65	0.7	0.8	0.9	1
—◆— Training set	0.127	0.222	0.237	0.242	0.242	0.244	0.243	0.244	0.239	0.235	0.217
—■— Test set	0.194			0.269	0.272	0.271	0.267				0.254

Fig. 5. Experimental results with different parameter ρ, $\delta = 1/15$, with stemming

the document collection and again observed the optimal value of $\rho = 0.6$. Then we applied the other 10% of query to the other half of the collection. Though the value of $\rho = 0.6$ was not optimal in this case, it was definitely better than the baseline values $\rho = 0$ (texts only) and $\rho = 1$ (MeSH only) and slightly better than the baseline value $\rho = 0.5$ (ignore text/MeSH distinction), see Figure 5. Stemming was used.

To compare our results with the Boolean model currently used in the search engine provided with MEDLINE, we extracted the MeSH terms from each query and searched the documents with these MeSH terms using the OR operation (if the AND operation is used, no documents are usually retrieved). Test result is shown in Table 1. Precision is used as the measure of quality for Boolean search, while R-precision for ranking method. As one can see, our method gives as much as 2.4 times better results.

5 Conclusions and Future Work

The search engine currently provided with MEDLINE uses Boolean search applied to only MeSH headings of the documents. We have shown that:

- Vector-space model, even if applied to MeSH headings only, gives much better results.
- Taking into account the full texts of the MEDLINE abstracts, and not only MeSH headings, significantly improves the results.
- Annotation with MeSH terms is important: without them, the abstracts alone do not provide enough information for search.
- Assigning greater weights to MeSH terms, and of them, somewhat greater weights to the MeSH major headings than to the MeSH minor headings, gives slightly better results.

Table 1. Boolean versus vector-space model

	Boolean	Boolean with stemming	Suggested method with stemming
Precision or R-precision	0.104	0.095	**0.353**

- The method is not very sensitive to specific values of the parameters used to modulate the term weights.

However, contrary to one's expectations, the improvement achieved with assigning greater weight to the more important MeSH headings proved to be rather insignificant in comparison with equal weights.

With the best combination of the parameters (vector space model, stemming, $\rho \approx 0.6$, $\delta \approx 0.07$) we obtained as much as 2.4 times better results than the system currently provided with MEDLINE.

In the future we plan to investigate the effects of automatic learning individual weights for each MeSH term instead of a common parameter ρ. Also, we plan to try semantically rich representations of the text structure, such as conceptual graphs [7].

References

1. Dhillon I. S. and Modha, D. S. *Concept Decomposition for Large Sparse Text Data using Clustering.* Technical Report RJ 10147(9502), IBM Almaden Research Center, 1999.
2. Dhillon I. S., Fan J., and Guan Y. Efficient Clustering of Very Large Document Collections. *Data Mining for Scientific and Engineering Applications*, Kluwer, 2001.
3. Baeza-Yates, R., and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley, 1999.
4. Frakes W. B. and R. Baeza-Yates. *Information Retrieval: Data Structures and Algorithms.* Prentince Hall, Englewood Cliffs, New Jersey, 1992.
5. Ide E. New experiments in relevance feedback. In G. Salton, editor, *The SMART Retrieval System*, 337-354, Prentice Hall, 1971.
6. Lowe H.J., Barnett O. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *J. American Medical Association*, 1995; 273:184.
7. Montes-y-Gomez, M., A. Lpez Lpez, A. Gelbukh. Information Retrieval with Conceptual Graph Matching. In Proc. DEXA-2000, Database and Expert Systems Applications. *Lecture Notes in Computer Science*, N 1873, Springer, 2000, pp. 312-321.
8. MEDLINE Fact Sheet. `www.nlm.nih.gov/pubs/factsheets/medline.html`.
9. Porter, M. An algorithm for suffix stripping. *Program*, 14, 1980, pp. 130-137.
10. Salton G. and. McGill M. J., *Introduction to Modern Retrieval.* McGraw-Hill, 1983.