Chapter 10
MODELING TEXT RETRIEVAL IN BIOMEDICINE

W. John Wilbur

National Center for Biotechnology Information, National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894

Chapter Overview

Given the amount of literature relevant to many of the areas of biomedicine, researchers are forced to use methods other than simply reading all the literature on a topic. Necessarily one must fall back on some kind of search engine. While the Google PageRank algorithm works well for finding popular web sites, it seems clear one must take a different approach in searching for information needed at the cutting edge of research. Information which is key to solving a particular problem may never have been looked at by many people in the past, yet it may be crucial to present progress. What has worked well to meet this need is to rank documents by their probable relevance to a piece of text describing the information need (a query). Here we will describe a general model for how this is done and how this model has been realized in both the vector and language modeling approaches to document retrieval. This approach is quite broad and applicable to much more than biomedicine. We will also present three example document retrieval systems that are designed to take advantage of specific information resources in biomedicine in an attempt to improve on the general model. Current challenges and future prospects are also discussed.

Keywords

relevance; probability; ranking; term weighting; vector model; unigram language model; smoothing

1. INTRODUCTION

Most papers written on the subject of text retrieval begin with the observation that the digital age has brought a deluge of natural language text and we are more or less overwhelmed by the amount of text available on the web and even in the specialized databases of interest to researchers. Certainly this is true in the field of biomedicine. The serious question is whether a researcher must personally read everything remotely related to her field of interest, or can technology rule some texts as not useful for her purposes and allow her to concentrate her search on a few texts where her efforts will have high yield. The answer is that, with a certain risk, technology can reduce the work load for information access in textual databases.

To better understand the risk of using technology in lieu of reading all the documents for oneself, it is helpful to think in terms of a simple model. Let the user be denoted by X, the information need state of X be denoted by S, and finally let the query by which X has expressed their information need be denoted by q. Typically for users of a search engine, q consists of one to three words or a short phrase. Such a short query is naturally quite inadequate to represent the need state S. In fact, an important study by Furnas et al. (1987) found that common objects are generally referred to by, on the average, five different names over a sample of references by different people. This same fact is underlined by the famous Blair and Maron (1985) study of retrieval in the area of legal documents. They discovered that legal experts, after a careful search using keywords and Boolean queries, felt they had found most of the relevant material pertaining to a case, but in fact more extensive and careful search showed they had only found less than 20% of what was relevant to the case. It proved impossible to predict the words people would use to describe relevant material. Further evidence on this point is provided by a study of MEDLINE® indexing. Funk et al. (1983) found inter-indexer consistencies ranging from 0.3 to 0.6 for different types of MeSH® term assignments. But to the extent we are unable to predict the indexing we are also unable to use it effectively for retrieval.

While we might try to improve indexing by expending more human effort on the process, there is an even more fundamental barrier to perfect retrieval. This stems from the variation observed in what people judge to be relevant to a query. Different judges agree on what is relevant to a query from 40% to 75% of the time (Saracevic, 1991; Swanson, 1988). This is true even for queries as long as the title and abstract of a MEDLINE document (Wilbur, 1998). If a human processing the query q can only find material relevant to user X with a precision of 75%, then that says something very important. An algorithm that is as "smart" as a typical human is also going

to find relevant material with a precision no better than 75%. Of course we do not have algorithms that can perform at a human level and probably will not for the foreseeable future (Shieber, 1994). Now one can lengthen the query q and thereby decrease the ambiguity. One way to accomplish this is to allow the user to choose a document that represents what they would like to see. Our research (Wilbur and Coffee, 1994) shows that one can make a substantial improvement in the retrieval process by this method. The chosen document becomes a query in its own right which is generally much longer and more detailed than one a user is willing to write. One can carry this idea even further by applying relevance feedback. Here a user makes judgments for the top few documents retrieved and these judgments are used to improve the ranking of the documents the user has not yet examined. Results of one study (Wilbur, 1998) show that if a user makes judgments on the top 50 documents, then a machine learning algorithm can convert that additional information into retrieval on new documents at a level at least as good as a human agent could accomplish based on the original query. Of course, even this is not as good as the user can do for himself and probably not as good as a human agent could do given the additional information consisting of the user's judgments. Furthermore there are practical limits in getting users to make multiple relevance judgments and the method has seen little use.

Based on the above described limitations it seems unlikely that any algorithm can ever remove the risk of missing important information. On the other hand it is clear that users must rely on algorithms because there are few topics in biomedicine where one could hope to read all the literature available. The best algorithms are those that minimize the risk of information loss.

2. LITERATURE REVIEW

One could say that the field of modern information retrieval began with the work of Maron and Kuhns (Maron and Kuhns, 1960) describing how to calculate probability of relevance of documents to a query. Their approach required that the individual documents have index terms assigned, each with a probability that if that document were retrieved this index term would be the term used to retrieve it. While such probabilistic indexing is doable in principle it is not very practical. However the approach clearly showed the way to a probabilistic treatment of information retrieval as reflected in later work by Sparck Jones and Robertson (Robertson and Sparck Jones, 1976; Sparck Jones, 1972). One of the problems with the probabilistic approach to information retrieval is a lack of the specific information needed to give the best possible estimates of the probabilities involved. Work by Croft and

Harper (Croft and Harper, 1979) pointed the way to giving reasonable estimates without detailed relevance information. We believe the traditional probabilistic approach to information retrieval has achieved its most mature statement in (Sparck Jones et al., 2000a, 2000b).

At about the same time as the efforts to understand information retrieval in terms of probability theory just described, there was a significant initiative to perform computerized retrieval experiments at Cornell University under the direction of Gerard Salton (Salton (Ed.), 1971). A retrieval system called the SMART system was under development and a number of different algorithms were tested on a variety of test databases to assess the value of different approaches to retrieval. Out of this effort came $tf \times idf$ term weighting (Salton, 1975) and the vector retrieval approach (Salton et al., 1975). Here tf stands for a factor related to the frequency of a term in a specific document (local factor) and idf stands for a factor related to the frequency of the term throughout the database (global factor). The vector approach assumes that each document can be represented by a vector in a space with as many dimensions as there are unique keywords throughout the database. A vector representing a particular document will have a co-ordinate value corresponding to a particular keyword that is the $tf \times idf$ weight for that keyword in that particular document. As a general rule tf (and hence $tf \times idf$) is zero when the keyword does not occur in the document. Thus documents are represented by sparse vectors in the vector space model.

Though the vector space model may seem to be fundamentally different than the probabilistic model, in the final analysis the probabilistic model may be seen to be a special case of the vector space model by simply choosing the formulas for tf and idf to be the values they receive in the probabilistic model. In fact, the vector space model is quite general and allows for the possibility of many different forms depending on how the tf and idf formulas are chosen. We shall subsequently give forms for these quantities which we have found very useful for retrieval in the PubMed database. However, Witten, Moffat, and Bell (Witten et al., 1999) make the following significant observations. Whatever formula is used for tf , within a single document a term with a higher frequency within that document should have a tf at least as great as any term with lower frequency. Likewise, globally a term with a lower frequency throughout the database should have an idf value at least as great as any term of higher frequency. They point out that hundreds of formulas that obey these constraints have been tested on the TREC (Text REtrieval Conference, http://trec.nist.gov/) data (Zobel and Moffat, 1998) and no one formula is best. Rather the choice of formula is a matter of taste and perhaps of the idiosyncrasies of the particular type of data at hand.

In 1998 Ponte and Croft (Ponte and Croft, 1998) introduced a new approach to information retrieval with what they termed language modeling. The idea is that given a query and a document in the database, one may use the frequencies of words in the document to estimate the probabilities of the words in the query and hence the likelihood that the query came from the same source as the document. It is assumed that the document which assigns the highest probability to the words in the query is the document most likely to be relevant to the query. One of the difficulties faced by the method is that not all the words in the query necessarily appear in the document. This is solved by a process of smoothing, which relies on the frequencies of words throughout the database to estimate the probability of seeing words in the query that do not occur in the document. This smoothing process is the real tie to language modeling. The language modeling approach is competitive with other methods and is an active area of research (Kurland and Lee, 2004; Zaragoza et al., 2003; Zhai and Lafferty, 2004). It remains to be seen whether it offers an advantage over other methods and whether there is one best way to do it, or many ways that each offer some small advantage for a particular type of text or a particular database.

In what follows we present an ideal model of information retrieval and then show how the different methods we have described can be seen as special cases of this ideal model. Finally we give some examples of systems that attempt to use particular resources and aspects of biology to advantage to provide a more convenient or more effective approach to information retrieval in limited subdomains.

3. AN IDEAL MODEL

Because of the inherent limitations of information retrieval from natural language texts the problem is most conveniently formulated in terms of probability theory. It is helpful to approach the problem by first describing an ideal retrieval model. We assume that people can be in any one of a set of mutually exclusive states of information need. Such a set of states can be denoted by $\{S_i\}$. Then given a document d there are three different probabilities that are important to consider:

$p(S_i)$ - The prior probability that the randomly chosen human X is in the information need state S_i. This is global information and has nothing to do with a particular person or a particular document.

$p(d|S_i)$ - The probability that a person in the information need state S_i would consider d relevant to that information need. This is local information about S_i and the probability that d gives useful information about the concern expressed by S_i.

The third probability can be expressed in terms of the first two through application of Bayes' theorem

$$p(S_i \mid d) = \frac{p(d \mid S_i)\, p(S_i)}{\sum_{j=1}^{N} p(d \mid S_j)\, p(S_j)}. \quad (1)$$

This is the probability that if a person has judged the document *d* relevant, that person is in information need state S_i.

Now we make an assumption about information need states. Namely, the need state of a user contains all the information about that user's need and once the need state is known the relevance of different documents to the need state become independent events. The probability that a person who sees document *d* as relevant will also see document *e* as relevant is an important quantity in the theory. By use of the probabilities just considered we may write this probability.

$$p(e \mid d) = \sum_{j=1}^{N} p(e \mid S_j)\, p(S_j \mid d) = \frac{\sum_{j=1}^{N} p(e \mid S_j)\, p(d \mid S_j)\, p(S_j)}{\sum_{j=1}^{N} p(d \mid S_j)\, p(S_j)} \quad (2)$$

Here the equality on the left follows from the assumed independence of the relevance of e and d given the information need state S_j. The right side equality follows from substitution of Eq. (1) into the middle term of Eq. (2). It is illegal to substitute d in place of e in Eq. (2) because the formula is only derivable if e and d are independent as assumed and d cannot be independent of itself. The value $p(d \mid d)$ is of course 1, while if one incorrectly substitutes d for e in Eq. (2) one generally obtains a number less than 1.

The information that a person has observed the document *d* could change the state of information need, but that is not dealt with in this model. It would require some modification of Bayes' formula as it appears in Eq. (1). To deal with this one can make the distinction of transient and stable states of information need. This introduces the concept of dynamics into the problem. In our formulation here we deal with only stable states. A person can deal with a change in state in a search for neighbors by simply dropping the search and perhaps taking up another thread of interest where there is still a need which is described by a different state. In this approach the human deals with the issue and it is not necessary to introduce this complexity into the computer model.

If we are given a document *d* and the knowledge that a user has found it relevant to their information need, then we may wish to find other documents most likely also relevant. For this purpose we may apply Eq. (2) to rank all the other documents. In this process *d* is constant and all we are concerned about is the relative ratings. Thus we may simplify the formula to

$$\begin{aligned} sim(e,d) &= \sum_{j=1}^{N} p(e \mid S_j) p(d \mid S_j)\, p(S_j) \\ &= \left(p(e \mid S_1)\sqrt{p(S_1)}, \ldots, p(e \mid S_N)\sqrt{p(S_N)} \right) \cdot \\ &\quad \left(p(d \mid S_1)\sqrt{p(S_1)}, \ldots, p(d \mid S_N)\sqrt{p(S_N)} \right) \end{aligned} \quad (3)$$

This formula has the advantage that it is symmetric in its arguments and can be written as a vector dot product of vectors that represent the two documents involved. These vectors are not normalized in general because they come from probabilities which need not obey such rules. On the other hand the possibility that they are normalized is not excluded.

The formula Eq. (3) may be applied to find the documents related to a given document d or it may be applied more generally when d is understood to represent some query text q. The key to its application is to identify some meaningful set of information need states that can represent the set $\{S_i\}$. How this may be done is the subject of the next section.

4. GENERAL TEXT RETRIEVAL

In general text retrieval, two kinds of information have proven useful. First, the frequency of a term throughout the database carries information about the general usefulness of the term. The less frequent the term is overall, the more informative that term tends to be. Second, the frequency of a term in a document and the overall size of the document combine to give an indication of the importance of the term within the document. The higher the frequency of a term relative to the frequency of other terms in the document the more important the term is likely to be in representing the document's subject matter. These two kinds of information together provide the raw material from which need states may be constructed. There are two important ways that this has been done.

4.1 Vector Models

The vector model assumes that each keyterm is weighted by a global weight gw_t for the term t and by a local weight that relates the term to the

document and may be denoted by lw_{td}. For any document d we can then construct a vector

$$v_d = (lw_{td} \times gw_t)_{t \in T} \tag{4}$$

where T represents the set of all keyterms used in the database. Typically the local weight lw_{td} is zero if the term t does not appear in the document. With this representation the similarity between two documents is given by

$$sim(d,e) = v_d \cdot v_e \tag{5}$$

Equations (3) and (5) will correspond if we identify the set of states of information need with the set of keyterms T and define the probabilities by

$$\begin{aligned} p(t) &= (gw_t)^2 \\ p(d \mid t) &= lw_{td} \end{aligned} \tag{6}$$

With these identifications we have an exact correspondence between the two equations. There is one minor problem with the correspondence. That is that $(gw_t)^2$ may not be a number between zero and one and further the sum of all such numbers may not be one. Both these problems can be corrected easily by making the definition

$$p(t) = (gw_t)^2 / \sum_{t' \in T} (gw_{t'})^2 . \tag{7}$$

This has no effect on the ranking because the normalization factor is a constant, but it endows the numbers with the correct formal properties to be probabilities. Thus the typical vector retrieval formula can be derived from the state space paradigm by making the correct identification of the probabilities involved.

One must ask how realistic it is to identify the set of states of information need with the set of keyterms. There are several pieces of evidence that favor this interpretation. First, it finds some justification in the fact that in search engines people typically express their information need with one or a very few terms (Silverstein and Henzinger, 1999). Thus in many cases a single word will express an information need effectively. Second, the formulation provides a natural probabilistic interpretation to vector retrieval, which has been viewed as ad hoc and empirical (Salton, 1991). Third, some of the local weight formulas that prove to be very effective in practice produce a number between 0 and 1 which is readily interpretable as a probability. This grows out of work by Harter who hypothesized (Harter, 1975) that important and unimportant terms follow two different Poisson distributions in their occurrence within documents. While this hypothesis did not initially lead directly to an advantage in information retrieval, Robertson and Walker

(Robertson and Walker, 1994) used the basic idea to design formulas for the local weighting of terms in documents. One of their more effective formulas appeared in (Ponte and Croft, 1998)

$$lw_{td} = tf_d / \left(tf_d + 0.5 + 0.5 * dlen / avedlen \right) \quad (8)$$

Here tf_d is the number of occurrences of t in d and $dlen$ is the length of d and equals the number of tokens in d while $avedlen$ is the average length of documents over the whole collection. Our own formulation is based on a more direct application of Harter's idea. Assuming two different rate constants, λ_i for important words in a document and λ_u for unimportant words, the probability that a word is important is given by

$$lw_{td} = \left[1 + Ce^{(\lambda_i - \lambda_u) dlen} \left(\lambda_u / \lambda_i \right)^{tf_d - 1} \right]^{-1} = \left[1 + e^{0.0044 dlen} \left(0.7 \right)^{tf_d - 1} \right]^{-1}. \quad (9)$$

Here the constants are determined by the data to obtain good performance. We find a slight (not statistically significant) advantage with Eq. (9) on our test data and also prefer it because of its sound theoretical basis in probability theory. It is used in computing the related documents in PubMed. For the global weight we use the traditional *IDF* weighting formula $\log\left(N / n_t \right)$ and set

$$gw_t = \sqrt{\log\left(N / n_t \right)} \quad (10)$$

where N is the total number of documents in the database and n_t the number of documents that contain the term t.

4.2 Language Models

Beginning with the seminal paper by Ponte and Croft (1998), unigram language models have become an important approach to textual information retrieval. Typically a language model is estimated from some corpus of text and used to estimate the probability of some new piece of text that is not a part of the corpus used to produce the language model. Bigram or trigram models involve the frequencies of word pairs or triples, respectively. They are useful in speech recognition or spelling correction tasks where one uses the most recent word or pair of words in an attempt to predict the next word. In a unigram model one simply uses the frequencies of words in an attempt to estimate the probability of seeing each word in a piece of text and thereby the probability of that piece of text. This approach to computing the probability of a piece of text naturally fits the paradigm of Eq. (3) provided we identify the states of information need with the possible language models that would be used to describe text in the area of need. This approach to

information retrieval has been articulated by Zaragoza et al. (2003). Assuming a Dirichlet prior distribution $p(S)$ and assuming a unigram language model (multinomial) the distribution $p(S \mid d)$ has a natural interpretation as the conjugate Dirichlet distribution. They are able to use this approach to compute $p(q \mid d)$.

$$p(q \mid d) = \frac{1}{p(d)} \int_S p(q \mid S) p(d \mid S)\, p(S)\, dS \tag{11}$$

This equation is just a form of Eq. (2) when one recognizes that the integral is a generalized sum. For ranking purposes this is equivalent to $p(d \mid q)$ (assuming a flat prior distribution $p(d)$). For further details we refer the reader to the original paper.

The more typical approach in language modeling for retrieval is to assume the distribution $p(S \mid d)$ is all concentrated in the single language model that maximizes $p(d \mid S)\, p(S)$. In this calculation the prior distribution $p(S)$ is assumed to be Dirichlet and is based on the collection frequencies of all terms. The resulting maximum likelihood language model blends the term counts in d with the collection frequencies and produces probabilities for individual terms given by

$$p(t) = \frac{tf_d + \mu\, p(t \mid C)}{dlen + \mu} \tag{12}$$

Here $p(t \mid C)$ is the fraction of tokens in the collection C that are t. This formula blends the estimate that would be based on the term counts in the document with the estimate that comes from the whole database. Terms that occur in the document would otherwise have their probabilities over estimated while terms that did not occur would have their probabilities under estimated. The result of the formula is a correction for this and is known as smoothing. Typically the parameter μ is several hundred to a few thousand (Zaragoza et al., 2003; Zhai and Lafferty, 2004). A method to automatically choose μ for good performance has been proposed in (Zhai and Lafferty, 2004).

Given the probabilities of individual words as in Eq. (12), the probability of a query text, $q = q_1 q_2 \ldots q_n$, is computed as

$$p(q \mid d) \propto \prod_{i=1}^{n} p(q_i). \tag{13}$$

Such numbers are equivalent to $p(d \mid q)$ (because $p(d)$ is assumed to be a constant over documents) and are used to rank the documents for retrieval.

One may naturally ask which approach to text retrieval, vector or language modeling, is best? We are not aware of any definitive comparison

of the two techniques. Researchers reporting on the language modeling approach have found it to perform well and it seems to be competitive with the more traditional vector approach of single term weighting. There are some differences in the two theories, in particular relating to how a user's information need state S is conceived (Robertson and Hiemstra, 2001; Sparck Jones, 2001). On the other hand it can be shown that in practice the way the two models are implemented produces results that are closely related (Zhai and Lafferty, 2004) and smoothing in the language model produces the equivalent of *IDF* weighting in the vector model.

We believe progress is possible in the general retrieval model, provided one can find a more realistic model for the information need states of a user. One can imagine that a more realistic way to represent an information need is in terms of concepts. However, it has not yet proved practical to represent the full scope of needs for a user of a large database with concepts. Concepts tend to be difficult to define and require a good deal of human curation. Even the concepts defined in the Unified Medical Language System (Humphreys et al., 1998) are not sufficient to represent all the different ideas that come into play in medical literature. An automatic way of finding concepts could lead to progress in this area.

5. EXAMPLE TEXT RETRIEVAL SYSTEMS SPECIALIZED TO A BIOLOGICAL DOMAIN

Given a large database in a medical or biological field as opposed to a general text collection such as the Brown Corpus or a collection of news articles from the Wall Street Journal, one might expect that there would be methods of retrieval in the area of biology in general that would work better for biology than for other areas. However, there is no approach that we are aware of that really makes information retrieval in the biological area better than general retrieval. This is true because the area of biology is simply too broad to allow any simplifications specific to biology. Just about any kind of text construction or topic that can appear in a large database of text in the field of biology can appear in any other collection, though the frequency of some types of text is less in documents on biology. As a consequence the PubMed (http://web.ncbi.nlm.nih.gov/PubMed/) search-engine-related documents function is based on a version of vector retrieval as outlined in the previous section. However, there are attempts to create databases in specialty areas of biology and medicine where retrieval can improve on the general model. We will describe several of those systems here.

5.1 Telemakus

The Telemakus system developed by S. Fuller and colleagues (Fuller et al., 2004) at the University of Washington represents research reports schematically with twenty-two fields or slots that contain information describing the research in different ways. Twelve slots are bibliographic and filled from PubMed, one is the Telemakus ID, and the remainder are extracted from the full text of the document. Among the fields filled from the document are items from the Methods section of a report that describe how the research was performed. Perhaps of most significance is the field that holds research findings. These are extracted especially from the captions of figures and tables and the extraction process makes use of the fact that the language in such captions is somewhat restricted and easier to process. Telemakus uses automated extraction to initially produce the schematic surrogate for a document. Then this automatically produced surrogate is displayed along with a marked up version of the original report so that a human expert can correct errors and finalize the schematic representation of the document.

Once data representing research reports has been entered into the system a user can access this information by keyword searching or in some cases browsing an index. When a particular study has been displayed, figures and tables can be accessed directly as can the full text document if available. Research findings are displayed for the study and can be queried for other studies reporting the same finding. In addition concepts that are represented in the database can be displayed in a window as a concept map. Such a map displays the concept along with other related concepts (measured by co-occurrence in research reports). One can then navigate by clicking on different concepts to search for concepts related to the original but perhaps more specific to the information need. A concept map for “neoplasms” is illustrated in Figure 10-1.

The Telemakus system is available at http://www.telemakus.net/ and currently comprises a database of research reports on Caloric Restriction and the Nutritional Aspects of Aging. A strength of the system is the ease with which one can examine the important findings in a report without having to read the whole report. A potential weakness is the need for a subject expert to examine each surrogate for a report and correct mistakes. Ways are being sought to make the system more nearly automatic. Currently the system is tied to the area of biology as a number of Unified Medical Language resources are used in its processing. However, there is in principle nothing to preclude its application in a wider context.

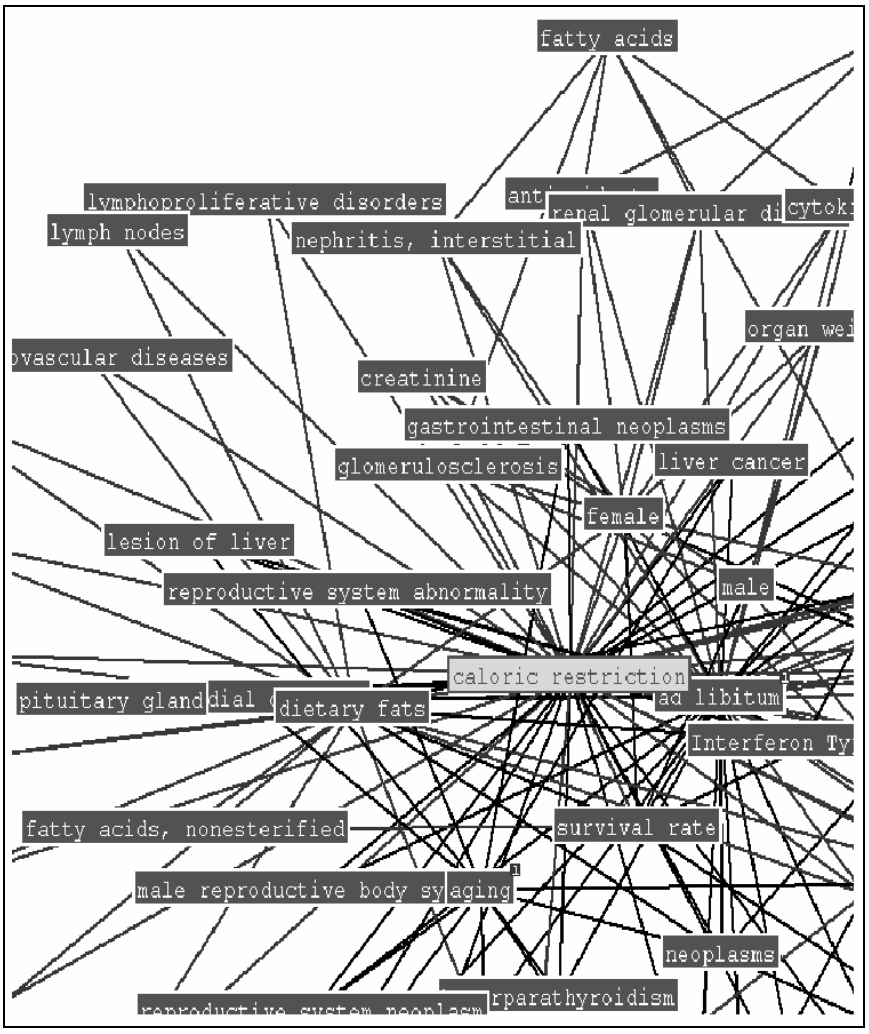

Figure 10-1. A concept map of research findings linked to "caloric restriction" (links reported by authors as statistically significant are in blue in original).

5.2 XplorMed

XplorMed is a system developed by Perez-Iratxeta and colleagues (Perez-Iratxeta et al., 2002; Perez-Iratxeta et al., 2001, 2003) for browsing the MEDLINE literature database. Given a set of MEDLINE abstracts (up to 500) the system computes the words with the strongest relations with other words (stop words excluded) as the keywords for the set. One keyword

is said to include another at a level greater than α if the presence of the first keyword in an abstract implies the presence of the second with a probability of at least α based on the data in the set of abstracts under analysis. The inclusion relation is used to define chains of words that all occur together in at least some of the documents. Keywords, included words, and word chains are displayed to the user during a session.

To begin a session the user of XplorMed may define a subset of MEDLINE abstracts by running a PubMed query or importing an already defined subset from another source or application. In the first step the set of abstracts is broken up into mutually exclusive subsets based on a set of broad MeSH categories. The user may then select a combination of the documents in any number of these categories to include in his analysis. Once the set of abstracts is finalized the system extracts keywords and computes their relationships to each other. Keywords are displayed as a list to the user who then has a number of options including asking to see a particular keyword in context in the abstracts in which it occurs, asking for the words implied by or included with a given keyword, or asking for all the word chains involving the keywords. Given a word chain one can then ask for a ranked list of those documents that contain the word chain. At the same time one may request to display links from the resulting set of documents to other databases such as OMIM, SwissProt, etc. It is also possible to display a listing of the MeSH terms that are contained in the resulting set of documents. Finally one may take the current set of documents and start the analysis cycle over. One also has the option at this stage to enlarge the set by pulling in related documents using the PubMed related documents function and to restart the analysis with this larger set.

A strength of the XplorMed system is its application of simple statistics on word use to find useful relationships between text words without reliance on a controlled vocabulary. A weakness is that such processing cannot guarantee that the relationships found are useful at the level that defined relationships between elements of a thesaurus are useful.

5.3 ABView:HivResist

In order to focus on a small set of the MEDLINE literature, Belew and Chang (2004) have developed a system called ABView:HivResist. This system is designed to provide an enhanced environment for the study of HIV drug resistance and the mutations that produce it. Currently the system contains 9,190 MEDLINE abstracts in the area of HIV protease inhibitors. Focusing on a limited set such as this reduces the ambiguity of terms, especially abbreviations, and makes practical the construction of a thesaurus which captures much of the synonymy in the domain. In particular the

thesaurus includes alternative ways of referring to mutations in the HIV protease molecule and the different names applied to the drugs studied for their inhibitory effects on this molecule. Relatedness of documents within the set may be assessed based on the citation of one by the other or by their relatedness as computed in the related documents function from PubMed.

The user of ABView:HivResist interacts with the system through a GUI and the results of a Boolean query appear in the main window (see Figure 10-2).

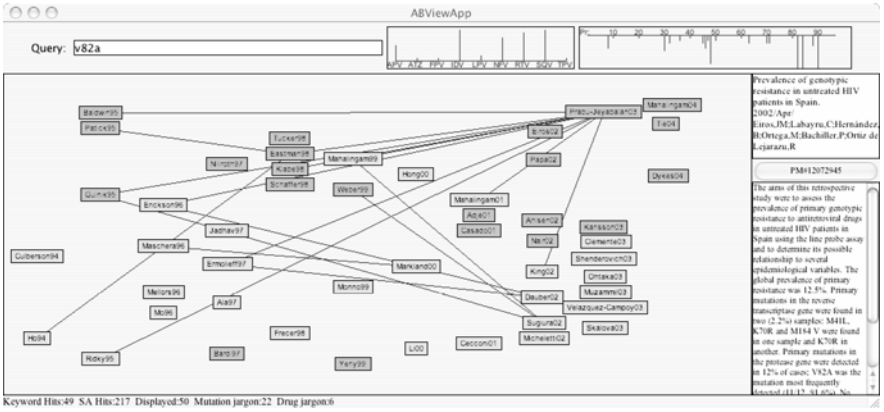

Figure 10-2. ABView:HivResist GUI allowing user to see the results of a search for documents that mention a mutation replacing valine by alanine at position 82 in the HIV-1 protease molecule.

Each document is represented as an icon and the most relevant documents appear highest in the window, while horizontal position in the window denotes time of publication over the past ten years. The citation of one document by another is denoted by an arrow from one icon to the other in the display. Documents directly retrieved by the query are displayed in dark green and those present only by virtue of a related document link to a direct hit are displayed in light green. There are also two histograms displayed, one showing which residues of HIV protease are mentioned and the other which inhibitory drugs are mentioned in the documents displayed. Clearly one of the strengths of this system is the narrow focus which allows one to develop a thesaurus tailored to a particularly important research problem. The drawback is the human effort required to develop this enhanced facility.

5.4 The Future

The three systems presented here each have the objective of using specialized information resources to enhance retrieval. We believe that such

approaches are the future of information retrieval. However, there are several problems that stand in the way of progress in improving information retrieval for specialized domains. The first problem is related to the cognitive effort required to create the structures that allow improved retrieval. Such cognitive effort is required in both Telemakus and ABView:HivResist. In Telemakus a human operator must examine the automatically produced surrogate document and make any necessary corrections. Ideally one would produce a completely reliable surrogate automatically, but computers do not have the necessary capabilities to understand language. Thus it is not possible to tell reliably what is a research finding without human intervention. Likewise ABView:HivResist required a human to describe the different ways a mutation is indicated in text. This is a specialized task and could not be done automatically based on the UMLS Thesaurus, for example. On the other hand the UMLS Thesaurus may be helpful in indicating alternative names for drugs used to treat drug resistant HIV. However, again one could not rely on any pre-constructed source to reveal just which drugs are important in treating drug resistance in HIV infection. The point is simply that in order to construct such specialized access tools a significant human effort is required and this will limit the more global application of such methods until the time we have more reliable automatic language processing tools.

A second problem standing in the way of progress is the lack of understanding of what the human mind is actually doing when one is searching for information. If we understood this we might be able to leverage the computer's strengths to help in the processing. Computers are very good at certain tasks, such as rapidly processing huge amounts of data looking for matching strings or strings satisfying simple criteria. A computer also has the ability to perfectly remember large amounts of information. But what are people actually doing when they look for information? We do not really know the answer. This is a problem in human cognition and its solution promises at some level to give guidance in how to perform better retrieval. For example, XplorMed uses simple statistics to provide terms that may be useful in refining a search. Is this really something that fits well what a user is trying to accomplish when he is searching? We do not know the answer to this, but it clearly would be helpful to know. We may hope that future research on human cognition will provide some answers.

Finally one of the important unsolved problems in this area of research is how to measure success. How can one accurately measure the utility of such a complicated system? Clearly successful usage will depend to a large extent on the knowledge and skill of the operator. Also one strategy for the use of a system may not be as good as another. Such heterogeneity makes

comparison of different systems difficult. As far as we are aware there are no published formal evaluations of the systems presented here.

REFERENCES

Belew, R. K. and Chang, M. (2004). "Purposeful Retrieval: Applying Domain Insight for Topically-focused Groups of Biologists," Paper presented at the *Search and Discovery in Bioinformatics: SIGIR 2004 Workshop*.

Blair, D. C. and Maron, M. E. (1985). "An Evaluation of Retrieval Effectiveness for a Full-text Document-retrieval System," *Communications of the ACM*, 28(3), 289-299.

Croft, W. B. and Harper, D. J. (1979). "Using Probabilistic Models of Document Retrieval Without Relevance Information," *Journal of Documentation*, 35(4), 285-295.

Fuller, S., Revere, D., Bugni, P., and Martin, G. M. (2004). "Telemakus: A Schema-based Information System to Promote Scientific Discovery," *Journal of the American Society for Information Science and Technology*, In press.

Funk, M. E., Reid, C. A., and McGoogan, L. S. (1983). "Indexing Consistency in MEDLINE," *Bulletin of the Medical Librarians Association*, 71(2), 176-183.

Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987). "The Vocabulary Problem in Human-System Communication," *Communications of the ACM*, 30(11), 964-971.

Harter, S. P. (1975). "A Probabilistic Approach to Automatic Keyword Indexing: Part I. On the Distribution of Specialty Words in a Technical Literature," *Journal of the American Society for Information Science*, 26, 197-206.

Humphreys, B. L., Lindberg, D. A., Schoolman, H. M., and Barnett, G. O. (1998). "The Unified Medical Language System: An Informatics Research Collaboration," *Journal of the American Medical Informatics Association*, 5(1), 1-11.

Kurland, O. and Lee, L. (2004). "Corpus Structure, Language Models, and Ad Hoc Information Retrieval," Paper presented at the *ACM SIGIR 2004*.

Maron, M. E. and Kuhns, J. L. (1960). "On Relevance, Probabilistic Indexing and Information Retrieval," *Journal of the ACM*, 7(3), 216-243.

Perez-Iratxeta, C., Keer, H. S., Bork, P., and Andrade, M. A. (2002). "Computing Fuzzy Associations for the Analysis of Biological Literature," *BioTechniques*, 32, 1380-1385.

Perez-Iratxeta, C., Perez, A. J., Bork, P., and Andrade, M. A. (2001). "XplorMed: A Tool for Exploring MEDLINE Abstracts," *TRENDS in Biochemical Sciences*, 26(9), 573-575.

Perez-Iratxeta, C., Perez, A. J., Bork, P., and Andrade, M. A. (2003). "Update on XplorMed: A Web Server for Exploring Scientific Literature," *Nucleic Acids Research*, 31(13), 3866-3868.

Ponte, J. M. and Croft, W. B. (1998). "A Language Modeling Approach to Information Retrieval," Paper presented at the *SIGIR98*, Melbourne, Australia.

Robertson, S. and Hiemstra, D. (2001). "Language Models and Probability of Relevance," Paper presented at the *First Workshop on Language Modeling and Information Retrieval*, Pittsburgh, PA.

Robertson, S. E. and Sparck Jones, K. (1976). "Relevance Weighting of Search Terms," *Journal of the American Society for Information Science*, May-June, 129-146.

Robertson, S. E. and Walker, S. (1994). "Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval," Paper presented at the *17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Salton, G. (1975). *A Theory of Indexing* (Vol. 18). Bristol, England: J. W. Arrowsmith, Ltd.

Salton, G. (1991). "Developments in Automatic Text Retrieval," *Science*, 253, 974-980.

Salton, G., Wong, A., and Yang, C. S. (1975). "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, 18, 613-620.

Salton, G. (Ed.). (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*, Englewood Cliffs, NJ: Prentice-Hall, Inc.

Saracevic, T. (1991). "Individual Differences in Organizing, Searching, and Retrieving Information," Paper presented at the *Proceedings of the 54th Annual ASIS Meeting*, Washington, D.C.

Shieber, S. M. (1994). "Lessons from a Restricted Turing Test," *Communications of the ACM*, 37(6), 70-78.

Silverstein, C. and Henzinger, M. (1999). "Analysis of a Very Large Web Search Engine Query Log," *SIGIR Forum*, 33(1), 6-12.

Sparck Jones, K. (1972). "A Statistical Interpretation of Term Specificity and its Application in Retrieval," *The Journal of Documentation*, 28(1), 11-21.

Sparck Jones, K., Walker, S., and Robertson, S. E. (2000a). "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments (Part 1)," *Information Processing and Management*, 36, 779-808.

Sparck Jones, K., Walker, S., and Robertson, S. E. (2000b). "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments (Part 2)," *Information Processing and Management*, 36, 809-840.

Sparck-Jones, K. (2001). "LM vs PM: Where's the Relevance?" Paper presented at the *First Workshop on Language Modeling and Information Retrieval*, Pittsburgh, PA.

Swanson, D. R. (1988). "Historical Note: Information Retrieval and the Future of an Illusion," *Journal of the American Society for Information Science*, 39(2), 92-98.

Wilbur, W. J. (1998). "The Knowledge in Multiple Human Relevance Judgments," *ACM Transactions on Information Systems*, 16(2), 101-126.

Wilbur, W. J. and Coffee, L. (1994). "The Effectiveness of Document Neighboring in Search Enhancement," *Information Processing and Management*, 30(2), 253-266.

Witten, I. H., Moffat, A., and Bell, T. C. (1999). *Managing Gigabytes* (Second ed.), San Francisco: Morgan-Kaufmann Publishers, Inc.

Zaragoza, H., Hiemstra, D., and Tipping, M. (2003). "Bayesian Extension to the Language Model for Ad Hoc Information Retrieval," Paper presented at the *SIGIR'03*, Toronto, Canada.

Zhai, C. and Lafferty, J. (2004). "A Study of Smoothing Methods for Language Models Applied to Information Retrieval," *ACM Transactions on Information Systems*, 22(2), 179-214.

Zobel, J. and Moffat, A. (1998). "Exploring the Similarity Space," *ACM SIGIR Forum*, 32(1), 18-34.

SUGGESTED READINGS

van Rijsbergen, C. J. (1979). *Information Retrieval*, Second Edition, London: Butterworths.
A classic in the field. Gives a highly readable account of fundamental topics such as indexing, file structures, clustering, term dependencies, probabilistic methods, and performance evaluation.

Salton, G. (1989). *Automatic Text Processing*, New York: Addison-Wesley.

The book has four parts and the third part of the book consists of three chapters on various aspects of information retrieval. The emphasis is on the vector model and tfxidf weighting, methods largely developed by the author and his students at Cornell University.

Sparck Jones, K. and Willet, P. (Eds.). (1997). *Readings in Information Retrieval*, San Francisco: Morgan Kaufman.
An important resource reprinting many of the most important papers detailing significant advances in the science of information retrieval over the years.

Witten, I. H., Moffat, A., and Bell, T. C. (1999). *Managing Gigabytes*, Second Edition, San Francisco: Morgan Kaufmann.
A very good treatment of basic vector information retrieval for text. Also an important resource for those who must manage large text files as it emphasizes compression methods and their practical implementation to construct digital libraries.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*, New York: Addison-Wesley.
A wide ranging coverage of all the basic approaches to information retrieval except the language modeling approach. Includes extensive treatment of user interfaces, the internet, and digital libraries.

Belew, Richard K. (2000). *Finding Out About*, Cambridge: Cambridge University Press.
A good introductory text with an emphasis on the World Wide Web and artificial intelligence.

ONLINE RESOURCES

Information Retrieval Links: Lists many resources related to the field of information retrieval. Included are links to access the software for the SMART retrieval system developed by Gerard Salton and his students:
http://www-a2k.is.tokushima-u.ac.jp/member/kita/NLP/IR.html

Information Retrieval Software: This site provides links to information retrieval software (some as freeware), to internet search engines and web directories, and to search engine optimization sites: http://www.ir-ware.biz

The Apache Jakarta Project: Jakarta Lucene is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform. Jakarta Lucene is an open source project available for free download from Apache Jakarta:
http://jakarta.apache.org/lucene/docs/index.html

QUESTIONS FOR DISCUSSION

1. Describe how one might use the World Wide Web to construct a representation for the information need states $\{S_i\}$. How might one estimate $p(S_i)$ and $p(d \mid S_i)$ used in equation (2) for such a model?

2. In choosing the features to represent documents in the vector method of retrieval it is generally found that single words work as well or better than single words plus phrases. Provide what you believe could be an explanation for this phenomenon.

3. In the MEDLINE database each document has on the average about a dozen MeSH headings assigned to help characterize the subjects discussed in the document. These MeSH terms make useful features for retrieval but they involve a significant expense and human effort to assign. Describe what you think the barriers are to making these MeSH assignments more useful.

4. The MEDLINE record of a document does not contain the list of citations or references which generally appear at the end of a document. However, some databases do have such information. Describe how these citations could be used as features in a vector retrieval system along with the words in the text. How would you weight them?

5. Describe one method that you feel would be appropriate to evaluate the Telemakus system and how this method could be used to make decisions about use of the system.

6. Describe one method that you feel would be appropriate to evaluate the XplorMed system and how this method could be used to make decisions about use of the system.

7. Some retrieval systems, such as ABView:HivResist, attempt to use graphical displays of documents in space to convey information. What do you see as problems with this approach? How do you think the relationships between documents could best be represented graphically or otherwise?