Personalizing Retrieval of Journal Articles for Patient Care

Simone Teufel, Ph.D.[*], Vasileios Hatzivassiloglou, Ph.D.[*], Kathleen R. McKeown, Ph.D.[*], Desmond A. Jordan, M.D.[†], Kathleen M. Dunn[*], Sergey Sigelman[*] and **André Kushniruk, Ph.D.[‡]**

[*]Department of Computer Science
450 Computer Science Building
Columbia University
New York, N.Y. 10027
{teufel, vh, kathy, kdunn, ss1792}@cs.columbia.edu

[†]Departments of Anesthesiology
and Medical Informatics
College of Physicians and Surgeons
Columbia University
New York, N.Y. 10032
daj3@columbia.edu

[‡]Cognitive Studies in Medicine
Centre for Medical Education
McGill University
Montreal, Canada
andrek@mathstat.yorku.ca

We present a system for patient-specific searches on a database of medical journal articles which uses natural language techniques to match search results against patient records. We performed an information retrieval experiment comparing the performance of this system to two strategies, one of which uses extensive medical knowledge, while the other uses the same patient information our system has. The results show that our system is useful in improving recall over the strategy simulating a human specialist, and clearly outperforms the strategy of using the patient record content without intelligent processing.

Introduction

Online search engines are notorious for overloading end users with irrelevant information. In healthcare settings, personalization of search to the individual patient can aid in filtering out unneeded results. Physicians, and in particular, physicians in training, need information that is clinically relevant to the patient under their care. By exploiting the online patient records at New York Presbyterian Hospital (NYPH) [1] as a sophisticated, pre-existing user model, we show how search results can be tailored to the needs of the clinician.

Our approach is implemented as a component of PERSIVAL (PErsonalized Retrieval and Summarization of Image, Video And Language) [7], a system designed to provide personalized access to a distributed digital library of medical literature. PERSIVAL includes facilities for distributed search over a variety of online sources [3]. Personalization is done by reranking the results returned by the search engine using a more intensive natural language analysis of the documents. While others [6] have shown the importance of providing the right information at the right time, they have yet to automate personalization for the patient.

Clinical studies reported in medical journals describe experimental results for a patient study population which is characterized in the article. Thus, we want to rank higher those articles that describe a patient population that is similar to the patient under the clinician's care. We do this by constructing an *article profile* containing a set of terms and values extracted from the article describing the patient study population (e.g., "high blood pressure", "ejection fraction of 30%", "congestive heart failure"). We also construct a *patient profile* by extracting terms and associated values from the patient record. An article is ranked higher when its profile is a better match to the patient profile; a main contribution of our research is on the algorithmic definition of a good match.

We evaluated the results of our re-ranking component by comparing it against two standards: one representing keyword searches on MEDLINE performed by experienced physicians, and another simulating an approach that uses medical terms from the patient's record but little knowledge of their relative significance. Our system clearly outperformed the second standard. It also improved recall relative to the experienced physician.

Algorithm Overview

The clinical scenario shown in Figure 1, based on two patients and six cardiology articles, illustrates our approach. Instead of the patient summaries shown, our system currently uses the full text from seven individual reports in each patient's record, and ultimately will use all the reports in a record. Patients A and B both had unstable angina, but A recently had a left ventricular assist device (LVAD) implanted while B did not undergo surgery. Thus, article 1, which discusses prognosis for patients with unstable angina, is relevant to both A and B, while article 5, which describes treatment using an LVAD, is relevant to A only. This scenario was used to determine what features could be used to best determine a match between article and patient record.

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.

Articles:

1: Clinical Predictors of In-Hospital Prognosis in Unstable Angina

2: ECLA3 Risks and Benefits of Combined Maze Procedure for Atrial Fibrillation Associated With Organic Heart Disease

3: Prognostic Value of Cardiac Troponin T After Noncardiac Surgery: 6-Month Follow-Up Data

4: Primary Pulmonary Hypertension: Improved Long-Term Effects and Survival With Continuous Intravenous Epoprostenol Infusion

5: Implantable Left Ventricular Assist Devices Provide an Excellent Outpatient Bridge to Transplantation and Recovery

6: Myocardial Viability in Patients with Chronic Coronary Artery Disease and Previous Myocardial Infarction: Comparison of Myocardial Contrast Echocardiography and Myocardial Perfusion Scintigraphy

Figure 1: Case scenario summaries compiled from the patient records, and titles of six relevant articles.

Our approach uses an efficient finite state grammar to extract terms, along with associated values, that describe the patient study population. For example, in article 5, patients in the study had evidence of "cardiogenic shock", measured by "capillary wedge pressure > 20 mm HG" and "cardiac index < 2.0 liters/min", all phrases that match the patient description for A shown in Figure 1. Terms may appear in sections of different importance, they may be associated with values, they may appear in a negative context (e.g., listed among the exclusion criteria), or combined in complex conjunctions; all these issues are handled by our term extraction software. Naturally, in addition to extracting terms and values matching the patient record, our procedure also extracts many terms and values that would never be found in the patient record. For example, in article 5, terms such as "pneumatic chamber" or "cam-follower bearings" are used to describe the LVAD. Our matching algorithm must heavily weight the first set of terms, while ignoring the latter set.

The Data

For the experiments reported in this paper and other work in the context of PERSIVAL, we collected a corpus of 29,784 medical articles in full text, either from the web with an automated crawler or via a licensing agreement with Ovid Technologies. The articles appeared in HTML format; we transformed them into XML using a pipeline we developed on the basis of publicly available XML tools. The corpus contains articles from 20 journals in cardiology from 1993 to 2000, comprising roughly 85 million word tokens (cf. Figure 2).

Name of Journal	Articles
Journal of the American College of Cardiology	1,816
Journal of the American College of Surgeons	453
American Heart Journal	926
American Journal of Cardiology	3,135
American Journal of Hypertension	643
American Journal of Medicine	821
American Journal of Surgery	570
Atherosclerosis	1,030
Annals of Thoracic Surgery	3,000
Annals of Vascular Surgery	216
Cardiovascular and Interventional Radiology	169
Circulation	13,516
Cardiovascular Surgery	304
European Journal of Cardio-Thoracic Surgery	1,003
International Journal of Cardiology	258
Journal of Clinical Anesthesia	249
Pediatric Cardiology	364
Trends in Cardiovascular Medicine	100
Thrombosis Research	715
World Journal of Surgery	496
Total	29,784

Figure 2: List of Journals in our Corpus

Extracting Terms and Values

A key element of our approach is to base relevance decisions on important medical terms rather than all words, as search engines typically do. To this end, we need to recognize terms in context, and also handle complexities such as associated values, negative context, conjunctions, and positional information.

We first use a finite state grammar we have developed, which detects noun phrases. Our grammar defines noun phrases as finite patterns over adjectives, quantifiers, determiners, and nouns. This step generates most of the terms in the medical domain, but also generates many phrases that are not medical terms. To solve the overgeneration problem, for each proposed term we consult a medical term database, the Unified Medical Language System (UMLS) [5]. UMLS assigns to each string an internal identifier (Concept Unique Identifier, or CUI). Several different strings that refer to the same concept may share the same CUI, thus linking synonymous terms. For instance, "atrial fibrillation", "auricular fibrillation" and "A-Fib" all share CUI C0004238. At the same time, a term may be

associated with multiple CUIs, if its meaning depends on context. For example, "MI" can mean "myocardial infarction" (C0027051) or "Mullerian duct inhibiting substance" (C0687670). For each CUI, UMLS also returns a *semantic type*, an indicator of the broad semantic class where the concept belongs (e.g., disease, symptom, demographic, time, etc.). We use a subset of the UMLS semantic types, obtained by consulting physicians at NYPH, thus removing terms with semantic types associated with general concepts (e.g., time, persons, and hospital and administrative terms).

Before lookup in the database, noun phrases containing coordination are broken down into smaller noun phrases by multiplying possibilities out; for example, "carotid or coronary arteries" is broken down into "carotid arteries" and "coronary arteries".

Acronyms are given special treatment. 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 (CUIs) 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.

Values associated with terms are identified by a subpart of our finite state grammar which looks for three kinds of context: a)linking verbs (*is*, *seems*, *appears*, ...) in all types of tense and voice combinations b)*of*-constructions ("blood pressure of 90 mm Hg") c) direct comparison operators (e.g., "blood pressure greater than 100 mm Hg").

We also developed methods to handle terms that appear in negative context, which is determined by pattern matching. We identify direct negations of terms, such as in "patients without myocardial infarction..." and "no atrial fibrillation". In addition, we capture exclusion criteria which are often given explicitly, e.g., "exclusion criteria were ...", or "we did not include patients who ..." Negative context information can prevent spurious matches between a term and its negated counterpart.

The article section where a term occurs is determined by an automated analysis of the article's structure. We do this on the assumption that certain article sections (e.g., Methods) are more likely to contain terms that describe the study population.

The patient records, consisting of many parts such as test results, operative reports, and x-ray reports, are processed in the same way. In the future, we plan to test an alternative approach, which uses a detailed semantic grammar to capture many of the idiosyncracies of the sublanguage in patient records, such as doctors' abbreviations, more frequent usage of acronyms, and additional negative constructions (e.g., "tumor can be ruled out"). We will use a version of MedLEE [2] specially tuned to the cardiology domain.

The Matching Algorithm

Once terms are extracted from both articles and patient records, our matching component uses information about shared terms to score numerically the similarity of an article to an patient. This relevance measure is then used to rerank the output of any search engine, so that articles that are ranked higher are selected for retrieval first. Matching is performed at the level of CUIs, including disambiguation of terms and weighting based on factors such as the semantic type of the term, the section of the article in which it is found, its context, and the frequency of its occurrence within the article or patient record.

Disambiguation

After term identification, each occurrence of a medical term is associated with a set of CUIs, a set of semantic types, and an exclusion context. Terms may be associated with multiple CUIs, which in turn may be associated with multiple semantic types. To reduce this to a single CUI and semantic type for each term, two levels of disambiguation are applied. The first 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); concepts expressed using one term are likely to also be expressed using another equivalent term within the same document. The highest frequency concepts are retained for the second level of disambiguation, which selects the CUI that is associated with the medically important semantic types (those with highest weight). If there are terms that are still ambiguous at this point, the first concept ID and its first associated semantic type are chosen.

Calculation of the Degree of Match between Article and Record

After the disambiguation stage, term information in the article and the patient record has been transformed into a vector of frequencies, a_i and p_i respectively. Our matching metric is based on the cosine between the two vectors, i.e.,

$$M_{base} = \frac{\sum_i A_i \cdot p_i}{\sqrt{\sum_i A_i^2} \cdot \sqrt{\sum_i p_i^2}} \qquad (1)$$

where i iterates over all concepts in the union of the article and patient record, and A_i is an adjusted version of a_i. The adjustment accounts for positional information, namely that terms in the abstract, methods, or results sections are more likely to be important for a match. Via experiments on our development set of articles and patient records, we have derived a set of weights s_j corresponding to each section in an article. Then, the modified frequency A_i used in equation (1) is

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

where a_{ij} is the frequency of term (concept) i in section j within the article.

We further modify equation (1) to account for other important information on terms. Each contribution $A_i \cdot p_i$ is further multiplied by a modifier v_i capturing the degree the associated values (if any) match, a modifier n_i representing any negative context information, and a modifier t_i representing the importance of the semantic type of concept i.

At this stage, our implementation matches only quantitative values for terms (numbers and ranges of numbers). If the two values overlap, we set v_i to 1, otherwise to -1. Terms with no values receive the same weight as terms with explicit and agreeing values. Because our current value matching is primitive, we do not employ the v_i's in the experiments reported in the Results section. Negative context such as explicit negation ("no CHF") reverses the sign of the contribution of term i, so that if the patient record ascertains a condition and an article explicitly excludes it the match would be lower. Negative contexts in both the patient record and the article cancel each other out, resulting in a positive contribution.

Semantic type weights are assigned to approximate the importance of the term. For example, a term with semantic type "Disease or Syndrome" (e.g., "congestive heart failure") is probably more relevant than a term with semantic type "Body Part, Organ, or Organ Component" (e.g., " left ventricle"). We are using predetermined values for t_i, chosen in part via experimentation and in part via consultation with physicians.

With these enhancements, our final matching formula becomes

$$M_{final} = \frac{\sum_i A_i \cdot p_i \cdot v_i \cdot n_i \cdot t_i}{\sqrt{\sum_i A_i^2} \cdot \sqrt{\sum_i p_i^2}} \qquad (2)$$

where i iterates over all concepts in the union of the article and patient record, as before.

Evaluation

We present an information retrieval experiment demonstrating the promise of our system for improving precision and recall of article searches in comparison with searches using regular search-engines. We consider two search strategies for comparison:

Random Term Pair Strategy: The keyword "treatment" plus two medical terms chosen at random from the patient record were used for a MEDLINE-style[1] full-text search on all articles of the journal *Circulation*.[2] This strategy uses the same information that our system has (the patient record) and simulates the approach of a searcher without medical knowledge, e.g., a lay person or a beginning medical student or an automatic process. It will act as our baseline. Note that trying more than two terms simultaneously very often returns no results, hence our choice of term pairs.

Expert Strategy: Several small sets of highly relevant keywords, carefully chosen by a medical expert on the basis of knowledge of the patient's situation were submitted to a MEDLINE search on all articles in our full collection (on titles and abstracts only). This procedure is very similar to the strategy and search environment an experienced doctor might choose in a real-world setting. Further, the results of the initial MEDLINE searches were refined by an expert on medical system evaluation, by expanding some of the query terms and selecting only a subset of the returned results. This strategy, including hand filtering of results, is expected to achieve near perfect precision/specificity, but its sensitivity might be limited by the number of queries the doctor tries.

We created a universe of articles used only for evaluation by merging results returned from the above strategies, which represent opposing points of medical sophistication, and restricting them to those in our 29,748 article corpus. This yields a mixed set of relevant and potentially irrelevant articles. It was necessary to restrict the universe to a number small enough to allow us to ask the medical expert[3] to read all abstracts in a reasonable time frame and make a relevance decision between the paper and the patient record given the "treatment" query. We chose articles for the evaluation set as follows:

- 40 articles per patient were chosen by the Random Term Pair Strategy. Four documents per query were randomly chosen and added to the data set. If a query returned less than four documents, that number was added to the set. We repeated this procedure until we obtained 40 documents per patient. These searches resulted in 77 articles (40 for patient A and 40 for patient B, with three articles appearing by chance in both sets).
- As many articles as returned were chosen by the Expert Strategy (informed search). A set of queries was issued, using keywords gained in an interview with the medical expert (3 keywords for Patient A and 5 for Patient B). Because the terms were much more selective, and because the expert only had a restricted search engine available (articles, titles, and keywords), there are less matching articles. We therefore ran these queries on our entire corpus (all 20 journals). The union of the individual query results and the subsequent expansion and filtering by the evaluation expert returned eight distinct documents per patient.

Our test corpus therefore consists of 93 articles.

The next step of the evaluation required relevance judgments for both patients on the full set of articles.

[1] The search engine is provided by Ovid Technologies.

[2] This journal makes up almost 40% of our corpus and is considered one of the highest quality publications in the field, with a high impact factor of 9.903 as reported by the Institute for Scientific Information.

[3] The fourth author of the present paper.

	Patient A			Patient B		
	P	R	F_1	P	R	F_1
Expert	100.0	44.4	61.5	100.0	50.0	67.0
System, T=.001	19.8	100.0	33.0	17.8	100.0	30.2
System, T=.005	24.5	66.7	35.8	21.7	62.5	32.3
System, T=.010	47.4	50.0	48.6	25.0	25.0	25.0
System, T=.015	100.0	22.2	36.4	50.0	6.2	11.1
System, T=.020	100.0	16.7	28.6	*	*	*
Random Pairs, avg.	14.7	14.6	8.5	11.6	3.6	5.5

Table 1: Precision, recall and F_1-measure in %, at model threshold 3 for various system thresholds (T). Starred entries had no relevant articles retrieved.

This took the medical expert about 4 hours. His judgment was applied *after* the universe of 93 articles had been determined. He judged relevance on a 0 to 5 scale, with 3 being fair and 5 a good match.

In our evaluation we convert judgments on this scale into binary relevance judgments (relevant / non-relevant) using different *model thresholds*. Our matching algorithm also produces graded relevance output (between 0 and 1), so by applying lower or higher *system thresholds* we can relax or tighten what our system would propose as relevant.

Results

We calculated results in two phases: First, we examined the quantitative impact in performance metrics of the various components of our matching algorithm. Space does not permit us listing here detailed scores, but we summarize our observations. We observed little effect from the weighing of terms according to section, which we attribute to our inability to correctly estimate the section weights from our small training data set. Negative context had a small but measurable effect, while we attained the most remarkable improvements (in the order of 30-40% in relative increase of the evaluation metrics) when using the weighing of terms according to semantic type.

The second phase of our evaluation involves comparing the quality of our matching algorithm to the alternative strategies offered by the Random Term Pair and Expert strategies. We measure precision, the percentage of relevant articles among those retrieved (a measure related to specificity); recall (also known as sensitivity), which rates how many of the relevant articles were retrieved; and F_1-measure, which combines precision and recall into a single number in a theoretically sound manner [8]. We performed several runs for different thresholds for our matching algorithm's output and for different ways of converting the expert's ratings of relevance to binary decisions. Partial results are shown in Table 1.

Conclusion

From the results in Table 1 we see that our technique significantly outperforms the baseline of the Random Term Pairs strategy. That baseline still uses the patient record, but clearly less effectively. Thus the improvement offered by our system is attributable to its more nuanced weighing schemas as well as its use of the entire patient record as a query mechanism.

The Expert strategy achieved perfect precision by hand-selecting good medical terms by a medical expert and filtering the results by another evaluation expert with medical domain expertise. Our system, however, not only surpasses traditional uninformed approaches in all measures, but appears to be of use even to the specialist in terms of recall. As it uses information from the entire patient record at once, it is able to recover as much as twice as many relevant articles at a respectable precision level (with approximately one of four suggested articles being truly relevant).

In the future, we are planning to refine further some of the components of our matching (utilizing value matches, for example) and introduce machine learning for the accurate estimation of the relevance of semantic types, once we have more rated articles. Currently, this estimation is done manually, which is one of the limitations of this work. We will also conduct a larger scale evaluation, using a universe of 1000 articles, more than 10 judges, and three patient situations instead of two.

References

1. P. D. Clayton, R. V. Sideli, and S. Sengupta. Open architecture and integrated information at Columbia-Presbyterian Medical Center. *M.D. Computing*, **9**(5):297–303, 1992.
2. C. Friedman, G. Hripcsak, W. DuMouchel, S. B. Johnson, and P. D. Clayton. Natural language processing in an operational clinical information system. *Nat. Lang. Engineering*, **1**(1):83–108, 1995.
3. N. Green, P. G. Ipeirotis, and L. Gravano. SDLIP + STARTS = SDARTS: A protocol and toolkit for metasearching. In *Proc. JCDL (to appear)*, 2001.
4. C. Grover, A. Mikheev, and C. Matheson. LT TTT version 1.0: Text tokenisation software. Technical report, Human Communication Research Centre, University of Edinburgh, 1999.
5. 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.
6. A. Kanter, F. Naeymi-Rad, and I. E. Buchan. Right information, right patient, right time: Intelligent content searching supporting point-of-care applications. In *Proc. Annual Symp. AMIA*, pages 403–407, 2000.
7. K. McKeown, S. Chang, and J. Cimino et al. PERSIVAL, a system for personalized search and summarization over multimedia healthcare information. In *Proc. JCDL (to appear)*, 2001.
8. C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition, 1979.