Personalized Search of the Medical Literature: An Evaluation

ABSTRACT

We describe a system for personalizing a set of medical journal articles (possibly created as the output of a search engine) by selecting those documents that specifically match a patient under care. Key element in our approach is the use of targeted parts of the electronic patient record to serve as a readily available user model for the personalization task. We discuss several enhancements to a TF*IDF based approach for measuring the similarity between articles and the patient record. We also present the results of an experiment involving almost 3,000 relevance judgments by medical doctors. Our evaluation establishes that the automated system surpasses in performance alternative methods for personalizing the set of articles, including keyword-based queries manually constructed by medical experts for this purpose.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering, selection process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

Keywords

Evaluation, Search, Re-ranking, Personalization, Natural Language Processing, Medical digital library

Eligible for Best Student Paper Award: No.

1. INTRODUCTION

Medical care providers often have to retrieve pertinent information for patients under their care from the online literature (whether specialized collections or internet databases). This task can be performed when the patient's prior medical history becomes first available; before or after a significant operation or change in the patient's status; and even, if technology allows, during the provision of care (e.g., in the doctor's office or the operation room). Searching databases of published results is a major activity for doctors in training (medical students, residents, and interns), and an activity in which even experienced doctors have to engage in order to keep up with the latest results.

The difficulty of the task is compounded by the large amount of information potentially relevant to a given patient. For example, in the narrow field of cardiac anesthesiology there are five regularly published scientific journals; but relevant information may appear in any of the 35 journals in anesthesiology, the 60 journals in cardiology, the 40 journals in cardiothoracic surgery, or even the more than 1,000 journals in the general field of internal medicine. On the other hand, a doctor querying an article database will often find that while an article may be relevant to a procedure or medication he or she is considering, the patient's particular circumstances limit the usefulness of the article. For example, the article may confine its analysis to specific demographic groups by gender or age, or pertain to patients with specific complicating factors (e.g., diabetes, high blood pressure) or prior medical treatment or incident (e.g., prior bypass or myocardial infarction). Thus, the doctor often has to wade through reams of information she cannot use in order to locate something that applies specifically to the case at hand.

We are developing an information access system, PERSIVAL (PErsonalized Retrieval and Summarization of Images, Video, and Language) for personalizing the retrieval and presentation of information to the needs of a specific patient. The system is intended to be usable initially by the health care specialist and eventually by the patient himself; health care personnel have specialized knowledge and different information needs than patients, so a separate mechanism is needed to filter information from the primary literature for patient use. PERSIVAL [1] uses data from the electronic patient record as a user model, to focus the selection of information to a specific patient. Figure 1 shows the architecture of the system; it contains modules for retrieving data from the electronic patient record, matching and reformulating queries according to rules of evidence-based medicine [13], querying disparate and distributed collections of data with multiple attribute schemas (federated databases), filtering the collected results according to the patient record information, and summarizing the results. PERSIVAL operates on text documents (medical article journals for specialists and articles written for lay persons, as well as support groups and health guides found on the world wide web), images from textbooks and lab procedures (e.g., x-rays), and specialized diagnostic videos (e.g., echocardiograms). It supports interaction in multiple modalities (text, speech, graphical manipulation) and co-ordinates output in multiple media.

In this paper we report on one of PERSIVAL's components, the reranking or natural language filtering module. This module represents the second half of our retrieval strategy

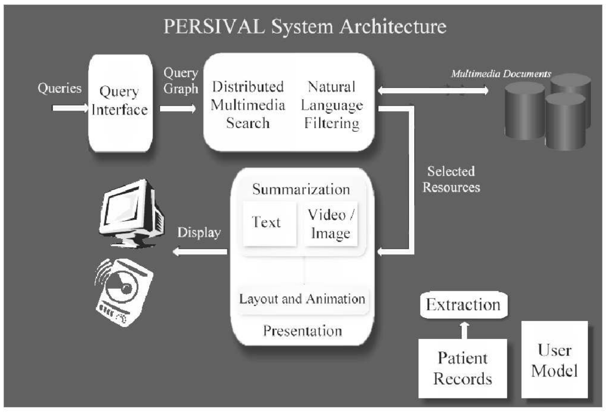

Figure 1: PERSIVAL's architecture.

for locating medical articles relevant to a specific patient. It takes as input a set of text documents deemed relevant to a query and filters that set to those documents that match a specific patient under care. It uses information in the patient record (demographics, prior operations, medications, concurrent diseases) to calculate the degree of match of a document with the patient, independent of the specific query that selected the article. Then the scores of the search engine that are based on the query can be combined with the match factors calculated by this module, resulting in a *re-ranked* list of relevant documents. This list focuses search results on the specific patient; it is used by PERSIVAL's subsequent summarization and presentation modules to generate a multimedia summary tailored to the patient at hand. Thus, not only are summaries tuned to the patient, but computationally intensive language analysis that is necessary for summarization needs be performed only on a small subset of the initial search results.

The patient record consists of multiple reports from different sources (text summaries of the patient's medical history, lab results, results from various diagnostic tests). Rather than using all the text and numbers in the record indiscriminately, we have developed methods to extract medical *terms* on which to base the comparison (e.g., *congestive heart failure*), and associate these terms with values whenever they appear together (e.g., *age–over 50* and *left ventricular ejection fraction–less than 35%*). In addition, we weigh the contribution of each term's occurrence according to several factors: its position in the document (for example, in a medical journal article terms that appear in the "Methods" section are more likely to describe the study population), any negative context (e.g., *no CHF*) that should reverse the contribution of a term to the matching, and the semantic class of a term (e.g., a disease is more important than a body part).

We describe first the algorithms used by our personalization/search filtering module. These include methods to process the electronic patient record and the input documents, identify medical terms, determine section context, negative context and the scope of conjunctions, extract values associated with terms, associate concepts and semantic types with terms, and calculate the degree of match between a document and the patient record. While our techniques can be expanded to general medical source documents, we have to date experimented with articles published in high quality medical journals.[1] We have built a large collection of medical articles, which we detail in Section 3. In Section 4 we discuss how we selected a sizable subset of these articles and had it labeled for relevance to three patients by medical experts. Our presentation of the results of this study follows in Section 5. We then compare the performance of our system to several alternative strategies that represent current search practice at varying levels of sophistication. The evaluation results clearly establish that our system outperforms even experts in the medical domain at the task of selecting the articles relevant to a patient with a single query or multiple but not linked queries.

This paper augments an earlier study of a prototype version of our system [2]. In that earlier study we used fewer patients, fewer reports from each patient record, only one doctor as judge of the relevance of search results, and substantially fewer articles (one tenth of the current study). We also used a more limited version of our system that did not include information about global term rarity or information about values linked to medical terms.

2. SYSTEM DESCRIPTION

2.1 Document Preprocessing

Our system takes as input two distinct pieces of information: published articles from medical journals, and the electronic patient record. Articles are available in several different locations (local databases, specialized collections, and the "deep web", i.e., returned by queries on medical search engines such as PubMed (`http://ncbi.nlm.nih.gov`)). One of PERSIVAL's earlier modules is dedicated to the tasks of locating sources of articles, deciding which ones to use, translating a query to the query language expected by each source, and unifying the returned results [3]. For our purposes, we assume that a set of articles in HTML have been collected and made available to the filtering module. We automatically transform the original HTML documents into a uniform XML format describing the article's structure [4], using a detailed DTD we defined for this purpose. We then use available XML tools such as TTT [8] to tokenize, segment, and annotate the text. This preprocessing needs to be done once for each article, and could be done offline for articles stored in a local collection. The output includes tags that denote section breaks, sentence breaks, tokens, and part of speech information.

Secure access to the electronic patient record is provided via the Clinical Information System in place at New York Presbyterian Hospital (NYPH) [6], which returns a number of reports. Some of these reports contain text (e.g., a patient's prior medical history), some have predefined fields filled in, and some consist of tables of numbers (e.g., microbiology lab reports and blood chemistry panels). Our system processes seven different kinds of reports; since some of these are re-issued at regular intervals (for example, electrocardiograms and chest x-rays are done once daily), the

[1] We intend to address issues of assessing source quality and translating between lay and technical terms, necessary for processing documents in the open web, in future work.

patient records we have worked with can contain upwards of 100 individual reports. For this study we limited the number of individual reports to seven representative reports (one from each category) per patient. We have developed specialized scripts that process each type of report, identify text, field, and tabular regions, and convert the data to a form suitable for preprocessing with the text annotation tools described earlier for the journal articles. The output is again an XML file. All records have been manually sanitized to remove personal identifying information for patients. We are in the process of enhancing the patient record processing with techniques from evidence-based medicine that will extract only the most important parts of the record [11].

2.2 Extracting Terms and Values

A key element of our approach is relying on select pieces of information in the patient record to determine relevance, rather than all words as in the typical information retrieval approach. To this end, we base our comparisons between articles and patient records on *terms*, the technical words and phrases in the domain. We measure similarity between articles and records by their common terms, weighted by contextual and semantic factors.

Rather than using statistical or distributional properties [5, 10] to identify terms, as is common in other applications, we rely instead on extensive knowledge sources that are available for the medical domain. We first process the XML representation of articles and patient records with a finite state grammar we developed, to detect simple (non-recursive) noun phrases. This is done on the basis of automatically assigned part of speech labels, using patterns over adjectives, quantifiers, determiners, and nouns for capturing common types of noun phrases. Each noun phrase is then matched against the Unified Medical Language System (UMLS) [9], a large knowledge base of terms and concepts in the medical domain maintained by the National Library of Medicine (**`http://www.nlm.nih.gov/research/umls/`**). Noun phrases are maximally matched to UMLS terms from the right, since a detected noun phrase can contain additional adjectival modifiers (e.g., "*severe* congestive heart failure", where "congestive heart failure" is the term and *severe* a modifier.

In addition to the noun phrases that are mapped to terms, a separate part of our grammar detects several types of associations between terms and values. We detect quantitative values (numbers, ranges of numbers, and comparative expressions such as "less than 50", all of which can be followed by units of measurement), and qualitative values expressed with adjectives. Values can be linked to a term either with pre-modification, as in the "severe congestive heart failure" example above, or with three kinds of post-modification: copular verbs (*is*, *seems*, *appears*, etc., as in "Blood pressure is 100 mm Hg"), direct comparison operations, or via an *of* prepositional phrase ("ejection fraction of 30%"). Extracted values are standardized in scale and marked in the output of our text analysis subsystem.

The grammar also performs expansion of conjoined terms, so that, for example, "carotid or coronary arteries" is broken down into "carotid arteries" and "coronary arteries". Multiple and nested conjunctions in the same phrase are handled by our system by heuristically ranking the combinations of the different possibilities.

Finally, our system detects and handles abbreviations in a special way. Even though the UMLS database contains many acronyms, its coverage in acronyms is lower than that of the corresponding full terms. Acronyms also show a higher degree of ambiguity concerning their interpretation than full terms do. We expand acronyms using a list of 2,011 acronyms in the cardiology domain collected from the internet, carrying on potential multiple matches for disambiguation at a later stage.

2.3 Determining Context

We weigh the significance of each term in a potential match by two contextual factors: its position within the article, and its potential occurrence in negative context (i.e., under negation or in the scope of an expression with exclusive effect). Our motivation for including position information is that certain sections in the highly structured medical articles are more likely to include descriptions of the patients that the article is about. For example, the "Methods" section is likely to list inclusion and exclusion criteria for the population in a study.

Section bounds are automatically detected during our preprocessing of the articles. We have assigned to each section type a weight in consultation with medical experts.

Negative context can also influence the match, as we do not want to select an article that mentions that "patients without myocardial infarction were sampled ... " if the patient record includes the term "myocardial infarction" or its synonyms (e.g., "MI"). As part of our finite state grammar, we identify direct negative operators (*no*, *none*, *without*, etc.). We also capture nine syntactic patterns for exclusion criteria, e.g., "exclusion criteria were ... " or "patients with ... were excluded".

The contribution of each term to the match is modified according to any detected occurrences in negative context. If a term occurs in normal (positive) context in the patient record and the article, or in negative context in both, there is no modification. If, however, the term occurs in negative context in one and positive context in the other, its contribution is reversed (subtracted from the running total of the match rather than added; see equation (2) in Section 2.5). When a term occurs multiple times in an article or patient record, we consider each combination of occurrences of the term (one in the article and one in the patient record), calculate the match contribution according to the above algorithm, and then average the match results across all such combinations for that term. In this manner, terms that occur in the same positive or negative context in both article and patient record contribute to a high degree of match, while terms that appear in different contexts actively penalize the match, helping prevent spurious matches between a term and its negated counterpart.

2.4 From Terms to Concepts

A well-known problem in information retrieval is the ambiguity of terms, and conversely, the existence of multiple terms that all refer to the same concept. For example, "MI" commonly refers to "myocardial infarction" (heart attack) in the cardiology domain, but occasionally may refer to "Mullerian duct syndrome". Conversely, the concept of "myocardial infarction" can appear as any of the surface terms "myocardial infarction", "infarct", or "MI". We therefore need to map different variants of the same concept to a common concept representation, and also disambiguate terms that

UMLS Semantic type	Weight
Disease or Syndrome	1
Therapeutic or Preventive Procedure	0.8
Diagnostic Procedure	0.6
Sign or Symptom	0.6
Laboratory or Test Result	0.4
Medical Device	0.4
...	...
Molecular Function	0
Cell Component	0

Table 1: A few of the semantic weights used by the filtering component.

can refer to multiple concepts.

For the first task, we again utilize the knowledge in the UMLS database. The UMLS links each term it contains to an underlying "semantic concept", to which a unique identifying number (CUI) is assigned. We use this mapping to handle synonymous terms. UMLS also provides a hierarchical organization of concepts, and a broad semantic class for each concept (such as "Disease or Syndrome" or "Diagnostic Procedure"). We use these class labels to weigh the relative importance of matching concepts, utilizing the weights in Table 1. These weights were derived in consultation with medical experts and from experimentation on earlier, held-out data.

For terms that have multiple concepts associated with them in the UMLS, or concepts with multiple semantic classes, we have implemented a local disambiguation procedure. The first disambiguation level takes the CUIs associated with a term, and retains those with the highest frequency of occurrence within the document being examined (article, or collective patient record). The reason behind this is that concepts expressed using one term are likely to also be expressed using another equivalent term within the same document. While one or both of these terms may be ambiguous (associated with multiple concept IDs), both terms will contribute to the frequency of the "correct" concept ID. It is less likely that other terms associated with the "incorrect" concepts will occur in the same document. Thus the frequency of the correct concept is expected to be higher than those of incorrect concepts that share the same term.

If two or more concepts are tied after the above calculation, the highest frequency concepts are retained for the second level of disambiguation. In that level, we look at the semantic types associated with the retained concepts, and select the CUI that is associated with the medically important semantic types (those with highest weight). If only one concept was the output of the first stage but it is associated with multiple semantic types, the same procedure is used to choose the most important semantic type. This ensures that important concepts contributing significantly to patient-article ranking will not be overlooked. If there are terms that are still ambiguous at this point, the first concept ID and its first associated semantic type are chosen.

2.5 Putting it All Together: The Match Formula

After the previous stages, we have extracted from each of the articles and the patient record a set of terms, with associated values, disambiguated concept identifiers, semantic weights, section weights, and exclusion contexts (values, section weights, and exclusion contexts may vary for each occurrence of the same term in the same article or patient record). To combine all this information into a single number, we first construct TF*IDF vectors of the terms in the article and patient record, and start with a simple cosine formula that measures their similarity [14]:

$$\frac{\sum_i a_i \cdot p_i \cdot \log^2(\frac{N}{DF(i)})}{\sqrt{\sum_i (a_i \cdot \log(\frac{N}{DF(i)}))^2} \cdot \sqrt{\sum_i (p_i \cdot \log(\frac{N}{DF(i)}))^2}} \qquad (1)$$

where a_i is the number of occurrences of term i in the article, p_i the number of occurrences of the term in the patient record, DF(i) is the number of articles in our collection that contain term i, and N is the total number of articles in the collection from which document frequency is calculated.

We then modify formula (1) to take account of the factors modifying a term's importance. First, we account for the influence of position information by replacing a_i with A_i,

$$A_i = \sum_{j \text{ over all section types}} (a_{ij} \cdot s_j)$$

where s_j is the weight for section type j and a_{ij} is the number of occurrences of term i in section j ($\sum_j s_j = 1$, and $\sum_j a_{ij} = a_i$). In other words, A_i is the normalized frequency of term i in the article according to section.

We further modify the contribution of each term ($A_i \cdot p_i \cdot 2 \log \frac{N}{DF(i)}$) by the following factors:

- t_i, a weight capturing the relative importance of different semantic types.
- n_i, measuring positive and negative context agreement for this term between the patient record and the article. For terms occurring once in the patient record and article, n_i is either $+1$ or -1 depending on whether the terms have been seen in similar (positive/positive or negative/negative) or different exclusion contexts. For terms with multiple occurrences in the patient record, the article, or both, we consider all combinations of these occurrences and average the $+1$ or -1 values assigned to each pair.
- v_i, which captures the similarity between observed values for term i in the article and the patient record.

The motivation and definition for t_i and n_i were discussed earlier. To calculate v_i, we collect all values extracted for that term from the article and the patient record, and consider all the pairs formed by taking one element from each of these lists. For each pair, we determine if the values are fully compatible, partially compatible, or incompatible. Currently, we consider two values fully compatible if they are identical, or if one of them is missing (this can happen, for example, if values have been extracted from only one of the documents).[2] If the values are both present, different, and not numeric, we consider them incompatible. We score fully compatible pairs with 1, and incompatible pairs with

[2]We are reluctant to penalize the match if no values are found in one case, as this may be due to either extraction errors or to the creators of one of the documents being less specific than in the other.

0. If the two values v_1 and v_2 are not identical and they are numeric, we consider them partially compatible and assign to that pair a scaled compatibility rating,

$$\frac{\min(|v_1|, |v_2|)|}{\max(|v_1|, |v_2|)}$$

Once we have calculated the pairwise compatibility of values for all pairs, we then determine v_i as their average. Note that this approach has several limitations, originating from lack of knowledge about both language and the domain. Non-numeric values that are different are not necessarily incompatible (e.g., *high* and *70* may be compatible depending on the term; also two different adjectives may be synonyms). Numeric values do not always represent ratio variables, as is assumed in our formula. We also do not compare ranges of values to single values or to other ranges for overlap (except if they happen to be identical). We plan to address these issues in future work focused on value compatibility.

With the modifications detailed above, our final formula for the degree of match between an article and a patient record becomes

$$\frac{\sum_i A_i \cdot p_i \cdot \log^2(\frac{N}{DF(i)}) \cdot t_i \cdot n_i \cdot v_i}{\sqrt{\sum_i (A_i \cdot \log(\frac{N}{DF(i)}))^2} \cdot \sqrt{\sum_i (p_i \cdot \log(\frac{N}{DF(i)}))^2}} \quad (2)$$

This ranges from -1 to $+1$, with $+1$ indicating total agreement, 0 indicating no overlap in terms between the documents, and -1 indicating active disagreement (i.e., the two documents share a lot of terms and disagree on the exclusion contexts or the values for those terms).

3. THE CORPUS

For the experiments reported here and other uses within the PERSIVAL project, we have collected a large corpus of high quality medical articles. We first determined availability of full-text articles, using a combination of automated web crawling and a licensing agreement with Ovid Technologies, a major publisher of medical journals. We then restricted this list of journals, those for which we could obtain full text articles, to a subset according to quality. We measured quality by a journal's impact factor [7] and incorporated suggestions by medical doctors and librarians on our team regarding which journals to include. In this manner, we selected 20 journals in the field of cardiology, from which we collected all articles between 1993 and 2000 which were electronically available.[3] This resulted in a collection of 29,784 articles containing 88,944,123 word tokens.

4. EVALUATION METHODOLOGY

We have designed an experiment to measure the performance of our system in selecting relevant articles. Several factors related to external evaluation resources (e.g., time) limit the scope of the experiment. Chief among these is the difficulty in obtaining properly qualified medical specialists that would rate the relevance of each article in an evaluation set to each patient. Although we do have access to a pool of residents, interns, and attending physicians at NYPH, their availability to participate in repeated ratings of articles is not inexhaustible.

[3]Excluding a few articles that our preprocessor failed to convert to XML due to HTML idiosyncrasies.

Patient A: Patient is a 45 year old female who came to the hospital because of shortness of breath, increasing dyspnea and chest pain. She had atrial fib. Her respiratory status acutely decompensated and she was intubated and emergently transferred to the OR for LVAD placement. On arrival to the OR it was determined that the patient was in cardiogenic shock with a MAP of 55, PCW of 45, cardiac index of 0.9 and on maximal cardiotonic drip support.

Patient B: Patient is a 47 year old man with recent MI complicated by cardiogenic shock requiring placement of intra-aortic balloon pump. He has a history of chronic renal failure, hypertension treated with atenolol, hypercholesterolemia, previous silent MI's by EKG and a family history of coronary artery disease. He went into the Emergency Room where he was found to have poor R wave progression on EKG and Q's in II, III and F.

Figure 2: Summaries compiled from two of the patient records.

To obtain realistic evaluation parameters and still have a sizable collection of articles, we considered a relatively small number of patients, reduced the number of articles in our evaluation universe (the set of articles for which relevance judgments will be sought), and performed a query-independent evaluation. Each of these points is elaborated below.

Three Patients We potentially have access to thousands of electronic patient records, each often containing over 100 individual reports. However, manually sanitizing hundreds of pages for each patient record to satisfy patient confidentiality requirements is a time-consuming task. Since testing our system on thousands of patients is impractical, we limited our experiment to three patients who exemplified significantly different circumstances of the same cardiac disease (unstable angina). For example, one of these patients (patient A) had a left ventricular assist device (LVAD, commonly known as pacemaker) implanted, while B has a history of silent myocardial infarctions (heart attacks that cause no pain and are undetected by the patient) and recently had an intra-aortic balloon pump implanted, and C has atrial fibrillation and underwent a maze procedure in the past. Figure 2 lists short summaries of the patient records for two of these three patients. The patients were chosen among many others by the physician whose care they were under at NYPH.

Query-Independent Evaluation We plan to use our system within PERSIVAL to filter the results that the search module produces from user-specified queries. The present personalization module limits the original search results, keeping the articles that match both the query and the patient record. However, evaluating the system's performance on several queries multiplies by the same number the total number of judgments needed by the doctors. Further, a fair evaluation requires a large number of queries to eliminate effects that the queries themselves may impart. We chose instead to evaluate our system in query-independent mode: Rather than measuring how well it filters the results for any particular query, we measure how well it selects articles that are in general relevant to the patient record for any conceivable query. Articles can be usefully retrieved for different purposes, including prognosis, diagnosis, and treatment, and we instructed the evaluators to consider an article as relevant when it would be relevant for any purpose they could think of. In this way, we separate query relevance from patient relevance and keep the number of judgments manageable.

Evaluation Universe Since it would be impossible (at least for a single-institution study) to find enough doctors to rate all 29,784 articles in our corpus for each of the three patients, we selected a smaller but still sizable part of the corpus as our evaluation universe. We applied the following procedure to guarantee that (a) the selected subset would include articles relevant for many different purposes, and (b) the subset would include a reasonable number of relevant articles (proportionally more than the original corpus) to make the measurement of recall reliable with a smaller universe size.

For the first purpose, we generated random pairs of terms from each of the three processed patient records, plus one of the *article type* keywords "treatment", "diagnosis", or "prognosis", and submitted the resulting three-term query to the Lucene search engine (**`http://jakarta.apache.org/lucene/docs/index.html`**), indexed over our collection of 29,784 articles. We selected the three article type keywords above to produce representative strata of the articles for different purposes, as these keywords correspond to the three most common high-level goals for which medical experts search article databases. The selection of random terms from the patient record captures knowledge of the patient's situation but no relative ranking of the importance of the various terms—in this way, it is akin to what a lay person or incoming medical student might do with the patient record information. For each of the nine subgroups defined in that manner (three patients times three article type keywords), we collected 110 articles from our corpus by submitting repeated random term pairs plus the article type keyword and keeping no more than 10 articles from each term pair. This was done to ensure, from a perspective additional to the three article types, that articles would be collected for a variety of purposes and that no term pair would dominate. After removing the duplicates among these nine sets of 110 articles each, we were left with 911 articles which we included in our universe.

To satisfy the second goal of including in the evaluation set at least some definitely relevant articles, an expert in cardiac anesthesiology[4] selected a few highly relevant articles for each patient out of our corpus of 29,784 articles. The expert constructed several queries involving up to six hand-selected terms, submitted them to the search engine, repeated the process revising the results, and filtered the final returned articles after reading them. Thus we obtained 8 articles for patient A, 9 for patient B, and 11 for patient C, which we added to the universe.

We subsequently asked doctors in NYPH to rate each of the 939 ($= 911 + 28$) articles in the universe for each of the three patients. We separated the articles in the universe into 19 piles of 45 to 50 articles each, randomly assigning articles from each stratum in the universe (defined by the patient and either the selected article type keyword or the expert) to each pile while balancing the number of articles from each stratum in each pile. Nine doctors participated in the experiment, each rating every article in one or more piles for each of the patients. Most did one pile only, but several did more and one physician did a total of ten. They included two residents, six attending physicians (all specialists in cardiac anesthesiology) and one physician's assistant. The doctors were given the full patient record, clinical summaries like the ones in Figure 2, and the full text of the articles, and were instructed to rate the relevance of each article to each patient on a scale 0 to 5, under the general, query-independent notion of relevance discussed earlier. In addition, we had three of the 19 piles rated twice, for the purpose of measuring agreement between doctors who rate the same articles. We use the collected judgments to rate the performance of our system and of alternative retrieval methods.

[4]The fourth author of this paper.

5. RESULTS

We measure the performance of our system by comparing its output for each patient to the judgments provided by doctors (when we have multiple judgments for an article and patient, we average the two doctors' scores). Neither the system's output nor the doctors' judgments are binary. The system produces a graded value between -1 and $+1$ (although in practice the values are almost always positive), and the doctors rate relevance on a scale between 0 and 5. We can set a threshold, t_d, on the doctors' scale for considering an article truly relevant; high thresholds correspond to more strict definitions of relevance. However, when multiple doctors participate in the evaluation, the meaning of the relevance scores is dependent on each evaluator; for example, we noticed that one of the doctors in our study rarely gave a relevance score higher than 3, while another doctor often assigned scores of 5. Doctors also have different spreads in their scores—some tend to use the evaluation scale more uniformly than others. To adjust for these dependencies on the evaluators, we replace each relevance judgment x_{ij} by evaluator i on article/patient combination j with the normalized value

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \tag{3}$$

where μ_i is the mean of doctor's i scores, and σ_i is their standard deviation. We calculate the mean and the standard deviation in equation (3) separately for each patient, to account for the possibility that the true proportion of relevant articles in our evaluation universe differs from patient to patient. A normalized score z_{ij} of 0 indicates that the doctor assigned to this article an average score, while positive scores indicate a more relevant than average article.

We therefore apply the threshold t_d on the normalized values of the experts' judgments. We also convert the system's scores to "yes"/"no" decisions by applying another threshold, t_s. Articles receiving a match score above t_s are treated as articles selected by our system for that patient. t_s controls the system's propensity to mark articles as relevant to a given patient, and has the usual effect on the recall/precision tradeoff (high t_s will give high precision, and low t_s high recall).

We ran our system with all possible combinations of features on top of the basic matching formula (equation (1)), and evaluated against the evaluation universe described in

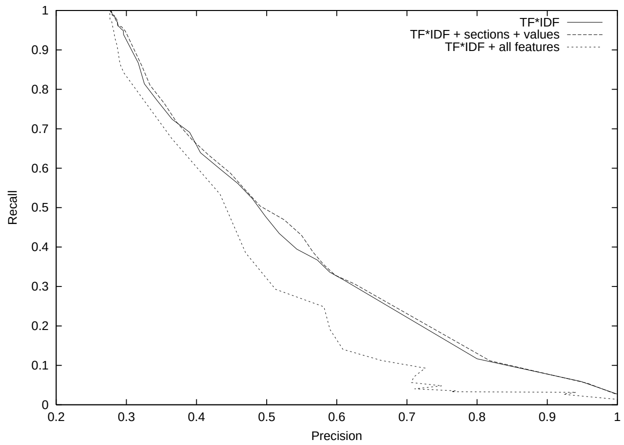

Figure 3: Precision-recall curves for three combinations of features (TF*IDF, TF*IDF plus section and value weighting, TF*IDF plus all our features) on average across all patients ($t_d = 0$).

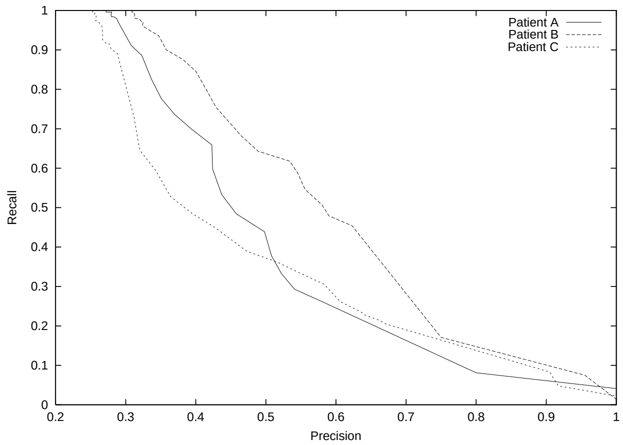

Figure 4: Precision-recall curves for the combination TF*IDF plus sections plus values, drawn separately for each of the three patients ($t_d = 0$).

Section 4. By comparing the performance of different combinations of features for different patients, as well as the average across patients, we noted considerable variability in the combination that performed best. The basic formula (equation (1)) of TF*IDF performed well, but was not the top one. On the other hand, features such as section weights and use of values had a generally positive effect on the evaluation scores, but sometimes their combination resulted in lower scores than TF*IDF alone. Figure 3 shows precision-recall curves for three combinations of features: basic TF*IDF, sections and values in addition to TF*IDF, and all four of our features plus TF*IDF, as an average across all three patients. The three curves are fairly close together. Contrast this with Figure 4, which displays separate precision-recall curves for the combination of TF*IDF, sections, and values for each patient. It appears that the effect of the patient on the scores is larger than the effect of the different features, and with only three patients in our sample, we may not be able to detect benefits of the features.

Table 2 lists the numeric scores obtained by the system for the above and additional combinations of features at the value of the threshold t_s that maximizes F-measure for each combination. This threshold value can be estimated during training, provided that we have a large enough sample of patients, or adjusted interactively during retrieval. We experimented with different values of the threshold t_d on the doctors' judgments—obviously higher values focus the results on the cases where the doctors are most sure of relevance, but also limit the number of articles considered relevant. We report results with $t_d = 0$ in Figures 3 and 4 and in Table 2 (i.e., we consider an article relevant if it exceeds the doctor's average rating across all articles for that patient).

The relative lack of differentiation in the results according to features is not what we expected, as intuitively we anticipated that semantic weights, values, negative context, and section weights would offer a larger gain to the system. We had also observed far more significant effects from some of these features, particularly semantic weights, in an earlier study [2]. In that study we conducted an evaluation against a separate, smaller document set, which included 93 articles generated by a combination of the random term pairs method and expert judgments for two of the patients. The variability of the results between the two studies and according to the set of articles rated, the evaluators, and the patients indicates that further study is needed to fully determine the effect of these features.

Regardless, the numbers in Figures 3 and 4 and in Table 2 show that the system does quite well in absolute terms. It achieves 40–45% recall and 60–65% precision, for an F-measure of about 50%. To further investigate whether these promising results are due to our use of the full patient record information, or a characteristic of this particular retrieval task, we compare the scores obtained by our system to those of several alternative search strategies. We first examine the random term pairs strategy, described in Section 4 and used in the construction of our evaluation universe. This strategy employs one of the article type keywords "treatment", "diagnosis", or "prognosis" (major classifiers of article type in the medical domain) plus two random terms from the patient record. Hence it utilizes the same kind of knowledge about the patient that our system has, only without modeling context, values, and semantic types, and without considering all the terms together. Rather, it simulates a user with access to the patient record but no ability to rank the relative importance of each term. We evaluate this strategy by taking the 330 articles it produced for each patient using the Lucene search engine and comparing them to the corresponding gold standard implied by the doctors' ratings for $t_d = 0$. The results, shown in the first row of Table 3, indicate that this strategy performs consistently at the 30% level for all three measures. Our system therefore outperforms this method by doubling the precision and increasing the recall by a relative 50%. This offers evidence that looking at the entire record rather than certain combinations of terms is more effective for retrieval.

We noted earlier that the random term pairs strategy may be an approximation to the strategy a non-expert would use when given the patient record. In order to generate a more sophisticated set of queries, we asked each of the doctors

Feature combination	Max. F-Measure	Precision at Max. F-Measure	Recall at Max. F-Measure
All patients			
TF*IDF	50.27%	61.00%	42.83%
TF*IDF, Sections, Values	50.87%	57.57%	46.13%
TF*IDF, Semantics, Context	47.57%	59.13%	40.07%
TF*IDF, all features	48.13%	77.27%	37.30%
Patient A			
TF*IDF	50.40%	65.40%	41.00%
TF*IDF, Sections, Values	51.50%	65.90%	42.30%
TF*IDF, Semantics, Context	49.70%	58.10%	43.50%
TF*IDF, all features	51.30%	58.10%	45.80%
TF*IDF, sections, values across patients			
Patient A	51.50%	65.90%	42.30%
Patient B	57.30%	61.80%	53.40%
Patient C	43.80%	45.00%	42.70%

Table 2: Evaluation scores for several feature combinations on all patients, the same combinations for patient A, and one of the combinations for different patients (in all cases, $t_d = 0$).

Alternative Search Strategy	F-Measure	Precision	Recall
Random term pairs	29.63%	29.02%	30.41%
Doctor queries submitted during evaluation (individual average)	17.84%	57.40%	19.28%
Doctor queries submitted during evaluation (OR'ed by patient)	39.44%	29.44%	39.44%
Queries produced from expert strategy (individual average)	22.62%	56.21%	25.33%
Queries produced from expert strategy (OR'ed by patient)	36.38%	36.93%	58.04%

Table 3: Evaluation scores for several alternative search strategies on all patients (in all cases, $t_d = 0$). All measures are averaged, either over all queries separately or over the three patients, so the average F-measure can be lower than both the average precision and the average recall.

participating in the study to construct a query for each patient that they would use to retrieve relevant articles from a search engine. This was done before the doctors saw the articles in our evaluation universe that they marked as relevant or not. Doctors were restricted to simple "and" operations between terms of their choice (which could be chosen from their expertise and did not have to appear in the patient record). Five of the nine doctors provided us with queries, which employed between 2 and 6 terms and varied in their level of specificity (e.g., from the very generic `abdominal pain` to the very specific `long term LVAD rematch trial`). The third row in Table 3 lists the scores obtained by this strategy, on average for each of the 15 queries generated by the doctors, and also when we take the union (i.e., OR) of the results of the 5 queries for each patient and average the scores across the three patients. As each doctor may have targeted a particular subclass of articles with their query, we expected high precision and relatively low recall for most of the individual queries. We also expected that the union of the queries for each patient would increase the recall but sacrifice some precision as what is relevant would vary from doctor to doctor. These expectations were validated during the evaluation (middle part of Table 3), with the individual doctors queries achieving respectable precision (in the 50–60% range and occasionally higher) but very low recall (even into single digits), while their union achieved higher recall with lowered precision (40% and 30% respectively). What is notable is that our system surpassed in performance both modes of using the doctors' queries. The latter can be assumed to be representative of typical queries that experts issue, establishing that our system is effective in utilizing the patient record information to tailor the search better than even experts' queries do.

We also compared our results against a search strategy defined by a senior medical expert. As before, the search strategy is single-shot, without allowing feedback in the search—this is done to avoid comparing with strategies that actively involve a human in the loop and therefore cannot be performed at least semi-automatically (i.e., with limited initial human input). The expert first extracted all valid terms from each patient record and selected the most important ones. Then we constructed a variety of searches by selecting two to four terms from that list, each of which came from different Medical Subject Headings (MeSH) [12] categories. MeSH provides a hierarchical classification of medical terms, so terms in different parts of the tree at the top level express very different types of information. It was the expert's expectation that this would help yield results that are more specific to the patient record. Each combination of terms gives a different feasible question related to the patient. The union of all query results from queries constructed this way would come as close as possible to covering all aspects of the record. We constructed five queries per patient, and evaluated them individually, and as an OR'ed set. As shown in the last part of Table 3, this strategy outperforms the other two doctor-based strategies on both precision and recall. It achieves higher recall than our totally automatic method but lower precision and F-measure.

6. CONCLUSION

We have presented a system for the filtering of medical journal articles to select those that match most closely a specific patient. Our system utilizes information from the online patient record, working with targeted pieces of text (medical terms) rather than all the words in it. The system also includes a number of linguistically motivated features (weighting by semantic type of the terms, by term position in various sections in an article, by exclusion/inclusion context, and by the compatibility of linked values).

We planned the evaluation experiments reported in this paper with two goals. First, we wanted to establish that using the patient record to form essentially a very complex query is more effective than creating one or multiple smaller queries, and that this strategy performs well enough to be useful to medical specialists. Our second target was to show that the additions to the standard information retrieval model via the four specialized weighting mechanisms contributed to increased performance of the system. Unfortunately, although intuitively we expect those features to be indeed useful, the experimental results are inconclusive on that account: While the additional features most times improve the evaluation scores (see Table 2), they also often decrease them. This may be in part due to imperfections in the modeling or extraction of the features, or the particular dataset analyzed (in particular, the patient records of which we had only three). In any case, we plan to further analyze the available data and complement the current experiment with smaller targeted ones to collect additional evidence on the potential usefulness of the features.

Despite the inconclusive outcome for the second part of our experiment, the primary goal of showing that the patient record can be effectively used for personalizing the search has met with unqualified success. We obtained high precision and recall scores, which compare favorably with all alternative search strategies we considered. These strategies include queries prepared by expert doctors who not only have access to the full patient record, but also to medical knowledge, both internalized and formalized in the MeSH taxonomic system. Despite this, the system was shown to perform better in the filtering task than a single query from a doctor (or a collection of queries across doctors) would do. This is not to say that our system would achieve equivalent performance with the experts in a session where the doctors would receive feedback from their first query—in such an environment the doctors would produce multiple queries, each dependent on the results of the earlier ones, and hone a search strategy adapted to the data in the patients' files. We plan to explore this direction as well, by incorporating feedback mechanisms into our system as part of the larger PERSIVAL system.

7. REFERENCES

[1] Author names witheld. Title withheld. In *Proceedings of The First ACM+IEEE Joint Conference on Digital Libraries*, Roanoke, West Virginia, June 2001.

[2] Author names withheld. Title withheld. In *Proceedings of the American Medical Informatics Association Fall Symposium*, Washington D.C., November 2001.

[3] Author names withheld. Title withheld. In *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries (JCDL 2001)*, Roanoke, Virginia, June 2001.

[4] Author names withheld. Title witheld. In *Proceedings of LREC-2002*, 2002.

[5] Kenneth W. Church. Empirical estimates of adaptation: The chance of two Noriegas is closer to $p/2$ than p^2. In *Proceedings of COLING*, Saarbrücken, Germany, 2000.

[6] P. D. Clayton, R. V. Sideli, and S. Sengupta. Open architecture and integrated information at Columbia-Presbyterian Medical Center. *MD Computing*, **9**(5):297–303, 1992.

[7] E. Garfield. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, **122**:108–111, 1955.

[8] Claire Grover, Andrei Mikheev, and Colin Matheson. LT TTT version 1.0: Text tokenisation software. Technical report, Human Communication Research Centre, University of Edinburgh, 1999. `http://www.ltg.ed.ac.uk/software/ttt/`.

[9] B. L. Humphreys, D. A. B. Lindberg, H. M. Schoolman, and G. O. Barnett. The Unified Medical Language System: An informatics research collaboration. *JAMIA*, **5**:1–11, 1998.

[10] J. S. Justeson and S. M. Katz. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, **1**(1):9–27, 1995.

[11] E. Mendonca, J. Cimino, and S. Johnson. Using narrative reports to support a digital library. In *Proceedings of the American Medical Informatics Association Fall Symposium*, 2001.

[12] National Library of Medicine. Medical subject headings, annotated alphabetic list, 1998, August 1997.

[13] David L. Sackett, R. Brian Haynes, G. H. Guyatt, and Peter Tugwell. *Clinical Epidemiology: A Basic Science for Clinical Medicine.* Little, Brown and Company, Boston and Toronto, 2nd edition, 1991.

[14] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, **25**(5):513–523, 1988.