What's Related? Generalizing Approaches to Related Articles in Medicine

Howard R. Strasberg, M.D., Christopher D. Manning, Ph.D.,
Thomas C. Rindfleisch, M.S. and Kenneth L. Melmon, M.D.

Departments of Medicine and Molecular Pharmacology
Divisions of Clinical Pharmacology and
Stanford Medical Informatics
Stanford, CA

ABSTRACT

INTRODUCTION: We did formative evaluations of several variations to the computation of related articles for non-bibliographic resources in the medical domain. METHODS: A binary model and several variations of the vector space model were used to measure similarity between documents. Two corpora were studied, using a human expert as the gold standard. RESULTS: Variations in term weights and stopword choices made little difference to performance. Performance was worse when documents were characterized by title words alone or by MeSH terms extracted from document references. DISCUSSION: Further studies are needed to evaluate these methods in medical information retrieval systems.

INTRODUCTION

"Related Articles" is a feature of many information retrieval systems, most notably PubMed and many search engines on the web. A "related articles" feature is closely related to the use of relevance feedback, which has been widely established as a successful way of query refinement[1], but pursues this goal in an interactive information retrieval context, where a user can explore articles related to another article via hypertext links. The feature allows users to enter a fairly general query (e.g. 2-3 words), and by bootstrapping, work their way towards relevant documents. Users look through the top of the list of retrieved documents until they find one of particular relevance to their information need. They then ask the system to see a list of "related articles", and proceed from there to converge, perhaps iteratively, on a set of documents that answer their question. Note that users never fully specify their question to the retrieval system, and in fact, they may not fully and consciously formulate the question even in their own minds until late in the search process. It is often very hard to specify an adequate Boolean expression that would select for the salient features of a relevant "seed" document, and "related articles" tools bypass this step. In an analogous way, similarity matches are used to locate DNA or protein sequences in databases based on a sample sequence of interest.

The algorithms used to compute related articles are based originally on Salton's work in information retrieval[2]. Documents are represented as vectors in n-dimensional space, where n is the total number of unique terms in the corpus. The vector components are based on various weighting functions that consider both the term frequency (the number of times a given term appears in a given document) and the inverse document frequency (a number inversely related to the fraction of documents in the collection containing a particular term). Highest weight is assigned to terms that appear frequently in a document and infrequently in the corpus. Similarity between two documents is measured by the vector dot product (cosine) measure, in which the cosine of the angle between two document vectors is computed. Compared to less similar documents, very similar documents will have a smaller angle between them and thus a larger cosine score. These models have been adjusted in various ways in attempts to optimize retrieval performance. For example, the algorithm used by PubMed is based primarily on work by Wilbur[3], whose algorithm carefully weights words with discriminating power above the average use of words in English scientific discourse.

In this paper we evaluate several variations in the computation of related articles, where the articles are chosen from full-text resources in the medical domain.

METHODS

Two corpora were used to conduct experiments. Several approaches were used to compute similarity scores for each corpus – seven for the first corpus and four of these for the second corpus. The seven approaches are summarized in Table 1.

Approach	Description (VS=Vector Space; B=Binary)
1	VS, Hersh weights, 7 stopwords
2	VS, Wilbur weights, 7 stopwords
3	VS, Wilbur weights, 290 stopwords
4	VS, Wilbur weights, title words only
5	VS, Hersh weights, 7 stopwords, double count title words
6	B, Dice coefficient, no stopwords
7	VS, Wilbur weights, MeSH terms from references, no other text words

Table 1. Approaches studied in computing related articles.

The *first corpus* consisted of all 186 articles (full-text subsections) from the Cardiology and Pulmonology sections of Scientific American Medicine 1998. One article was chosen arbitrarily as the seed article. A medical expert agreed to read 90 articles randomly chosen from the 186 available articles, and to indicate which of these 90 articles were related to the seed article. The expert was told that "related" is defined as any article he would like to see if the related articles link were invoked when viewing the seed article. The expert read each of the 90 articles in detail and appeared to be extremely conscientious in making similarity judgments.

Approaches 1-5 and 7 used some form of the vector space model. Approach 1 used weights suggested by Hersh[4], in which TF' = $1+\log_{10}(TF)$, and IDF=$1+\log_{10}$ (N/n), where TF is the term frequency, TF' is the term frequency weight, IDF is the inverse document frequency weight, N is the number of documents in the corpus and n is the number of documents in the corpus containing a given term *t*. We used somewhat different weight functions in Approach 2 to see the effect on the results. Specifically, we used weights suggested by Wilbur, in which TF'=0.5+0.5x(TF/TF_d), where TF_d is the maximum TF over all terms *t* that occur in document *d*. This measure is known as the augmented normalized term frequency, and it attempts to normalize the term frequency to the length of the

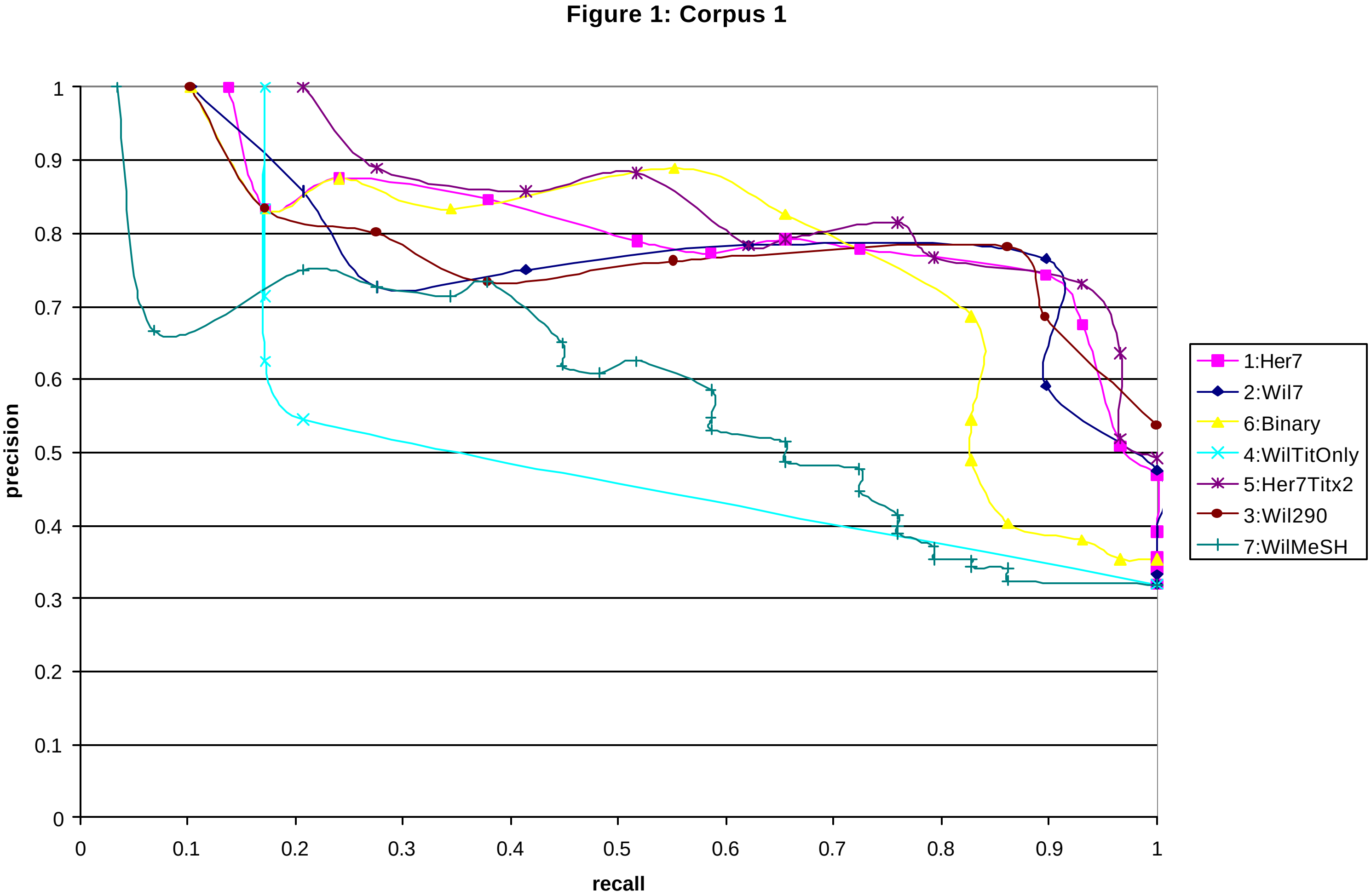

Figure 1. Precision-recall curves for seven approaches to related articles in corpus 1.

document. Approach 2 used IDF = $\log_{10}$ (N/n). Both approaches 1 and 2 used seven stopwords, which are {and, an, by, from, of, the, with}, as suggested by Hersh. Stopwords are entirely excluded from the calculations.

Approach 3 used the weighting formulae of Approach 2, but included 290 stopwords (instead of 7) in an attempt to evaluate the effect of stopwords on the results. A medical student chose these stopwords as words that had little contribution to finding relevant hits in the Stanford Health Information Network for Education (SHINE)[5].

Approach 4 again used the weighting formulae of Approach 2, but this time included only the words in the document titles, with the exception of the seven stopwords above.

Approach 5 used the weighting formulae of Approach 1, but double counted words appearing in the document title. That is, a word in the title was counted as if it appeared twice in the document, rather than once. A word appearing once in the title and 3x in the document would get a frequency of (2x1)+3=5. While titles should give high value keywords, their sparseness limits their utility. One might hope to get the best of both worlds by this combination.

Approach 6 was designed to compare the vector space model used in the above approaches with a binary method of computing similarity[6]. In this case document vector components were restricted to 0 or 1, where 0 meant the term did not appear in the document and 1 meant the opposite. The Dice coefficient was used to compute similarity, which for a pair of document vectors was calculated as twice the number of 1's in common divided by the total number of 1's.

Approach 7 differed from the others in that the documents were represented by MeSH terms rather than text words. Since the documents did not have MeSH terms assigned, MeSH terms from the document references were used as a surrogate measure of MeSH index terms for the document. That is, for a given document MeSH terms were

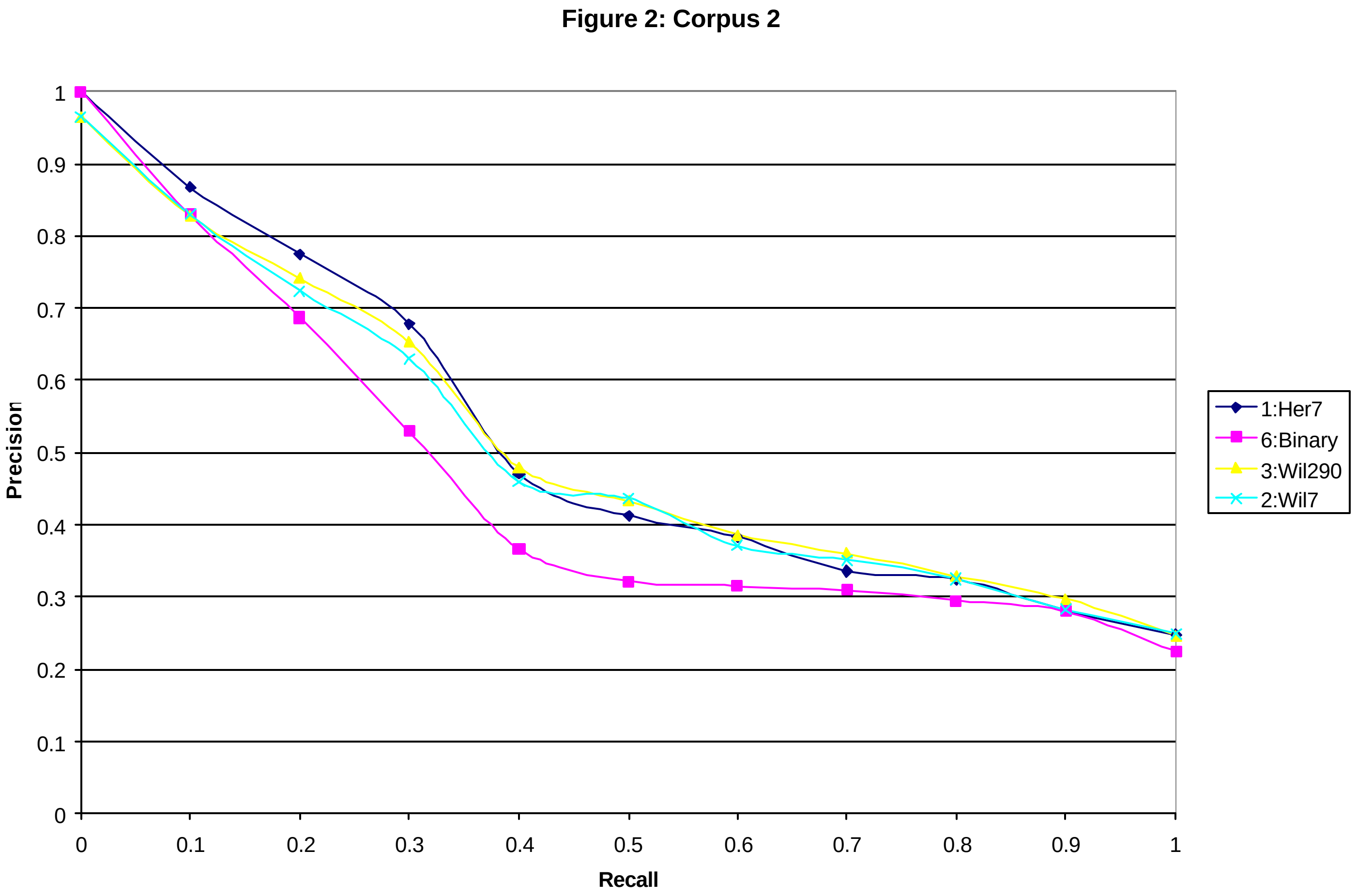

Figure 2. Precision-recall curves for four approaches used in corpus 2.

obtained for each journal article referenced by that document. The document was then described by the set of non-duplicated MeSH terms from all its references. For those documents with no MeSH terms, the MeSH terms were taken from the most similar document with MeSH terms, as computed by the vector space model. Once the documents were characterized by MeSH terms, Approach 2 was applied to compute similarity.

For the first corpus, each approach provided a score for similarity between the seed article and the 90 other articles that had been read by the medical expert. The gold standard for relevance was the physician reader. A precision-recall graph was created by assigning a range of thresholds to the computer output, such that documents with scores equal to or greater than the threshold defined the set of related articles as chosen by the computer.

The *second corpus* was assembled by asking a physician expert to use SHINE to find 20 articles, all related to each other. The expert chose articles from a textbook, a drug database, full text journals and guidelines. 80 additional, unrelated articles were randomly chosen from SHINE content, such that the distribution of sources of these 80 unrelated articles was in proportion to the distribution of sources in the sample of 20. In the event that a randomly chosen article was related to the set of 20 related articles, another article was randomly chosen in its place.

Each article was used in turn as a seed article with the assumption that it was related to the 19 other articles in the set chosen by the expert, and that it was unrelated to the 80 additional articles. For each seed article, precision was calculated at fixed levels of recall. The precision was averaged over all 20 seed articles at each level of recall, thus permitting the creation of a precision-recall curve and the calculation of the 11-point interpolated average precision[a] for the approaches (1,2,3,6) applied to the second corpus.

RESULTS

Precision-recall curves for Approaches 1-7 for the first corpus appear in Figure 1.

Precision-recall curves for the second corpus appear in Figure 2. Table 2 provides the 11-point interpolated average precision for the four approaches used. The 11-point average precision for Approach 1 and Corpus 1 was 0.818.

Approach 1	0.525
Approach 2	0.511
Approach 3	0.520
Approach 6	0.469

Table 2. Interpolated average precision for Corpus 2.

DISCUSSION

We have described approaches to compute measures of similarity (relatedness) for full-text documents in the medical domain and have done a series of formative evaluation experiments on these methods.

Text word-based approaches to similarity assessment all performed comparably. That is, the various weighting formulae used and the different numbers of stopwords made little difference to the general results. This insensitivity of performance to weighting approaches was seen in both corpora. In general, title words alone and MeSH terms from references did not perform as well, particularly at higher recall levels.

Using title words alone is extremely effective at moderate levels of recall, but degrades sharply. One accurately finds related articles that share title words – and this is an effective technique because titles are carefully chosen to summarize an article – but it is extremely ineffective once one is at higher recall levels and needs to notice similarities not conveyed by title words.

Similarly, using MeSH terms from references produced good results at low levels of recall, but this method also degrades sharply. Presumably many related articles do not share MeSH terms in their references. Another approach might look at the citations themselves as a measure of similarity between documents, as described by Kessler[7].

Approach 5, generally effective at getting the best of both worlds (title and non-title words), performed best at most recall levels.
These methods all performed well at low levels of recall for both corpora. These results are encouraging, since users are likely to look primarily at the top 10 or 20 related articles returned.

The major purpose of this study was to compare various methods of computing similarity. We were primarily concerned with the relative performance of these methods with respect to each other, rather than

[a] precision at 11 fixed levels of recall from 0 to 1 in increments of 0.1

with their absolute performance. Average precision depends on the overall proportion of relevant documents in the corpus, with a higher proportion of relevant documents resulting in a higher average precision. We deliberately chose corpora with relatively high proportions of relevant documents to maximize the ability of these experiments to distinguish between the different methods to compute similarity. Corpora 1 and 2 differed in their overall proportions of relevant documents, and therefore differences in average precision were observed. We were not concerned about these differences, since our focus was on the relative performance of the different methods *within* each corpus, rather than *between* corpora.

Only one seed article and one expert were used in the study of Corpus 1. While 20 seed articles were used in the study of Corpus 2, this approach to assembling a corpus has two problems. First, even though the 80 unrelated articles were selected randomly, the specific subject matter of these unrelated articles could nevertheless affect the results. The second problem is that among the 20 related articles, not all pairs of articles were related to the same degree. We used a convenient method to obtain this sample, but its limitations are clearly illustrated by the relatively poorer performance seen in Figure 2 compared to Figure 1. We may have seen better results in Figure 2 if we had experts willing to spend the time to rate relatedness of the 99 other articles for *each* of 20 seed articles.

Future studies should look at several seed articles, and have them rated by several experts. Doing so would overcome the limitations of both corpora in our study. Unfortunately, the task of reading a large number of articles is very time-consuming, and practical considerations may limit, but not eliminate, the feasibility of more elaborate studies.

Further assessment of these methods could be done by their actual implementation in an information retrieval system, and a subsequent analysis of system retrieval performance and user satisfaction.

CONCLUSION

We evaluated the relative performance of various methods to compute related articles in the medical domain. Variations in term weights and stopword choices made little difference to performance. Performance was worse when documents were characterized by title words alone or by MeSH terms extracted from document references. Further studies are needed to evaluate these methods more summatively in medical information retrieval systems.

ACKNOWLEDGMENTS

This research was supported in part by NLM training grant LM07033.

REFERENCES

[1] Salton G, Buckley C. Improving Retrieval Performance by Relevance Feedback. Journal of the American Society for Information Science 1990;41(4):288-97.

[2] Salton G. Introduction to Modern Information Retrieval. McGraw-Hill. New York, 1983.

[3] Wilbur WJ, Yang Y. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. Comput Bio Med 1996;26(3):209-222.

[4] Hersh WR. Information Retrieval: A Health Care Perspective. Springer-Verlag. New York, 1996.

[5] Hubbs PR, Tsai M, Dev P et al. The Stanford Health Information Network for Education: integrated information for decision making and learning. Proceedings AMIA Annual Fall Symposium 1997:505-8.

[6] Manning CD, Schuetze H. Foundations of Statistical Natural Language Processing. MIT Pr 1999.

[7] Kessler MM. Bibliographic coupling between scientific papers. American Documentation 1963;14:10-25.