Advanced Relevance Feedback Query Expansion Strategy for Information Retrieval in MEDLINE

Kwangcheol Shin,[1] Sang-Yong Han,[1+] Alexander Gelbukh,[1,2*] Jaehwa Park [1]

[1] School of Computer Science and Engineering,
Chung-Ang University, 156-756, Seoul, Korea
kcshin@archi.cse.cau.ac.kr, {hansy,jaehwa}@cau.ac.kr

[2] Center for Computing Research,
National Polytechnic Institute, Zacatenco 07738 DF, Mexico
www.Gelbukh.com

Abstract. MEDLINE is a very large database of abstracts of research papers in medical domain, maintained by the National Library of Medicine. Documents in MEDLINE are supplied with manually assigned keywords from a controlled vocabulary called MeSH terms, classified for each document into major MeSH terms describing the main topics of the document and minor MeSH terms giving more details on the document's topic. To search MEDLINE, we apply a query expansion strategy through automatic relevance feedback, with the following modification: we assign greater weights to the MeSH terms, with different modulation of the major and minor MeSH terms' weights. With this, we obtain 16% of improvement of the retrieval quality over the best known system.

1 Introduction

Relevance feedback is a classic information retrieval (IR) technique that reformulates a query based on documents selected by the user as relevant [10]. Relevance feedback techniques have been recently an active research area in IR.

We experimented with the MEDLINE database maintained by the National Library of Medicine, which is widely used in medical research. It contains ca. 12 million abstracts on biology and medicine collected from 4,600 international biomedical journals. To each document in this database, keywords called MeSH (Medical Subject Headings) are manually added to describe its content for indexing in a uniform manner. This is a specific features of MEDLINE that other databases do not have [5].

In this paper we suggest new a retrieval technique for MEDLINE based on relevance feedback using modulating MeSH terms in query expansion. We show that our technique gives 16% improvement in the quality of retrieval over the best currently known system.

The paper is organized as follows. Section 2 explains the MEDLINE database and MeSH indexing, as well as the vector space model and the relevance feedback

[+] Corresponding author.

[*] The third author is currently on Sabbatical leave at Chung-Ang University.

K. Shin, S.-Y. Han, A. Gelbukh, J. Park. Advanced Relevance Feedback Query Expansion Strategy for Information Retrieval in MEDLINE. In: A. Sanfeliu, J. F. Martínez Trinidad, J. A. Carrasco Ochoa (Eds.) Progress in Pattern Recognition, Image Analysis and Applications (CIARP 2004). Lecture Notes in Computer Science, N 3287, Springer-Verlag, 2004, p. 425–431.

technique. Section 3 discusses related work. Section 4 describes the proposed technique to modulate the MeSH terms' weights in relevance feedback-based query expansion. Section 5 presents our experimental results, and Section 6 draws conclusions.

2 Background

2.1 MEDLINE and MeSH

MEDLINE, a premier bibliography database of National Library of Medicine (NLM, www.nlm.gov), covers the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, the preclinical sciences, and some other areas of the life sciences. It contains bibliographic citations and author abstracts from over 4,600 journals published in the United States and in 70 foreign countries. It has approximately 12 million records dating back to 1966 [5].

MeSH is the acronym for *Medical Subject Headings*. It is the authority list of the vocabulary terms used for subject analysis of biomedical literature at NLM [6]. The MeSH controlled vocabulary, a distinctive feature of MEDLINE, is used for indexing journal articles. It imposes uniformity and consistency to the indexing of biomedical literature. MeSH is an extensive list of medical terminology. It has a well-formed hierarchical structure. MeSH includes major categories such as anatomy/body systems, organisms, diseases, chemicals and drugs, and medical equipment. Expert annotators of the NLM databases, based on indexed content of documents, assign subject headings to each document for the users to be able to effectively retrieve the information that explains the same concept with different terminology [5].

MeSH terms 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].

To use the current MEDLINE search engine, users give their keywords as a query to the system. The system automatically converts such a query to a Boolean query and retrieves data from the MeSH field of the documents. The current system does not use the full text of the documents.

2.2 Vector Space Model

The vector space model has the advantage over the Boolean model (used currently in the search engine provided with MEDLINE) 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 [8] 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. Namely, the following procedure is used for converting documents into vectors:

The vector space model for the entire document collection is determined by the $d \times n$-dimensional matrix $\| w_{ij} \|$, where d is the number of significant words in all documents of the collection (stopwords, i.e.., the functional words and the words with too high and too low frequency, are excluded), n is the number of documents in the collection, and w_{ij} is the weight of the i-th term in j-th document. For these weights, usually the *tf-idf* (*term frequency–inverse document frequency*) value is used:

$$tf\text{-}idf = \frac{f_{ij}}{\max f_{ij}} \log \frac{n_i}{n} \tag{1}$$

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

Using such vectors to represent documents, we can measure the similarity between two documents (vectors) using the cosine measure (the cosine of the angle between the two vectors) widely used in information retrieval. This measure is easy to understand and its calculation for sparse vectors is very simple [8]. 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.

The cosine similarity between two vectors x_i and x_j is calculated as their inner product:

$$s(x_i, x_j) = \frac{x_i^T x_j}{\| x_i \| \| x_j \|} = \cos \theta \tag{2}$$

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. The similarity is in the range between 0 and 1. If the two documents have no words in common, the similarity is 0; the similarity between two copies of same document is 1.

2.3 Query Expansion using Relevance Feedback

To improve the quality of ranking, a number of strategies is used, among which is query expansion: the system automatically adds to the user query certain words (in some very broad sense synonymous to the original ones) that bring relevant documents not matching literally with the original user query.

In the relevance feedback technique, the query is modified using information in a previously retrieved ranked list of documents that have been judged for relevance by the user. A number of methods, such as those suggested by Rocchio and Ide, have been studied within this broad strategy. Using Rocchio's method [11], the new query is derived from old query according to the below formula:

$$\vec{q}_m = \alpha \vec{q} + \frac{\beta}{|D_r|} \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|} \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j \tag{3}$$

D_r : *Set of relevant documents, as identified by the user, among the retrieved documents*
D_n :*Set of irrelevant documents, as identified by the user, among the retrieved documents*
α, β, γ :*Tuning parameters*

The parameter α represents the relative importance of terms in the original query; β and γ are parameters regulating the relative importance of relevant irrelevant information for query expansion.

Ide [10] uses a slightly different formula:

$$\vec{q}_m = \alpha\vec{q} + \beta \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \gamma \max_{irrelevant}(\vec{d}_j) \cdot \tag{4}$$

2.4 Related Work

The best known retrieval technique for MEDLINE is the one introduced by Srinivasan in a series of recent articles focusing on two areas of the retrieval process, indexing [1] and query expansion [2, 3]. Here we briefly introduce this method.

Srinivasan constructs two index vectors for each document: a vector of the (significant—not stopwords) words in the title and abstract (*ta*-vector) and a vector of the (significant) words of the MeSH terms (*m*-vector). With a title-abstract vocabulary of *p* words and a MeSH vocabulary of *q* words, a document is represented as:

$$\vec{d}_j = (w_{1j}, w_{1j}, ..., w_{pj}); (c_{1j}, c_{1j}, ..., c_{qj}) .$$

She generates a single *ta*-vector for each query, since she considers the user's initial free-text query more suitable for searching the title and abstract field.

$$q_{old} = (w_{1q}, w_{2q}, ..., w_{tq}) .$$

Her query expansion strategy consists in adding an *m*-vector to each query representation. This expanded query is used to compute the ranking as a weighted sum of the vector inner products of the corresponding vectors in the documents and queries:

$$\text{Similarity(d, q)} = \sigma * \text{similarity(ta - vectors)} + \text{similarity(m - vectors)} \ , \tag{5}$$

where σ is a parameter that allows one to change the relative emphases on the two types of vectors during retrieval.

Thus, her query expansion consists in adding the MeSH terms of the retrieved documents to the original query:

$$q_{new} = (w_{1q}, w_{2q}, ..., w_{tq}); (c_{1q}, c_{2q}, ..., c_{pq})$$

Note that the *ta*-vectors in q_{old} and q_{new} are always identical. The retrieval process considers both *ta*-vectors and *m*-vectors from the documents and queries as in (5).

3 Modulating MeSH Term Weights

As explained before, MEDLINE data contains MeSH keywords classified for each document into major (more important) and minor (less important) ones as shown in Table 1.

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...

Table 1. A sample of MEDLINE data.

Our idea is to modulate the weight of MeSH terms in each document vector in query expansion, since these terms are more important than the ordinary words in the text of the document. Indeed, a keyword assigned by the reviewer "stands for" several words in the document body that "voted" for this generalized 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 occurrences in the text, namely, *allergic*, *reaction*, and *itch*. Our hypothesis is 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.

It is well known that relevance feedback, which uses the terms contained in relevant documents to supplement and enrich the user's initial query, gives better results than first retrieval result [10]. In this paper, we use a modified relevance feedback model:

$$\vec{q}_m = \alpha\vec{q} + \sum_{\forall \vec{d}_j \in D_r} (\vec{\beta}_j + \vec{d}_j) \qquad (6)$$

D_r : *Set of relevant documents, as identified by the user, among the retrieved documents*

Here by × we denote coordinate-wise multiplication of the two vectors. We give different weights to MeSH terms and to general terms:

$$\beta_{ij} \leftarrow \begin{cases} (\delta + \tau\delta) & : term\ i\ is\ MeSH\ Major\ term\ \ in\ \ \vec{d}_j \\ (\delta - \tau\delta) & : \ term\ i\ is\ MeSH\ Minor\ term\ \ in\ \ \vec{d}_j \\ 0 & : otherwise \end{cases} \tag{7}$$

$$\delta, \tau : Tuning\ \ parameters$$

4 Experimental Results

For the experiments we use the well-known Cystic Fibrosis (CF) dataset, which is a subset of MEDLINE. It has 1,239 medical data records and 100 queries with relevant documents provided for each query. A sample query is shown in Table 2, with relevant document numbers (e.g., 139) and the relevant scores ranging from 0 to 2 obtained from 4 specialists manually evaluating the query and the data (e.g., 1222 stands for the score 1 assigned by the first expert and 2 by all others).

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

Table 2. Part of CF queries.

We used the MC program [8] to produces vectors from the documents. Stopwords and the terms with frequency lower than 0.2% and higher than 15% were excluded. With this, the CF collection had 3,925 terms remaining. Then the *tf-idf* value was calculated for each document and the vectors were normalized; this produced 1,239 document vectors.

For applying the user's initial query, we formed two datasets, one consisting of only abstracts and another one consisting of abstract and MeSH terms, to compare our technique with Srinivasan's one which searches only in the abstracts when performs retrieval according to the initial query.

	Abstract	MeSH+ Abstract
R = the number of relevant docs in collection	4819	4819
#R = number of relevant docs among best R	1343	1511
Sum of scores of #R	4819	6155
#R / R = R-precision	0.279	**0.314**

Table 3. Test result by applying initial query.

Table 3 shows the results on first iteration (the original, not expanded query). We show the average R-Precision on 100 queries and the total value of the relevant scores (taken from the CF collection) of the *R* highest-ranked documents, where *R* is the total number of the documents known to be relevant in the collection. We considered a document to be known to be relevant if it was assigned non-zero score by at least one of the four human experts, see Table 2. One can note 12.51% of improvement in R-precision on MeSH+Abstracts data.

Now, to verify our query expansion technique, we used the documents known to be relevant within the *R* highest-ranked ones (their number is denoted #R), thus simulating the user's relevance feedback. Table 4 shows the result using Srinivasan's technique, and Table 5 shows the result using our technique.

σ	0.0	0.4	0.6	**0.7**	0.8	1.0	1.2
R = number of correct in the collection	4819	4819	4819	**4819**	4819	4819	4819
#R = number of correct docs among best R	2094	2110	2115	**2117**	2115	2108	2100
sum of scores of #R	8424	8528	8559	**8571**	8568	8528	8481
#R / R = R-precision	0.435	0.438	0.439	**0.439**	0.439	0.437	0.436

Table 4. Query expansion results with Srinivasan's technique.

δ ($\tau = \delta / 20$)	0.3	0.4	0.5	0.6	**0.7**	0.8	0.9	1.0
R = number of correct in collection	4819	4819	4819	4819	**4819**	4819	4819	4819
#R = number of correct among best R	2446	2452	2450	2455	**2456**	2448	2441	2440
sum of scores of #R	9445	9466	9411	9455	**9416**	9373	9321	9312
#R / R = R-precision	0.508	0.509	0.508	0.509	**0.510**	0.508	0.507	0.506

Table 5. Query expansion results with our technique.

One can note a 16% improvement over Srinivasan's technique, which is the best currently known technique.

5 Conclusions

We have shown that assigning different weights to major and minor MeSH headings in relevance feedback technique on MEDLINE data gives the results superior to the best known technique, which ignores the difference between the major and minor MeSH heading, treating them in the same way. Our technique shows a 16% improvement in *R*-precision.

References

1. Srinivasan P. Optimal document-indexing vocabulary for MEDLINE. Information Processing and Management, 1996; 32(5):503-514.

2. Srinivasan P. Query expansion and MEDLINE. Information Processing and Management, 1996; 32(4): 431-443.

3. Srinivasan P. Retrieval feedback in MEDLINE. Journal of the American Medical Informatics Association, 1996; 3(2):157-167.

4. MEDLINE Fact Sheet. www.nlm.nih.gov/pubs/factsheets/medline.html.

5. 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.

6. 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.

7. Dhillon I. S., Fan J., and Guan Y.,: Efficient Clustering of Very Large Document Collections. Data Mining for Scientific and Engineering Applications, Kluwer Academic Publishers, 2001.

8. Frakes W. B. and Baeza-Yates R., *Information Retrieval: Data Structures and Algorithms*. Prentince Hall, Englewood Cliffs, New Jersey, 1992.

9. Ide E., "New experiments in relevance feedback" In G. Salton, editor, The SMART Retrieval System, 337-354, Prentice Hall, 1971.

10. Salton G. and. McGill M. J., *Introduction to Modern Retrieval.* McGraw-Hill Book Company, 1983.

11. Rocchio, J. (1971). Relevance feedback in information retrieval, In G. Salton (Ed.), *The SMART Retrieval System-Experiments in Automatic Document Processing* (Chap 14). Englewood Cliffs, N J: Prentice Hall.