Leveraging medical thesauri and physician feedback for improving medical literature retrieval for case queries

Parikshit Sondhi,1 Jimeng Sun,2 ChengXiang Zhai,1 Robert Sorrentino,2 Martin S Kohn2

ABSTRACT

Objective This paper presents a study of methods for medical literature retrieval for case queries, in which the goal is to retrieve literature articles similar to a given patient case. In particular, it focuses on analyzing the performance of state-of-the-art general retrieval methods and improving them by the use of medical thesauri and physician feedback.

Materials and Methods The Kullback–Leibler divergence retrieval model with Dirichlet smoothing is used as the state-of-the-art general retrieval method. Pseudorelevance feedback and term weighing methods are proposed by leveraging MeSH and UMLS thesauri. Evaluation is performed on a test collection recently created for the ImageCLEF medical case retrieval challenge.

Results Experimental results show that a well-tuned state-of-the-art general retrieval model achieves a mean average precision of 0.2754, but the performance can be improved by over 40% to 0.3980, through the proposed methods.

Discussion The results over the ImageCLEF test collection, which is currently the best collection available for the task, are encouraging. There are, however, limitations due to small evaluation set size. The analysis shows that further refinement of the methods is necessary before they can be really useful in a clinical setting.

Conclusion Medical case-based literature retrieval is a critical search application that presents a number of unique challenges. This analysis shows that the state-of-the-art general retrieval models are reasonably good for the task, but the performance can be significantly improved by developing new task-specific retrieval models that incorporate medical thesauri and physician feedback.

In this paper we present our work on finding relevant full-text articles from the literature in response to a medical case query submitted by a healthcare professional. We call this task case-based document retrieval. There are important uses for such capabilities, both educationally and in the clinical setting. In current evidence-based training programmes, medical students often find it useful to view information from teaching files containing both detailed case histories and outcomes, as well as test results and imaging studies. This allows individual physicians to enhance their diagnostic skills through self-paced learning. In clinical practice, the capability to review case files and related medical articles from curated collections assembled by prominent clinical institutions may significantly improve physicians’ ability to make a diagnosis in complex or puzzling cases.

From a technical perspective, case queries are often long unstructured natural language text, containing arbitrary combinations of patient background information, symptoms, test results or diagnosis information, etc. A set of medical images such as x-ray scans may also be included to provide additional information. Figure 1 shows two examples from the ImageCLEF 2010 case retrieval dataset.

These are in contrast with general informational queries in the biomedical domain (eg, the query: ‘review article on cholesterol emboli’), which tend to be shorter and more open in their scope. They also differ from typical queries used in genomics information retrieval in which the goal is often to answer a particular question (eg, finding the molecular function of a gene).

The problem of medical literature retrieval for case queries has gathered interest due to the wide acceptance of evidence-based medicine, and interest in the development of medical case-based reasoning systems. To promote further research in the area, the ImageCLEF medical case retrieval task was introduced in 2009 and continued in 2010 and 2011. It contributed to the development of a standardized test collection for the problem. Based on the performance of participating systems, it has been observed that text-based approaches, based on standard document retrieval algorithms, tend to outperform multimodal methods, due to the poor performance of image content-based retrieval. It is thus reasonable to expect that performance could be improved if the standard text retrieval methods were tailored specifically for handling case queries.

Our main contribution in this paper is to evaluate the utility of general retrieval methods for the task and develop new state-of-the-art text retrieval approaches specific to case queries. Our results show that while well-tuned general retrieval methods work reasonably well, they have several limitations. In particular, the problems of vocabulary gap (ie, a user may use a different vocabulary in the query than that used by authors of relevant documents to describe the same concept), and non-optimal query term weighting (ie, failing to differentiate the important query keywords from the less important ones). To address these limitations, we propose to extend a general retrieval method by
performing query term reweighting based on the Unified Medical Language System (UMLS) thesaurus and pseudorelevance feedback based on the Medical Subject Headings (MeSH) thesaurus; both were found to improve retrieval accuracy. Additional information from physicians in the form of related query keywords was also found to be helpful, while relevance feedback improved performance moderately. Our system based on these strategies achieved the best performance in the ImageCLEF medical case retrieval challenge 2010. In this paper, we refine our methods and report more comprehensive results on the ImageCLEF dataset.

**BACKGROUND**

**Related work**

The problem of information retrieval for biomedical literature has gathered tremendous interest since 2003, when the TREC Genomics Track was introduced. Research has focused on both utilizing general retrieval methods and leveraging semantic resources for the problem. Siadaty et al. present a high precision retrieval method using sentence-level co-occurrence of query terms to generate relevance scores. Can and Baykal present MedicoPort, a biomedical search engine that uses TF-IDF-based ranking and incorporates semantic knowledge in the form of the UMLS metathesaurus. Lu et al. found TF-IDF-based methods to be better than sentence-level co-occurrence-based methods using the TREC Genomics Track dataset. Huang and Hu present a Bayesian learning method for achieving result diversity in biomedical information retrieval.

Query expansion techniques, including relevance feedback and pseudorelevance feedback, have been widely used. Lu offers a survey of multiple biomedical search systems that use these feedback techniques. Relevance feedback involves expanding the search query with relevant keywords from the retrieved relevant documents. Pseudorelevance feedback expands the query with additional keywords from top-ranked documents from an initial search. In both cases, results are re-ranked by searching again with the expanded query. While both methods are generally known to improve performance, the latter is more popular as it does not require labeled relevant documents. Other techniques for query expansion include using ontology (e.g., Bhogal et al.) or global corpus analysis (e.g., Xu and Croft). Previous work on feedback and query expansion has not been tailored for dealing with long and complex case queries, which is what we study.

PubMed maintained by the National Library of Medicine is a database of biomedical research articles containing over 21 million citations. PubMed also runs a clinical query search service that allows users to refine search results in specific clinical research areas. Their approach differs from ours in that the refinement is achieved by pre-categorizing the documents in the collection to different clinical research areas by the use of study-type filters. They do not attempt to develop case query-specific methods.

Another related and more challenging task is biomedical question answering, in which the goal is to retrieve precise answers to natural language biomedical questions instead of retrieving entire documents as in our task. In case-based document retrieval, research has focused on developing multimodal approaches, which utilize both text and image similarity to solve the problem. Related query keywords was also found to be helpful, while relevance feedback improved performance moderately. Our system based on these strategies achieved the best performance in the ImageCLEF medical case retrieval challenge 2010. In this paper, we refine our methods and report more comprehensive results on the ImageCLEF dataset.

**Dataset description**

The ImageCLEF dataset comprises 5585 research articles from the Radiological Society of North America Radiographics journal (http://radiographics.rsna.org/) and 1 of 19 case queries with relevance judgments provided. Five of the 19 queries are from the 2009 task and 1 from the 2010 task. The case queries were formulated based on cases from the teaching file Casimage. This teaching file contains cases (including images) from radiological practices that are used by clinicians mainly for teaching purposes. Any information regarding diagnoses and treatments was removed from these queries, so as to simulate the situation of the clinician who has to diagnose the patient. For the judging process, however, the relevance judgements (expert
in the field) were provided with complete information in order to perform accurate evaluation. Additional details are available in Müller et al.3 Some sample queries are shown in supplementary appendix I (available online only). We used the five queries from the 2009 task for parameter tuning and the remaining 14 queries from 2010 for evaluation. The evaluation set contained on average 37.2 relevant documents per query (min 2, max 97, SD 31.01).

Design objectives
As much of the task setup in case retrieval is similar to general retrieval, we hypothesize that the state-of-the-art general retrieval models may achieve reasonable performance. On the other hand, domain/task-specific limitations must be addressed by using semantic resources and getting additional information from the users. Based on this intuition, the main objectives of our experiments are to test the following hypotheses:
2. H2: Performance can be improved by systematically addressing limitations of the state-of-the-art methods, by the use of medical thesauri.
3. H3: Performance can be improved by user feedback, in the form of additional query keywords and relevance judgments.

We began by analyzing the performance of one of the best performing general retrieval methods: Kullback–Leibler (KL)-divergence retrieval model with Dirichlet smoothing and pseudorelevance feedback41 to test H1. To the best of our knowledge, it has not been applied to the case-based document retrieval task before. Then using it as a baseline, we developed additional methods to test H2 and H3.

METHODS
In this section we start with a discussion of the standard retrieval model that forms our baseline. We then discuss our proposed task-specific approaches based on thesauri and physician feedback. Our strategy will be to use the baseline search method as a black box (for searching full-text articles) and implement additional methods on top of the baseline method, by either appropriate expansion and weighing of queries or re-ranking of results.

Standard retrieval models

KL-divergence retrieval model with Dirichlet smoothing
Language modeling41 provides a sound statistical framework for designing retrieval models. One of the best-performing retrieval models based on language modeling is the KL-divergence retrieval model.41 Given a query Q and a document D, this model would first estimate a unigram query language model θQ (ie, a word distribution) based on a given query and a document language model θD for document D, and then score the document D with respect to query Q based on negative KL-divergence between the two language models, −D(θQ || θD), defined below:

\[ -D(\theta_Q \parallel \theta_D) = - \sum_{w \in V} p(w|\theta_Q) \log \frac{p(w|\theta_Q)}{p(w|\theta_D)} \]

where V is the set of words in our vocabulary, and p(w|θQ) and p(w|θD) are the probabilities of word w given by the two language models, respectively. The negative KL-divergence intuitively measures the similarity of the query language model and the document language model. It would thus favor a document that matches more query words.

The document language model θD characterized as the word distribution \( p(w|\theta_D) \) is usually estimated using Dirichlet prior smoothing41:

\[ p(w|\theta_D) = \frac{c(w,D) + \mu p(w|C)}{|D| + \mu} \]

where \( c(w,D) \) is the count of word w in document D, \( p(w|C) \) is a background/reference language model estimated based on all the documents in the collection and helps providing probabilities for words unseen in a document, and \( \mu \) is a smoothing parameter, which was tuned using training data. The optimal value was found to be \( \mu = 4800 \).

The simplest way to estimate the query model \( \theta_Q \) is to set \( p(w|\theta_Q) \) to the relative frequency of a word in the query:

\[ p(w|\theta_Q) = \frac{c(w,Q)}{\sum_{w \in Q} c(w,Q)} \]

As this approach assigns zero probability to words not in the query, a potentially better way to estimate this model is to use a technique called pseudorelevance feedback, which we discuss next.

Pseudorelevance feedback
The basic idea of pseudorelevance feedback is to treat a small number of top-ranked documents in the initial retrieval result as if they were relevant documents and extract useful terms from these feedback documents to improve the estimate of a query language model. In our experiments, we used the mixture model approach described in Zhai and Laﬀerty,42 which is one of the best-performing approaches to pseudorelevance feedback. This method first obtains a word distribution characterizing the topic in the feedback documents. Then, it interpolates that word distribution from the feedback documents with the word distribution estimated based on the relative frequency of words in the query. The mixture model pseudorelevance feedback method has a few parameters, which were tuned using the five queries from the ImageCLEF 2009 dataset. The best results were found when using only the top two documents for feedback.

Thesaurus-based approaches
We propose two approaches for exploiting available thesauri in the medical domain to improve the general retrieval model.

Semantic query weighing
Not all case query keywords are equally useful in identifying relevant documents. General retrieval uses inverse document frequency45 as a critical heuristic for weighing keywords, which assumes that it is more important to match a rare term than a frequent term. However, this general heuristic is insufficient for our task because keywords belonging to certain semantic categories such as disease names, symptoms, drugs etc (see figure 2) are more representative of a medical case and must be assigned high weights regardless of their inverse document frequency.

To this end, we propose to map all query keywords to UMLS15 semantic types using the MMTx toolkit39 and assign weights according to their semantic types. Based on our analysis of all the semantic types appearing in the training queries, the following were among the most discriminative for retrieving a relevant case:
Our second approach.

Top N documents are relevant. We overcome this limitation in the top N documents and may not perform well if none of the

1. Make a list of potential diagnoses for the case query at hand.
2. Assign a high relevance score to all documents discussing these diagnoses.

Assuming the case query is sufficiently descriptive, there would only be a small number of potential diagnoses possible. If we can predict these conditions and increase the rank of the documents that primarily talk about them, we should be able to improve performance. This breaks down into two problems:
1. How to find out which conditions a given document talks about?
2. How to find out what conditions the query case is likely to represent?

This is a harder problem. In the following discussion, we present two ways of dealing with it.

Top-N-based MeSH feedback
This approach is similar to pseudorelevance feedback. We make a list of all condition-related MeSH terms present in the top N=10 documents in the initial ranked list generated by the baseline method. We then slightly reduce the weight of any documents below these top N that do not share any MeSH terms with this list (see figure 3).

The approach does have limitations in that it cannot re-rank the top N documents and may not perform well if none of the top N documents are relevant. We overcome this limitation in our second approach.

Distribution-based MeSH feedback
This method is based on the intuition that a MeSH term assigned to a document that contains a large number of query keywords is more likely to represent the query. Therefore, for each MeSH term, we first identify all the documents indexed with it, and then count the number of unique query keywords present in these documents. Finally we pick the 25 highest scoring MeSH terms for our filtration list.

This approach is useful in that it can also allow us to re-rank the top results. This becomes important when a high precision is required. A more detailed algorithm is provided in supplementary appendix II (available online only).

Note that for both MeSH-based approaches, we do not take into account the hierarchical relationships between concepts. Incorporating the hierarchy will introduce new parameters that may lead to overfitting on our small training set.

Physician feedback
In an interactive retrieval system, a physician can potentially provide feedback in two ways: (1) additional keywords can be provided to focus the query better; (2) relevance judgments on retrieval results can be provided as examples for the system to use for better inference of relevance. We now discuss how we can leverage these two different forms of feedback.

Additional query keywords
While submitting a case query, physician users often have a number of additional keywords in mind that they feel are relevant to the case, but they avoid using these keywords as they may influence the results too strongly. However, if used appropriately, these keywords can be very useful in retrieving the right documents. To examine this hypothesis, we asked two physicians to look at each case query along with its associated images (without looking at any relevant documents) and provide additional keywords they thought were relevant. These were then merged to form the additional physician keywords for each query. An example is shown in figure 4.

We added them to the original query with comparatively low weights to keep them from dominating the results (ie, their query count was modified as $c'(w,Q) = c(w,Q) * 0.3$). This helped greatly in improving performance. Adding these keywords directly to the query with equal weights did not do as well.
Relevance feedback
To leverage feedback information in an interactive retrieval system, we also experimented with physician relevance feedback. The idea was to ask a physician to judge 20 top-ranked documents as relevant or non-relevant and then use these relevant documents for relevance feedback, ie, the system would learn from these examples to improve retrieval performance. More specifically, our system can use the same mixture model that we had used for pseudorelevance feedback to improve the estimation of the query language model based on judged relevant documents.

EXPERIMENTS AND RESULTS
Experimental design
Our first set of experiments was to evaluate the performance of the baseline retrieval method using the standard implementation provided in the LEMUR retrieval toolkit (http://www.lemurproject.org/). This gave an estimate on the best performance that general retrieval methods can achieve for this task. Subsequently, we experimented by adding our proposed methods. Figure 5 provides a summary of all the different experiments. The light colored boxes show experiments in which performance was improved by adding a new method on top of all parents. The dark-colored boxes show experiments in which the performance dropped. More details are available in table 2.

Evaluation criteria
The performance of each method is measured using precision at 10 (P@10), recall at 30 (R@30) and mean average precision (MAP). P@10 represents the percentage of relevant documents in the top 10 results. R@30 represents the percentage of all the relevant documents in our collection that are present in the top 30 results. MAP is the arithmetic mean of average precision values over a set of queries. Suppose, for some ranking ri and the MAP over a set of rankings R are then given by:

\[
AP(ri) = \sum_{j=1}^{k_i} \frac{P(rank(j))}{k_i}; \quad MAP(R) = \frac{\sum_{r_i \in R} AP(r_i)}{|R|}
\]

AP intuitively captures the average of precision at every point when a new relevant document is retrieved.

Results
Standard retrieval method
The query-specific performance results are shown in table 1. As users are more likely to formulate a new query rather than search for relevant documents beyond the third page or so, R@30 helps give a good estimate on practical recall. The P@10 varies from 0 to 0.9, with a mean of 0.4286 (SD 0.26). Therefore, on average, the baseline shows approximately four relevant documents in the top 10. This suggests that our first hypothesis of baseline performance being reasonable holds. We did, however, notice a fairly high variation in the performance values. We present a detailed failure analysis of the baseline and some of our other approaches in the discussion section.

Thesaurus-based methods
A performance comparison of the semantic weighing and MeSH-based methods with the baseline is shown in table 2. We observe that semantic weighing (run O1) generally improves
performance across all measures. The improvement in recall is the most pronounced.

Of the two MeSH-based approaches, the top-N approach (run O2) with N=10 and penalty factor set at 0.1 (tuned using ImageCLEF 2009’s five queries) improves MAP by 2.5%. The P@10 does not change, as the top 10 documents are not re-ranked in this case. On combining with the semantic weighing method (run O4), we get the best thesaurus-based run in terms of MAP, with an improvement of 6.8%.

The distribution-based MeSH feedback approach (run O3) performs worse than the baseline. However, it starts to show improvement when combined with other methods (runs O5, K4, K6, R5). This approach also re-ranks the top 10 documents, and when combining it with semantic weighing (run O5), we observe the best performance in terms of P@10 among all the thesaurus runs.

**Physician feedback**

**Additional query keywords**

The results in table 2 show that using additional keywords from physicians (run K1) generally improved performance. The performance improved for nine cases and was hurt in three. The magnitude of improvement was higher (40.1%) compared with the previous methods. It thus suggests that a better way to submit a case query is not only to provide the natural language case description, but also any additional ad hoc keywords that the user feels may be moderately relevant. These keywords are often helpful in sufficiently focusing a query, especially when physicians provide potential diagnosis keywords such as ‘lung cancer’, which tend to be highly discriminative.

Combining methods also resulted in an improvement in performance. This is expected, as the methods address different baseline limitations. Also the MeSH-based methods are likely to

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Baseline performance of individual queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query ID</td>
<td>Related documents</td>
</tr>
<tr>
<td>1</td>
<td>31</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>97</td>
</tr>
<tr>
<td>9</td>
<td>59</td>
</tr>
<tr>
<td>10</td>
<td>93</td>
</tr>
<tr>
<td>11</td>
<td>59</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>45</td>
</tr>
<tr>
<td>14</td>
<td>35</td>
</tr>
<tr>
<td>Average over all 14</td>
<td>37.2 (± 31.01)</td>
</tr>
</tbody>
</table>

* Values in brackets indicate the SD.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Combination results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run ID</td>
<td>Run name</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>Baseline</td>
</tr>
<tr>
<td></td>
<td>Thesaurus-based runs</td>
</tr>
<tr>
<td>O1</td>
<td>Sem. Wt.</td>
</tr>
<tr>
<td>O2</td>
<td>Top-N MeSH (10,0.1)</td>
</tr>
<tr>
<td>O3</td>
<td>Dist. MeSH (40,0.1)</td>
</tr>
<tr>
<td>O4</td>
<td>Sem. Wt. + top-N</td>
</tr>
<tr>
<td>O5</td>
<td>Sem. Wt. + Dist. MeSH</td>
</tr>
<tr>
<td>K1</td>
<td>Phy. Keys</td>
</tr>
<tr>
<td>K2</td>
<td>O1 + Phy. Keys</td>
</tr>
<tr>
<td>K3</td>
<td>O2 + Phy. Keys</td>
</tr>
<tr>
<td>K4</td>
<td>O3 + Phy. Keys</td>
</tr>
<tr>
<td>K5</td>
<td>O4 + Phy. Keys</td>
</tr>
<tr>
<td>K6</td>
<td>O5 + Phy. Keys</td>
</tr>
<tr>
<td>R1</td>
<td>Relevance feedback (N=20)</td>
</tr>
<tr>
<td>R2</td>
<td>O4 + Rel. fb.</td>
</tr>
<tr>
<td>R3</td>
<td>O5 + Rel. fb.</td>
</tr>
<tr>
<td>R4</td>
<td>K3 + Rel. fb.</td>
</tr>
<tr>
<td>R5</td>
<td>K4 + Rel. fb.</td>
</tr>
</tbody>
</table>

Maximum improvements are highlighted in bold.

All improvements are over the baseline run B1. Statistically significant improvements in mean average precision (by Wilcoxon signed rank test) are highlighted with superscripts.

*Significant using Wilcoxon signed rank test at level p<0.01.
†Significant using Wilcoxon signed rank test at level p<0.025.
‡Significant using Wilcoxon signed rank test at level p<0.05.
§Significant using Wilcoxon signed rank test at level p>0.1.
perform better as the performance of the underlying system improves. The only exceptions are runs that combine semantic weighing with physician keywords (runs K2, K5, K6). In these cases the performance is generally better than their counterparts without physician keywords (runs O1, O4, O5), but worse than those without semantic weighing (runs P1, K3, K4).

Overall, the best improvement was 42.4% (run K3) in terms of MAP, 11.7% in terms of P@10 and 37.7% in terms of R@30.

Relevance feedback
For the relevance feedback experiments, a physician user was shown the top 20 documents for each query. The user was asked to judge the presented documents as relevant or non-relevant. The subsequent (unseen) documents were then re-ranked using the relevance judgments provided by the user. The P@10 of all relevance feedback runs remained the same as that of the original ranking as the top 20 results were not re-ranked.

While relevance feedback is known to improve performance, in our case it helped moderately in some cases and not at all in others. Runs R1 to R5 in table 2 show the results of incorporating relevance feedback to some of the major thesaurus-based and additional keyword-based runs. Run R1, which merely adds relevance feedback on top of the baseline, achieves a 3.1% improvement. On the other hand, runs R2 and R3 show that adding relevance feedback on top of the best performing thesaurus runs (O4 and O5) degrades performance. In the case of the best performing additional keyword runs (K3 and K4) we again observe a small improvement (R4 and R5).

Overall, we achieve the best performance in terms of all the different measures when applying relevance feedback over a combination of additional keywords and distribution-based MeSH feedback. Performance of the top-N-based approach is also similar.

Details on individual query performances of some top performing methods are shown in supplementary appendix III (available online only).

Algorithm recommendation
Table 3 presents the most suitable method configuration depending upon application settings. It also suggests the possibility of constructing a system that dynamically picks up the most suitable configuration. For example, if the user supplies additional keywords and indicates that P@10 needs to be optimized, the system can generate a ranking using configuration K4. Such a hybrid system would presumably have a higher overall utility for users than any of the individual methods.

DISCUSSION
In this paper we present a study of methods for retrieving medical literature articles for case queries. While our results are encouraging, there are still some challenges that need to be dealt with before the methods can be really useful in a clinical setting. We discuss them below.

Small dataset size
The main limitation of our work arises from the small size of the evaluation set. A realistic system in a clinical setting will need to deal with hundreds of thousands to millions of documents. The creation of such a large labeled training collection requires time and expertise. Even though the ImageCLEF 2010 dataset is realistic in that the queries are based on real user needs, at 19 queries and 5585 documents, it is fairly small and does not capture the wide spectrum of documents and queries expected in real life. Currently, we deal with the issue by performing statistical significance tests (Wilcoxon signed rank test) to ensure the performance improvements are significant.

Suitability for clinical search tasks
The precision and recall numbers give us some sense of the utility of our methods in a clinical setting. For our best performing methods, average P@10 is approximately 0.48 (SD 0.24) and R@30 approximately 0.47 (SD 0.24). While the figure is reasonable (five relevant documents in the top 10% and 50% of all relevant documents in the top 30%), and may be enough to support search tasks in which, for example, the user is looking for a few good literature references on frequently seen cases, it may not be enough for difficult queries or high recall tasks such as literature reviews. In particular, the high variation in performance suggests that our methods need further refinement before they can be clinically deployed. As it is hard to define a standard for clinically acceptable values of precision and recall, we will ultimately explore this in future by user studies in a real clinical environment.

Failure analysis
We analyzed individual queries with poor retrieval accuracy (as reflected by low precision) to understand why our methods did not work well with some of them. In general, it seems that poorly performing queries were often ambiguous. In particular two kinds of failures stood out.

- Difficulty in recovering new treatments and rare alternative diagnoses.

Such documents usually have low keyword overlaps with the query. Our MeSH-based and semantic weighing methods are tuned towards retrieving documents with similar diseases and thus do not do as well in such scenarios. The problem is compounded by the fact that, in cases in which the disease itself is not the focus of the document, relevant disease MeSH terms may not have been assigned, making the MeSH-based pseudorelevance feedback methods less effective. Physician keywords help greatly when they overlap with such documents, but this does not always happen.

- Confusion between similar diseases/conditions

For example, in a query related to ‘arm fractures’, some ‘neck fracture’-related documents were also retrieved, due to high keyword overlap. In general, it was difficult to guess automatically when certain diseases must be logically ruled out.

The two failure modes are related in the sense that, while the first requires methods to have a high recall, the second requires higher precision. The first failure can be potentially alleviated by diversifying the retrieval results so that documents with a rare diagnosis would have a better chance to be ranked on the top, while the second problem can be addressed by using more specific units than single words (eg, phrases) for indexing. Ultimately, however, overcoming these challenges will require us...
to incorporate more complex concept models for relationships between biomedical concepts. For example, we can use the hierarchical relationships between concepts within MeSH for more refined feedback.

Physician keywords

Among the many different methods we tried, the incorporation of appropriately weighted physician keywords proved to be the most useful in improving performance. Our physicians were able to provide them fairly quickly, without looking at any relevant documents. This suggests that the technique would be relatively easy to apply in a clinical setting. While these keywords were provided by experts, a natural future task is to generate them automatically. Our methods based on extracting useful MeSH terms for a query are a step in this direction. However, MeSH is a controlled vocabulary, while physician keywords are often ad hoc. As a result, for our queries, we found only a limited correspondence between them. One possibility may be to use the co-occurrence of MeSH descriptors in document collections to suggest related keywords.

CONCLUSION

In this study, our focus was primarily at identifying the major challenges arising from the differences between general retrieval and medical case-based document retrieval, and then developing methods for addressing them. The results show that state-of-the-art retrieval methods perform reasonably well for this task, but performance can be improved by selective query term weighing based on medical thesauri and physician feedback. Overall, most of our strategies helped improve performance and the best performing method was up to 44.5% better than the baseline.

In the future, we plan to acquire a larger labeled dataset for some specific diseases for evaluation. In addition, we would like to investigate the operational aspects of the system by conducting user studies with real physician users.

Funding

This paper is based upon work supported in part by an IBM faculty award to investigate the operational aspects of the system by conducting user studies with real physician users. This work was also supported in part by a grant from the National Library of Medicine, National Institute of Health. The CLEF 2010 Medical Image Retrieval Task was also supported in part by a grant from the National Library of Medicine, National Institute of Health.

REFERENCES

11. Rothstein HR, Berkman ND, Glanz K.绝对的不相关性: 索引词的实践应用。The CLEF 2010 Medical Image Retrieval Task was also supported in part by a grant from the National Library of Medicine, National Institute of Health.