Exploring Query Expansion Strategies for MEDLINE

Padmini Srinivasan*
Department of Computer Science
Cornell University, Ithaca, NY 14853

June 22, 1995

Abstract

This paper evaluates the retrieval effectiveness of query expansion strategies on a MEDLINE test collection using Cornell University's Smart retrieval system. Two expansion strategies are tested on their ability to identify appropriate MeSH terms for user queries: expansion using an inter–field statistical thesaurus and expansion via pure retrieval feedback. These expansion strategies do not require prior relevance decisions. The study compares retrieval effectiveness using the original unexpanded and the alternative expanded user queries on a collection of 75 queries and 2,334 MEDLINE citations. Retrieval effectiveness is assessed using eleven point average precision scores (11-AvgP). Expansion by retrieval feedback gives the best improvement of 16.4% over a baseline performance of 0.5169 11-AvgP. Query expansion using the inter–field thesaurus gives a significant but lower performance improvement (9.9%) over the same baseline. This study recommends query expansion using retrieval feedback for adding MeSH search terms to a user's initial query.

1 Introduction

One problem apparent in query construction is that most user queries are rather small in size consisting of at most one natural language sentence. Since small query size is likely to handicap the search for relevant documents, query expansion is a vibrant research focus in information retrieval. This paper addresses a sub–problem of the overall query construction problem. We examine the more specific challenge of automatically constructing effective queries for MeSH indexes

*On sabbatical leave from the University of Iowa. The author thanks Professor Salton, Amit Singhal, Chris Buckley, and Mandar Mitra for their generous comments and kind hospitality.

of documents. MeSH concepts, known to be important for searching [9], have unique characteristics from the other searchable parts of MEDLINE documents such as titles and abstracts. Unfortunately end–users who are not trained in MeSH rules are unable to identify proper MeSH terms for their queries [8]. Therefore effective and automatic query construction strategies designed for MeSH have the potential to make this important tool more accessible to users.

Other researchers have also investigated the construction of MeSH based queries for MEDLINE searching. However our effort is unique in its focus on relevance independent expansion strategies. Strategies investigated here include: expansion via a statistical thesaurus and expansion via retrieval feedback. Proposed strategies are compared to each other as well as to appropriate control strategies.

1.1 Query Construction using Thesauri

There is a significant body of literature on thesauri and query design with some paying specific attention to MeSH. Hersh et al. test a novel approach of using the titles and abstracts of documents to extract concepts from the Metathesaurus component of the Unified Medical Language System (UMLS), a National Library of Medicine product [8]. The authors discover that SAPHIRE performs similar to both novice physicians and expert physicians. However, there are significant differences with the performance of librarians.

Aronson et al. [1] explore a different method of using the UMLS Metathesaurus for retrieval based on "underspecified syntactic analysis". They test their approach with Smart [2] as their retrieval system using a test collection of 3000 MEDLINE documents and 150 queries in three subject areas. The authors report an increase of 4% in average precision when compared with retrieval using unmodified queries and documents.

Yang and Chute investigate MeSH query construction as part of a larger investigation to solve for vocabulary differences between the queries and the documents [17, 18, 16]. They use mappings from query words to MeSH concepts for query expansion. Test queries expanded using these learned mappings give improved retrieval performance. When applied to a reduced version of the MEDLINE collection used in the study by Hersh et al. [8] as well as in this study, Yang and Chute observe a 32.2% improvement in average precision over their baseline of 0.412. The Yang and Chute study certainly emphasizes the observation that MeSH terms are important for retrieval. A practical disadvantage of their approach is its unrealistic reliance on the availability of a training set of queries and relevant documents, which limits the applicability of their method.

In summary, query construction efforts for MEDLINE using the UMLS Metathesaurus have not yet yielded positive returns. Successful statistical mapping functions unfortunately rely on training relevance data and expect a high degree of similarity between the new query and those of the training set. Our first objective is to continue the exploration of statistical thesauri. However in contrast to the Yang and Chute work, the thesaurus we propose does not require any prior relevance decisions. Thus there is no extra cognitive load on the current user or any reliance on prior users. We suggest that statistical query expansion approaches tied to strong retrieval models continue to have significant potential for MEDLINE.

1.1.1 Structure of the Statistical Thesaurus

Different types of statistical thesauri have been explored for information retrieval [4, 5, 12, 15, 19] with some efforts being reasonably successful. The straightforward approach is to construct a thesaurus using information regarding the co–occurrence of terms in the documents. However many variations in methodology exist within this general model. The thesaurus proposed here specifies associations across two fields of a document: title–abstract and MeSH. These links are identified based on the connected entities' co–occurrence patterns in the database, and are therefore, independent of relevance judgements. The free–text title–abstract mappings will be used to identify the MeSH terms associated to a user's original free–text query.

1.2 Query Expansion using Retrieval Feedback

Our second expansion technique derives from a well established query expansion technique called relevance feedback [3, 10, 11, 14]. In relevance feedback, a 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 Rocchio's [11] and Ide's [10] have been studied within this broad strategy.

In Ide's method, the weight for a term t_i in the old query is modified as:

$$wt_{Qnew}t_i = alpha * wt_{Qold}t_i + beta * \sum_{k=1}^{R}(wt_{Doc_k}t_i) - gamma * wt_{Doc_{NR_1}}t_i \tag{1}$$

where alpha represents the relative importance of term t_i in the original query. Beta and gamma are parameters designating the relative importance of relevant vs non–relevant information for query expansion. R is the number of relevant documents and NR_1 is the top–most non–relevant document retrieved.

Retrieval feedback is a derivative of relevance feedback [6]. Here an initial retrieval run is conducted, and the top few documents retrieved are assumed to be relevant and are used for query expansion. Therefore in the feedback equation above, the negated term drops off and R now specifies the number of documents considered from the initial run for feedback.

Retrieval feedback works very well in the full–text TREC experimental environment producing improvements as high as 20% as shown by the Cornell group using Smart [6]. However, the technique remains untested for bibliographic databases such as MEDLINE which differ significantly from the TREC collection. Thus our second objective is to test query expansion using retrieval feedback. This technique will also be compared to the previous expansion method using the thesaurus. The comparison will tell us if relevance independent query expansion based on inter–field co–occurrence patterns derived from the entire database differs from relevance independent query expansion constrained to the local context of a single query.

2 Methods

These experiments were run using Cornell's SMART retrieval system [13], a sophisticated and powerful research system based on the vector space model and designed for testing ideas pertaining to information retrieval. The measure of performance used here is the standard 11 point average precision or 11-AvgP.

A test collection of 75 queries and 2,344 MEDLINE documents available for retrieval research is used in this study. It was produced by Haynes and colleagues [7] in a study evaluating MEDLINE. All queries have some relevant documents in the collection. All documents include abstracts.

2.1 Indexing Strategies

In Smart, documents and queries are automatically indexed to yield a weighted vector of index terms (words or phrases), as shown below for a document D_1. Term weights reflect the relative importance of terms in content representation of the document.

$$D_1 = (wt_{D_1}t_1, wt_{D_1}t_2, ..., wt_{D_1}t_m)$$

where $wt_{D_1}t_i$ represents the weight of term t_i in document D_1 and an indexing vocabulary of size m is assumed. Smart conducts retrieval by computing the similarity as the vector inner product of the document and the query vectors resulting in a ranked list of documents for a given query.

Word based index vectors are used in this study. In order to create the statistical thesauri mapping from the title–abstract vocabulary to the MeSH vocabulary, we derive two index vectors for each document. A vector from the non trivial words in the title and abstract (ta–vector) and a vector from the non trivial words of the MeSH concepts (m–vector). Assuming a title–abstract vocabulary of p words and a MeSH vocabulary of q words[1], a document may be represented as:

$$D_1 = (wt_{D_1}w_1, wt_{D_1}w_2, ..., wt_{D_1}w_p),\ (wt_{D_1}c_1, wt_{D_1}c_2, ...wt_{D_1}c_q) \quad (2)$$

We generate a single ta–vector for each query, as we consider the user's initial free–text query more suitable for searching the title and abstract field.

$$Q_1 = (wt_{Q1}w_1, wt_{Q1}w_2, ..., wt_{Q1}w_p)$$

Our query expansion strategies add an m–vector to each query representation. Retrieval using this expanded query is a weighted sum of the inner vector products of corresponding vectors in documents and queries.

Similarity(D, Q) = delta x similarity (ta–vectors) + similarity (m–vectors) (3)

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

Smart allows a wide variety of strategies for computing term weights. Each strategy is represented by a triple: ABC. A represents the term frequency component, i.e., the number of times a term occurs in the document. B represents the inverse document frequency component which increases with the rarity of the term in the database. C represents the normalization component

[1] For ease of distinction we will henceforth refer to words of the titles and abstracts as *words* and words of the MeSH concepts as *concepts*.

Dimension	Symbol	Interpretation
A	b: binary	1 if term is present and 0 if term is absent in document
A	a: augmented	$(0.5 + 0.5 * tf/max\_tf\_in\_document)$
A	l: logarithmic	1 + ln(tf)
A	n: none	tf
B	t	ln(N/n); N = # docs. in database; n = # docs. with term.
B	n	idf not used
C	c: cosine	$\sqrt{w_{D_1}c_1^2 + \ldots + w_{D_1}c_m^2}$; m = vocabulary size
C	n	no normalization

Table 1: Term Weighting Strategies in Smart.

Retrieval Strategy	11-pt. Average Precision
atn.atc	0.5169
atn.ntc	0.5158
ann.atc	0.5149
ann.ntc	0.5130
atc.atc	0.5027
atc.ntc	0.5016

Table 2: Baseline Performance of Control Retrieval Strategies.

for the length of the document. Options are available for each of these three dimensions of the triple as shown in Table 1.

A retrieval strategy is represented by a pair of triples, eg., atc.atn, where the first and the second triples represent the document and the query weighting strategies respectively. Given the indexing options in Table 1, there are 256 (16 x 16) retrieval strategies that may be explored within Smart. The six strategies of Table 2 serve as our control strategies as they provide the best baseline performance. For this baseline study, document–query similarity is computed as the inner vector product of their ta–vectors alone.

2.2 Query Expansion

Given a user's initial query Q_1 represented by a single ta–vector:

$$Q_1 = (wt_{Q_1}w_1, wt_{Q_1}w_2, \ldots, wt_{Q_1}w_p) \tag{4}$$

our objective is to add an appropriate m–vector which transforms Q_1 into Q_2:

$$Q_2 = (wt_{Q_2}w_1, wt_{Q_2}w_2, \ldots, wt_{Q_2}w_p),\ (wt_{Q_2}c_1, wt_{Q_2}c_2, \ldots, wt_{Q_2}c_q). \tag{5}$$

Note that the ta–vectors for Q_1 and are Q_2 are always identical in this study. In other words only MeSH concepts are added during query expansion.

2.2.1 Expansion via Statistical Thesauri

Given the indexing vector for a document, as shown in Equation 2, the association strength between a word w_i and a concept c_j may be measured as the product of their weights in the document. Therefore the net association strength between w_i and c_j in the database,

$$Net\ Association(w_i, c_j) \ = \sum_{K=1}^{M} (wt_{Doc_K} w_i * wt_{Doc_K} c_j)/M$$

where M is the number of documents in the collection in which they co–occur. Two different thesauri were created in this study: TH(atn) and TH(ntn) using atn and ntn document indexing strategies, respectively.

Given a thesaurus, we need query expansion strategies to support the query transformation shown from Equation 4 to Equation 5. There are two steps to undertake for this transformation. First appropriate concepts have to be selected for the query from the thesaurus. Second weights have to be computed for these new concepts.

Concept selection begins by considering all concepts in the thesaurus that are associated with at least one of the words in the query's ta–vector. Each of these concepts is ranked by its overall association with the query, computed as the sum of its associations with individual query words. That is, for a concept c_y,

$$Assocation(Q_1, c_y) = \sum_{j=1}^{r} Net\, Association(w_j, c_y) \tag{6}$$

where r is the number of words in the query's ta–vector.

Concepts are selected from the top of this ranked list based on a parameter specifying the number of new MeSH concepts to add to queries. Next, the weight for each added concept is computed. Two strategies have been designed for this. In both strategies

$$wt_{Q_2} c_y = wt'_{Q_2} c_y / \sqrt{wt'_{Q_2} c_1^2 + wt'_{Q_2} c_2^2 + ... + wt'_{Q_2} c_m^2}$$

where it is assumed that an m–vector of size m is added.

In weighting strategy 1 (W1),

Retrieval Run	Baseline 11-AvgP	Query Expansion 11-AvgP (% increase)	M–vector Size	Delta
atn.atc.TH(atn).W1	0.5169	0.5620 (+8.7%)	15	1.3
atn.atc.TH(ntn).W1	0.5169	**0.5681 (+9.9%)**	15	1.3
atn.atc.TH(atn).W2	0.5169	0.5605 (+8.4%)	10	1.3
atn.atc.TH(ntn).W2	0.5169	0.5650 (+9.3%)	5	1.3
atn.ntc.TH(atn).W1	0.5158	0.5587 (+8.3%)	15	1.3
atn.ntc.TH(ntn).W1	0.5158	0.5643 (+9.4%)	10	1.3
atn.ntc.TH(atn).W2	0.5158	0.5570 (+8.0%)	10	1.3
atn.ntc.TH(ntn).W2	0.5158	0.5617 (+8.9%)	5	1.3
ann.atc.TH(atn).W1	0.5149	0.5576 (+8.3%)	5	0.8
ann.atc.TH(ntn).W1	0.5149	0.5576 (+8.3%)	5	0.8
ann.atc.TH(atn).W2	0.5149	0.5590 (+8.6%)	10	0.8
ann.atc.TH(ntn).W2	0.5149	0.5582 (+8.4%)	5	2
ann.ntc.TH(atn).W2	0.5130	0.5577 (+8.7%)	10	1.3
atc.atc.TH(ntn).W1	0.5027	0.5560 (+10.6%)	10	2

Table 3: Performance of Query Expansion via Statistical Thesauri.

$$wt'_{Q_2} c_y = \sum_{j=1}^{r} Net\,Association(w_j, c_y)$$

while in weighting strategy 2 (W2),

$$wt_{Q_2} c_y = \sum_{j=1}^{r} wt_{Q_1} w_j * Net\,Association(w_j, c_y).$$

We test the four combinations of thesauri and weighting strategies under each of the six control retrieval strategies of Table 2. This gives us 24 different retrieval runs. During query expansion, the parameter controlling the number of MeSH concepts to be added varied from 5 to 20 concepts in steps of 5. Finally, during retrieval the relative emphasis on ta–vectors and m–vectors also varies as the controlling parameter delta (see Equation 3) varied from 2, 1.75, 1.5, 1.3, 1, 0.8, 0.66, and 0.57.

Table 3 provides the best performance data for the top 14 retrieval runs. (All tables present performance data averaged over all 75 queries). A row representing a retrieval run is designated by a combination of a retrieval strategy (the first two triples) then the thesaurus identifier followed by the weighting strategy.

2.2.2 Expansion via Retrieval Feedback

In retrieval feedback, the goal is again to select concepts for the new m–vector and compute their weights.

All MeSH concepts in the initially retrieved documents of a query are considered to be candidates for addition to the new m–vector. The first term in Equation 1, representing Ide's method, drops

Retrieval Strategy	Baseline 11-AvgP	Query Expansion 11-AvgP (% increase)	Size Initial Set	M–vector Size	Delta
atn.atc	0.5169	**0.6018 (+16.4%)**	10	20	0.66
atn.ntc	0.5158	0.6011 (+16.5%)	10	20	0.57
ann.atc	0.5149	0.5745 (+11.6%)	10	10	2
ann.ntc	0.5130	0.5725 (+11.6%)	10	10	2
atc.atc	0.5027	0.5935 (+18.1%)	5	10	0.8
atc.ntc	0.5016	0.5935 (+18.3%)	5	20	0.8

Table 4: Performance of Query Expansion via Retrieval Feedback.

off since there is no original m–vector for the query. Similarly gamma has no role since all retrieved documents are marked relevant. Consequently we get,

$$wt_{Q_2} c_y = \sum_{K=1}^{R} (wt_{Doc_K} c_y) \tag{7}$$

MeSH concepts in the initially retrieved documents are ranked by the measure in Equation 7 and concepts are selected from the top of this list according to a parameter governing size of the new m–vector. The final weight of an added concept is also specified by the same equation. Table 4 provides the best performance data for this expansion technique applied to the six baseline retrieval strategies of Table 1.

3 Results

Our first objective is to assess the potential of inter–field statistical thesauri for query expansion. Table 3 shows that improvements between 8.0% and 10.6% result from this query expansion technique. The best strategy atn.ntc with TH(ntn) and the W1 weighting strategy yields a 11-AvgP of 0.5681. However the improvements and 11-AvgP values across all strategies are very similar. Expansion using TH(ntn) produces an average improvement of 9.2% across all retrieval runs while the TH(atn) thesaurus averages at 8.78%. W1 and W2 retrieval runs average 9.13% and 8.88% improvements, again revealing no significant differences. For most of the retrieval runs the best improvements are achieved when the original ta–vector is viewed as at least 1.3 times as important as the newly added m–vector. Also two–thirds of the runs add at least 10 concepts to the indexing for best performance.

Table 4 displays the results for our test of retrieval feedback for query expansion. The best

Retrieval Strategy	Baseline 11-AvgP	Thesaurus 11-AvgP(% increase)	Retrieval 11-AvgP (% increase)
atn.atc	0.5169	**0.5681 (+9.9)**	**0.6018 (+16.4%)**
atn.ntc	0.5158	0.5643 (+9.4)	0.6011 (+16.5%)
ann.atc	0.5149	0.5592 (+8.6)	0.5745 (+11.6%)
ann.ntc	0.5130	0.5576 (+8.7)	0.5725 (+11.6%)
atc.atc	0.5027	0.5560 (+10.6)	0.5935 (+18.1%)
atc.ntc	0.5016	0.5548 (+10.6)	0.5936 (+18.3%)

Table 5: Comparison of Different Query Expansion Strategies.

improvement (18.3%) is seen for atc.ntc while the highest 11-AvgP value (0.6018) is achieved by atn.atc. Interestingly the original ta–vectors are only 0.66 times as important as the added m–vectors in the best retrieval runs. Finally the lowest 11-AvgP (0.5725) is higher than the best value achieved when expanding the query using the thesaurus (0.5620). This indicates that retrieval feedback offers a more effective route for query expansion than our statistical thesaurus.

Table 5 focuses on the best performance of the two techniques studied here for each of the six control retrieval strategies of our study. First this table indicates that it is feasible to effectively add MeSH concepts to an original query and expand the search to include the MeSH field. Table 5 also shows that although the thesaurus method produces what may be significant improvements, these are overshadowed by retrieval feedback. It appears reasonable to conclude that the improvements obtained by the thesaurus are almost completely subsumed by the improvements obtained by retrieval feedback. Since the thesaurus generation and utilization processess are more complex, this study recommends the simpler retrieval feedback method as the optimal approach for query expansion.

3.1 Comparison with Previous Research on MEDLINE Query Construction

Some comparisons of our results with the results of [1, 6, 16] are possible. Key methodological differences across studies, especially with regards to measures used are identified where appropriate.

The NLM study [1], employs the Smart system to compare retrieval using original queries and documents with retrieval using modified queries and documents. The improvements obtained here (10% to 18.3%) over our best baselines are much higher than the 4% improvement they report.

Comparison with SAPHIRE's performance poses some problems in terms of measurement units. We derive a 11-AvgP score of 0.3549 from their table 2 (page 56) which is substantially lower than any of our baseline strategies. However methodological variances in measures used suggest that the difference in performance computed is a somewhat tentative indicator of the relative merits of the compared approaches.

Comparison with Yang's work [16] is also a bit challenging. Although they use the same MEDLINE test database, they test their technique only on a subset of this MEDLINE test database. The remaining portion is used as a training set to create the mappings. Second their performance measure is the average of precision scores over a 10 point recall scale (0.1, 0.2, ..., 1.0). Hence we first convert our 11-AvgP scores to 10-AvgP for compatibility. This conversion does not change any of our results in terms of the relative ranking of techniques. However, it does bring down our average precision scores, since the 10-AvgP measure ignores the ranks of relevant documents retrieved before the 0.1 recall point. Interestingly our worst baseline strategy (atc.ntc) has a 10-AvgP of 0.4815 which is 16.9% higher than their baseline of 0.412. Using learned mappings their query construction method produces a final 10-AvgP of 0.545 [16]. Our best performance using query expansion via retrieval feedback yields a 11-AvgP of 0.6018 equivalent to a 10-AvgP of 0.586. This shows a 7.5% improvement over their approach, an improvement achieved without using any thesaurus or any learned mappings across vocabularies.

In summary, our results show that query expansion to add MeSH concepts to an initial free–text query may be done automatically and effectively. More importantly, this process can be relevance independent. That is, it need not involve user relevance decisions. Thus the user is not given the added responsibility of suggesting relevant documents. Also the current user can operate independently of previous users of the system. In other words, success in the construction of a query can be independent of relevant items retrieved previously.

4 Discussion

Query expansion strategies are needed to improve users' original queries to search MEDLINE. In this research, original queries were used to search the title and abstract field while terms for searching the MeSH field of documents, were identified using different expansion strategies. The strategies explored indicate that both an inter–field statistical thesaurus as well as retrieval feedback based query expansion improve performance significantly even over a high baseline of 0.5169. Previous studies have also shown improvements but on much lower baselines.

References

[1] Aronson A, Rindflesch T and Browne A. Exploiting a large thesaurus for information retrieval. RIAO-94, 197-216.

[2] Buckley C, Salton G. Improving Retrieval Performance by Relevance Feedback. Technical Report 88-898. Department of Computer Science, Cornell University, Ithaca, NY 14853-7501. February 1988.

[3] Buckley C, Salton G. Optimization of Relevance Feedback Weights. SIGIR 1995.

[4] Crouch C. An approach to the automatic construction of global thesauri, IP&M, 26(5),629-640, 1990.

[5] Crouch C. and Yang B. Experiments in automatic statistical thesaurus construction. SIGIR 1992, 77-88, 1992.

[6] Harman D. Overview of the Third Text REtrieval Converence (TREC-3). Proceedings of Text Retrieval Conference 3, 1994.

[7] Haynes R, McKibbon K, Walker C, Ryan N, Fitzgerald D, Ramsden M. Online access to MEDLINE in clinical settings. Ann Intern Med. 1990;112:78-84.

[8] Hersh W, Hickam D, Haynes R, McKibbon K. A Performance and Failure Analysis of SAPHIRE with a MEDLINE Test Collection. JAMIA 1(1);51-60, Jan/Feb 1994.

[9] Hersh W, Buckley C, Leone T, and Hickam D. OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. SIGIR 1994. 192-200.

[10] Ide E. New experiments in relevance feedback. In Gerard Salton, editor, The SMART Retrieval System-Experiments in Automatic Document Processing, chapter 16. Prentice Hall, Englewood Cliffs, NJ, 1971.

[11] Rocchio J. Relevance feedback in information retrieval. In Gerard Salton, editor, The SMART Retrieval System-Experiments in Automatic Document Processing, chapter 14. Prentice Hall, Englewood Cliffs, NJ, 1971.

[12] Salton G. automatic information organization and retrieval, McGraw-Hill Book Company, 1968.

[13] Salton G. (Editor). The SMART Retrieval System-Experiments in Automatic Document Processing, Prentice Hall, Englewood Cliffs, NJ, 1971.

[14] Salton G. and Buckley C. Improving retrieval performance by relevance feedback. JASIS, 41(4):288-297,1990.

[15] Sparck Jones K. and Needham R. Automatic term classification and retrieval, IP&M, 4:91-100,1968.

[16] Yang Y. Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval. SIGIR 94, 13-22. Ireland

[17] Yang Y and Chute C. Words or concepts: the features of indexing units and their optimal use in information retrieval. SCAMC 93, 685-689.

[18] Yang Y and Chute C. An Application of Least Squares Fit Mapping to Text. SIGIR 93, 281-290. Pittsburgh

[19] Yufeng J and Croft B. An association thesaurus for information retrieval. RIAO 94. 146:160.