Leveraging a Common Representation for Personalized Search and Summarization in a Medical Digital Library

Kathleen R. McKeown Noemie Elhadad Vasileios Hatzivassiloglou

Department of Computer Science
Columbia University
New York, NY 10027, USA
E-mail: {kathy, noemie, vh}@cs.columbia.edu

Abstract

Despite the large amount of online medical literature, it can be difficult for clinicians to find relevant information at the point of patient care. In this paper, we present techniques to personalize *the results of search, making use of the online patient record as a sophisticated, pre-existing user model. Our work in* PERSIVAL*, a medical digital library, includes methods for re-ranking the results of search to prioritize those that better match the patient record. It also generates summaries of the re-ranked results which highlight information that is relevant to the patient under the physician's care. We focus on the use of a common representation for the articles returned by search and the patient record which facilitates both the re-ranking and the summarization tasks.This common approach to both tasks has a strong positive effect on the ability to personalize information.*

1. Introduction

The medical field publishes a high volume of research articles every year and many of these are now available online. While increased availability of online literature suggests that it should be easier to access information, in practice online searches often provide users with more information than needed, much of it irrelevant. This may be caused in part by the fact that often search queries contain only a few words [16]; users are notoriously tight-lipped when providing clues about what they are interested in. If information about the end user could be taken into account when searching and presenting results, a system would be able to better filter results to improve relevance. In this paper, we present methods to re-rank search engine results based on user-specific knowledge, highlighting information in which the user is more likely to be interested through automated generation of a personalized summary of the re-ranked search results.

Our research on personalized search and summarization is part of PERSIVAL (PErsonalized Retrieval and Summarization over Images, Video and Language) [11], a medical digital library. For physicians, experienced or in training, PERSIVAL will provide access to literature that is clinically relevant to the patient under their care at the point of patient care. In this scenario, the patient record can provide information about articles that are likely to interest the healthcare specialist. This information includes the medical history, laboratory results, procedures performed and diagnoses, which can be used to pinpoint articles that can provide the physician with the latest results relevant to the patient under care. Similarly, the patient record can also be used to determine which information *inside* the articles is likely to be of interest and to highlight it as part of the summary of search results.

We use a unified approach to re-ranking of search results and personalized summarization, using the same representation and basic tools to determine relevance to a patient. 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. Article relevance is determined by "matching" the article profile against the patient profile, with higher rankings given to articles with a better match. Summary content is determined by matching article sentences against the patient profile, retaining sentences that match well as potential summary content.

Profiling and matching are the basic primitives from which re-ranking and personalized summarization are built. Re-ranking enriches this process using additional features associated with terms that can help determine when an article is a good fit. For example, we might expect that the

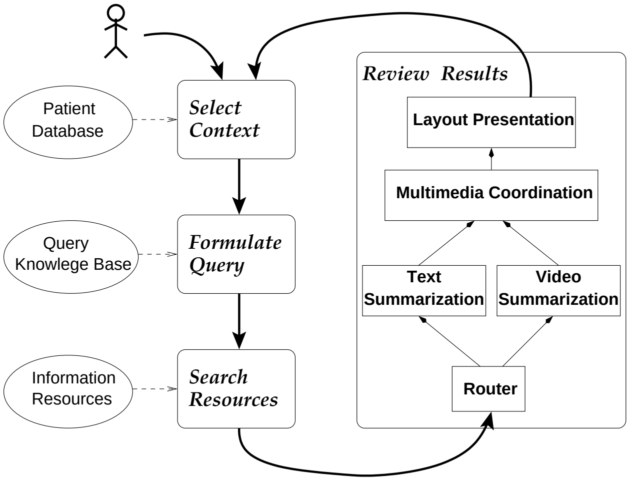


Figure 1. PERSIVAL system architecture with focus on the summarization modules.

semantic category of a term (e.g., *disease* vs. *body part*) would influence how important it is to the match as might the article section in which a term occurs (e.g., "Methods" vs. "Related Work").

Summarization further tailors the information presented to the user by identifying specialized pieces of information within the article, and keeping only those specialized pieces that match the patient's background and current status. User feasibility studies at the early stage of the project [12] indicate that physicians are interested in getting information primarily from technical articles such as clinical studies; further, the information should be tailored to the specific patient at the point of care, and presented in summary form whenever possible. These three desiderata have motivated our approach to personalized summarization. First, an automatic classifier of the relevant articles is needed to filter out articles that were not determined to be of interest (e.g., letters to the editor, case studies); we have built such a classifier for clinical studies, which represent a major portion of the technical articles published in medical journals. Second, the summarizer satisfies the other two user needs: Starting from clinical studies, it selects the sentences that report results using information extraction techniques, and matches them to the patient's profile to keep only the results relevant to that specific patient. However, even including just the results that matched would produce a lengthy and repetitive summary, since we are extracting facts from multiple relevant articles. Instead, we merge matching sentence pieces across articles, identifying repetitions and contradictions and grouping information in a coherent way.

This two-pronged approach of re-ranking results at the article level and then selecting and merging pieces of information across articles to satisfy specific information needs contrasts with current search practice, both in general search engines and in search mechanisms employed in the medical domain. Search engines will retrieve all relevant articles for a given query, and present them as entire documents to the user. In contrast, PERSIVAL utilizes re-ranking to prioritize documents that match the current patient, not just the query, and summarization to only display the most pertinent information.

In the following sections, we first overview the PERSIVAL architecture. We then turn to the themes of this work, personalization and the unified representation and approach to re-ranking and summarization. Following that, we discuss the three major components of the system: article classification, re-ranking, and summarization. We close by showing how the unified approach yields more personalized results than would either approach alone.

2. PERSIVAL: Personalized Access to Medical Literature

PERSIVAL is designed to provide personalized access to a distributed digital library of multimedia medical literature. It is an interdisciplinary project that involves researchers in computer science, electrical engineering, medical informatics and library and information science.

A key feature of PERSIVAL is the user's ability to ask questions and receive related literature within the context of patient information. PERSIVAL links to the large online patient record database available at the New York Presbyterian Hospital, which serves as part of the user model [1, 6]. As shown in Figure 1, interaction with PERSIVAL begins with access to a specific patient record. From the context of the patient record, the user may decide to access the online literature and pose a question. The Query Formulation module helps a user to formulate a good question related to the patient information within the context [13]. The query, along with related patient information, is then sent to the search engine, which allows access to distributed online textual resources [4] as well as a library of digital echocardiograms. The results of the text search are re-ranked by matching the articles returned against the patient record, scoring those articles which discuss results related to the patient's case as more relevant [17]. A textual summarizer [3] and a video summarizer [2] each generate a summary of the relevant results and a multimedia coordination component produces explicit links between the two. The resulting multimedia summary and search results are presented to the user by a sophisticated layout component [10].

User Type: {physician, lay}
Context: {<patient record number>, self}
Access Task: {browse, search, get briefing}

Figure 2. The PERSIVAL user model, its dimensions and their possible values.

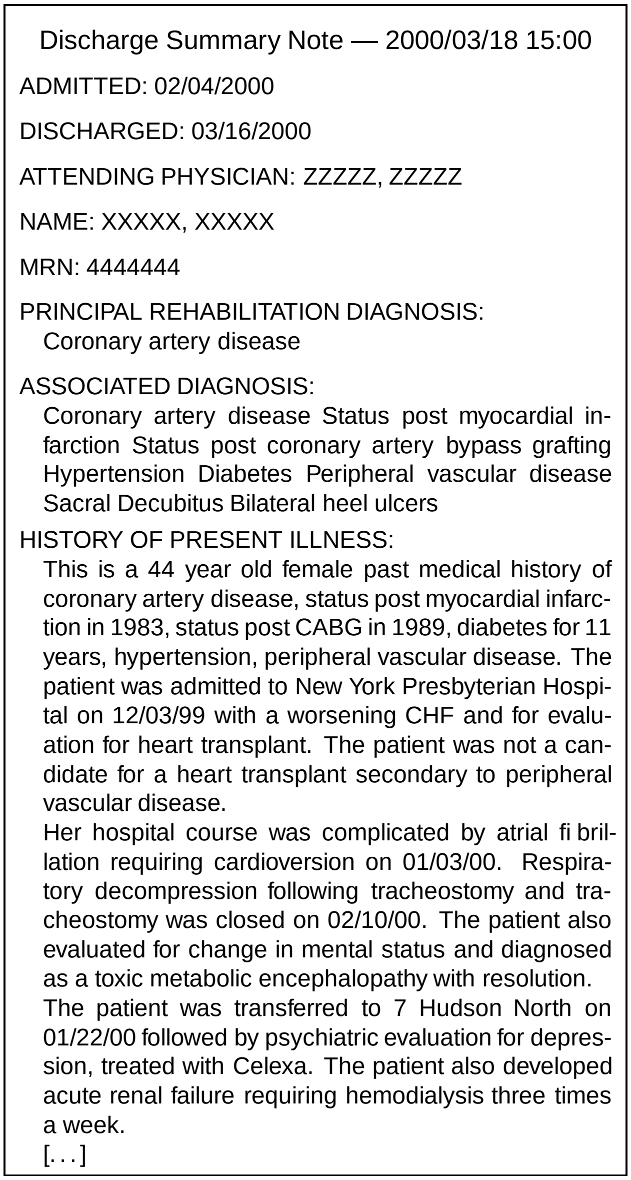
3. Personalization and a Scenario

Digital libraries often serve a wide spectrum of users, varying in level of expertise, interaction goals, and context of question. In the medical setting, users can range from naïve lay consumers, to educated non-specialists, to medical students in training to specialized clinicians.

The user model in PERSIVAL is designed to capture these differences in information need and access strategies between lay people and physicians. Based on our user population analysis done in accordance with the guidelines in [9], we represent three basic dimensions in the user model: (1) the domain expertise of the end user (physician versus layperson); (2) for input, the identity of the patient being treated (a patient of the physician, or the patient end user himself); (3) for output, the user's access task (browsing, searching, or getting a briefing). The attributes and their possible values in our implemented user model are shown in Figure 2. In practice, some dimensions are more salient than others depending on the type of user. When information is needed at the point of patient care, clinicians will prefer just one access task (get briefing), while lay people will tend to look for information for themselves (self).

In this paper, we limit ourselves to physicians and physicians in training (i.e., residents and interns) as end users. As domain experts, physicians are highly knowledgeable about their field of practice; they need access to the latest findings published in the medical literature to keep abreast of new developments in the field. Providing patient-specific information can have enormous benefits, particularly when supporting evidence-based practices [14]; physicians often search for clinical studies with a specific patient they are treating in mind. Thus, to provide relevant information, it is critical for PERSIVAL to take information from the patient record into account.

We use the online patient records at New York Presbyterian Hospital (NYPH) [1] to provide this aspect of the user model. Note that while the physician might be able to provide this information as part of an extended query (e.g., providing background of illness), it would considerably lengthen the time of interaction and if the system can take advantage of the fact that it is already available, there is no need for this additional user input. A patient record for any single patient consists of many individual reports, collected during a visit to the hospital. For some patients, this can be up to several hundred reports. While some of the reports are in tabular format, and thus similar to a database entry, many of the reports are textual (e.g., they may be the result of dictation) and thus require natural language processing in order to be useful for further processing.


Discharge Summary Note — 2000/03/18 15:00

ADMITTED: 02/04/2000

DISCHARGED: 03/16/2000

ATTENDING PHYSICIAN: ZZZZZ, ZZZZZ

NAME: XXXXX, XXXXX

MRN: 4444444

PRINCIPAL REHABILITATION DIAGNOSIS:
Coronary artery disease

ASSOCIATED DIAGNOSIS:
Coronary artery disease Status post myocardial infarction Status post coronary artery bypass grafting Hypertension Diabetes Peripheral vascular disease Sacral Decubitus Bilateral heel ulcers

HISTORY OF PRESENT ILLNESS:
This is a 44 year old female past medical history of coronary artery disease, status post myocardial infarction in 1983, status post CABG in 1989, diabetes for 11 years, hypertension, peripheral vascular disease. The patient was admitted to New York Presbyterian Hospital on 12/03/99 with a worsening CHF and for evaluation for heart transplant. The patient was not a candidate for a heart transplant secondary to peripheral vascular disease.
Her hospital course was complicated by atrial fibrillation requiring cardioversion on 01/03/00. Respiratory decompression following tracheostomy and tracheostomy was closed on 02/10/00. The patient also evaluated for change in mental status and diagnosed as a toxic metabolic encephalopathy with resolution.
The patient was transferred to 7 Hudson North on 01/22/00 followed by psychiatric evaluation for depression, treated with Celexa. The patient also developed acute renal failure requiring hemodialysis three times a week.
[. . .]

Figure 3. Part of the discharge summary for Patient A.

As an example, consider the case of Patient A, an anonymized, true case drawn from the online clinical database. She comes to the hospital because of shortness of breath and chest pain. She already has a patient record online, and the discharge summary from the last visit in-

Table 1. Extract from the profile for Patient A.

Term	UMLS Concept ID	Semantic Type	Positive/Negative	Report
hemodynamics	C0019010	Organ or Tissue Function	+	Cath Lab
conduit	C0441247	Medical Device	+	Cath Lab
aorta	C0003483	Body Part or Organ Component	+	Cath Lab
artery	C0003842	Body Part or Organ Component	+	Cath Lab
peripheral vascular disease	C0085096	Disease or Syndrome	+	Cath Lab
diagnosis	C0011900	Diagnostic Procedure	+	Discharge
ischemic cardiomyopathy	C0349782	Disease or Syndrome	+	Discharge
diabetes	C0011847	Disease or Syndrome	+	Discharge
atrial fibrillation	C0004238	Finding	+	Discharge
cardioversion	C0013778	Therapeutic Procedure	+	Discharge

dicates to the examining physician that she has a history of coronary artery disease, diabetes, hypertension, smoking, and atrial fibrillation. Figure 3 shows an extract from the latest discharge summary; an additional 840 words are present in the full discharge summary. Patient A has a total of 125 reports in her record (110 lab and microbiology reports, 9 cardiology and radiology reports, and 6 admit/discharge summaries). Now, her left ventricular ejection fraction is 35%, which indicates that there is a chance of recurrent atrial fibrillation. Given these pieces of information, the physician wants to know what is the best treatment for recurrent atrial fibrillation. In the remainder of the paper, we will use Patient A and this physician question to show how personalization works.

4. Common Methods and Representations

PERSIVAL's user model is made accessible to all components and thus, personalization for re-ranking and summarization is based on identical information. In addition to the user model, re-ranking and summarization both use the same set of primitives for processing the patient record and the articles under consideration. Personalization is carried out by first producing the article and patient record profiles and then matching either the entire article (for re-ranking), or sentences within the article (for summarization) against the patient record.

A key element of our approach is to base relevance decisions on important medical terms rather than all words, as search engines typically do. Both the patient and article profiles consist of the set of all medical terms found in the documents. To build the profile, we use an efficient finite state grammar to extract terms (e.g., "left ventricular ejection fraction"), along with associated values (e.g., "low" or "35%"), that describe the patient study population. The grammar defines terms as noun phrases which are encoded as finite patterns over adjectives, quantifiers, determiners, and nouns. Conjunction between terms is removed and the two separate terms are generated. Negation is also noted. We filter the non-medical terms by consulting a medical term database, the Unified Medical Language System (UMLS) [7]. UMLS assigns to each string an internal identifier (Concept Unique Identifier, or CUI). 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 remove all terms with semantic types associated with general concepts (e.g., time, persons, and hospital and administrative terms). Acronyms are expanded to the full medical term using a list of 2,011 acronyms in the cardiology domain collected from the Internet. Finally, 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"); or (c) direct comparison operators (e.g., "blood pressure greater than 100 mm Hg").

The resulting profile is thus a list of terms with the associated CUI and semantic type for each term. Table 1 shows a portion of the profile that is constructed for Patient A from a catheterization laboratory report and the discharge summary. The UMLS links terms that refer to the same concept by assigning them the same CUI. For instance, "atrial fibrillation", "auricular fibrillation" and "A-Fib" all share CUI C0004238.

Determining relevance of an article or sentence to the patient record is based on a primitive match function. Matching takes two terms, each possibly with associated values, and matches the CUIs of the terms and their values. This means that two terms match if they are synonymous, whether or not they use exactly the same form. This primitive matching is then extended in various ways by re-ranking and summarization, as discussed below. For example, note that many of the terms shown in Table 1 would not be indicative of a match because they don't refer to disease or treatment (e.g., "diagnosis," or "conduit"). Further extensions deal in part with weighting to place more stress on the more important terms.

As an example, PERSIVAL finds many matching terms given the journal article "Patient Characteristics and Underlying Heart Disease as Predictors of Recurrent Atrial Fibrillation After Internal and External Cardioversion in Patients Treated with Oral Sotalol" from the *American Heart Journal*, which our physician informants indicated is a good article when treating Patient A. Terms such as "atrial fibrillation," "cardioversion," and "coronary artery disease" occur in both patient and article profiles, indicating the overlap in diseases and methods (cardioversion is a method used to treat atrial fibrillation).

We now turn to categorization, re-ranking and summarization, showing how they use the article and patient profiles and embed the basic match primitive in their operation. When possible, we use the scenario presented here to illustrate our approach.

5. Categorization

Our user studies revealed that clinical studies are of more importance to physicians than many other article types. During search, our distributed search component accesses different databases depending on whether the user is a lay person or a physician. In the case of a physician, the search is performed on technical article collections, which include medical research publications. However, there are several possible types of technical medical publications, ranging from the very general (clinical trials and review articles) to more specific (case reports) to miscellaneous publications (such as letters to the editors).

We implemented a categorizer which automatically detects the type of an article. From our local collection of 35,000 journal articles, we selected a subset of 7,000 cardiology articles from PubMed[1] for training and testing of our categorization system. We used 6,000 articles for training and 1,000 for testing. All the articles indexed in PubMed have meta-data tags available, among them the type of publication.[2] We used this field to automatically label each article. In the training data, 59% of the articles were clinical studies. We took advantage of the preprocessing of articles into an XML format where the different sections are identified, along with their titles, to provide the features used for categorization. We use simple features such as the length of the document (number of words), the presence of an Abstract section, the presence of sections with title containing the words Methods or Results, as well as the presence of some key terms such as "trial" or "randomized". The categorization achieves 91.8% precision at 97.8% recall for the "Clinical study" category.

6. Re-ranking of Search Results

The re-ranking component receives as input a patient profile and a set of articles that need to be personalized to that patient. The patient profile consists of a set of attribute-value pairs, which are extracted from the various reports and tables in the patient record. The set of articles is typically the result of a distributed search over large collections of online articles. This search is performed with keywords that a physician or patient specifies. Although the particular choice of keywords is naturally influenced by the patient's current condition and prior medical background, the search procedure itself is not informed by the patient record or any other user model. As is the case with most search engines, the search component in PERSIVAL produces the same set of articles for a given query and collections to search; it is the task of the re-ranking module to take into account known information about the patient to produce a modified list of search results that varies from patient to patient for the same query. We have experimented with modifications that reorder the articles originally retrieved from the search so that articles that are more likely to apply to the patient under consideration are ranked near the top.

To achieve this reorganization of the search results, we take advantage of our common representation of medical information and tools for operating on this information (detection of terms, mapping of terms to concepts, semantic categorization of concepts, and primitive matching between concepts). The re-ranking module views the articles and the patient record as sets of attribute-value pairs, where the attributes are the medical terms after they have been disambiguated and mapped to concept identifiers (CUIs), as described in section 4. In addition to the semantic type that is provided by our term disambiguation module using information from the UMLS, the re-ranking module uses several additional features, also extracted during the processing of the articles and patient record when our common representation is constructed:

- *Negation* of terms, which can be explicitly signaled by words such as "no", "none", "without", etc. as in "patients without myocardial infarction were sampled ..." and "no atrial fibrillation was observed", or implicitly specified when the term occurs in an exclusion context, such as "we did not include patients who ...".

[1] PubMed is a search engine for medical publications provided by the National Library of Medicine. It is available at `http://www.ncbi.nlm.nih.gov/PubMed/`.

[2] Note that PERSIVAL uses a distributed search engine that is not limited to PubMed alone. Thus, the hand labeled categories in PubMed are not sufficient for direct use in PERSIVAL. However, they are helpful to build training data.

We have implemented a set of pattern-based rules that recognize the most common constructs that introduce negative context, and each recognized term is assigned either a positive or negative label. The intent of this feature is to prevent spurious matches where a term occurs in both the article and the patient record, but in a negative context in one and positive context in the other. This is exactly the situation that a typical search engine cannot recognize—it would return an article discussing women who *did not* have peripheral vascular disease or a prior heart attack for Patient A. Negation does not occur frequently (less than 0.5% of the cases we detect are negated), but changes meaning when it does.

- *Section information* provides clues as to the relative importance of terms in articles. Certain sections (e.g., Introduction) are more likely to provide general background information that may not apply to the patients in a clinical study, while sections such as Methods more often specify the characteristics of the population under study. We take advantage of the relatively rigid structure of journal papers in medicine to segment the articles into sections, and give priority to primitive matches involving terms in more privileged sections of each article.

- *Values* for recognized terms (such as "blood pressure over 100mm Hg") help to further assess the compatibility of terms that appear both in the patient record and an article. Certain demographic and medical attributes such as age, heart rate, or ejection fraction will appear in many articles in the cardiology domain; however, the fact that both the patient record and an article mention "age" should not influence our assessment of the match between them. It is the compatibility between the values associated with these attributes that determines whether they should contribute to the match in a positive or negative direction. We extract values using our finite-state grammar to recognize modification relationships, as illustrated in Section 4. Presently, our strategy for determining the compatibility of values is rather naïve: values are compatible if identical, partially compatible if they are numeric depending on how close they are, and incompatible otherwise. However, even this simple tactic offers a small improvement over not using the value information at all—solving the hard and interesting issues in comparing values (such as comparing a numeric value to a qualitative expression such as "high") remains a high-priority direction of our future research.

- *Inverse document frequency* (IDF) for each term (the negative of the logarithm of the ratio of the number of articles in our entire collection that contain this term versus the total number of articles (35,000) in that collection) helps locate the rarer terms, which are presumably more informative when they do occur. Primitive matches are weighted so that those involving terms of high IDF influence the overall match more.

Once terms have been extracted and annotated with the above features, they are collected into one vector representing the patient record and a similar vector for each article in the set supplied by the search module. The re-ranking module uses this information to calculate a numeric value for the compatibility of any two such vectors (in practice, we are only interested in the compatibility of each article with the fixed patient record). We base this compatibility value on a modified cosine measure, which takes into account frequency information (to weigh more the concepts that appear more often in either the patient record or the article) as well as the modifying factors expressed by our features. 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 [15]:

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

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. This basic matching formula utilizes as a basic building block the primitive matching between single instances of terms, which links terms that are expressed differently in the text but all correspond to the same concept (see Section 4).

Given the formula (1), we can modify this basic matching function to take account of the factors modifying a term's importance. First, we account for the influence of section information by replacing term frequency over the entire article by the sum of term frequencies for each section and weighting each such frequency by a weight representing the importance of that section. This results in the normalized frequency of a term according to section information, accomplished by replacing a_i (the term frequency) in the formula above 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 $\sum_j a_{ij} = a_i$). Our evaluation of re-ranking alone shows that section weights have a small positive influence on overall results [5].

We further modify the contribution of each term by weights representing the following factors:

- a weight capturing the relative importance of term's i semantic type, represented as t_i in the formula.

- a weight capturing negations when present. For terms occurring once in the patient record and article, this 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. This weight is represented by n_i in the formula. For example, if the same term is seen twice in the patient record and three times in the article, one of the latter in a negative context, this yields six pairs of terms. Four of these pairs contain two positive contexts and thus each provides a score of $+1$. In two cases, we have a mismatch, with one positive and one negative context. This gives $(4+(-2))/6$ to yield the final weight of 1/3.

- a weight which captures the similarity between observed values for term i in the article and the patient record. As in the case of n_i, for terms occurring multiple times in the article or the patient record we assign separately a similarity to the values associated with the terms in each pair, and subsequently average these similarities to obtain v_i. In our current implementation, a pair of values is deemed either fully compatible with a similarity of 1 (if they are identical, or if one or both terms have no values assigned to them), incompatible with a similarity of -1 (if both values are present, they are not identical, and at least one of them is non-numeric), or partially compatible (if both of them are present and numeric). In the latter case, the similarity for the pair is based on how much apart the two numbers V_1 and V_2 are, namely,

  $$\frac{\min(V_1, V_2)}{\max(V_1, V_2)}$$

  This is represented by v_i in the final formula.

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

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

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).

The weights s_i and t_i in the above formulas represent the relative significance of different sections and different semantic types. Currently, we have empirically determined "good" values for these weights through experimentation on small sets of articles and in consultation with medical experts. We plan to eventually use machine learning techniques to determine optimal values for these weights.

An earlier version of our re-ranking component[3] was evaluated using a set of 93 articles and two patient records [17]. The articles were selected not as a response to a particular query (which would bias the evaluation towards that query type) but by combining articles known to be relevant to each of the three patients (as determined by a medical expert) with articles that randomly matched some of the terms in the patient record. A specialist in cardiology assigned relevance scores for each of these 93 articles and each patient, and we compared the scores assigned to each article by the system to the expert's relevance score. We used different thresholds to convert the relevance scores produced by both the system and the expert to binary judgments ("relevant article for this patient or not"). Our evaluation (see [17] for a full description of experimental setting and results) showed that the re-ranking strategy significantly outperformed our baseline strategy that determined relevant articles by randomly selecting terms from the patient record and submitting them to a standard search engine. Compared to the expert, the re-ranking module achieved as expected lower precision (about 50%) but located many relevant articles that the expert himself did not find using standard queries on PubMed.

More recently, we have collected the data from a large-scale evaluation using the latest version of the re-ranking module. For that evaluation, we expanded the number of patients to three (from two), the number of articles to 939 (from 93), the number of article types to three (prognosis, treatment, and diagnosis from treatment only), and the number of physician evaluators to nine (from one). These results also show that our model outperforms the baseline, and several competitive models for searching medical collections. The full results (available in [5]) will be reported in a future paper.

An example of the effect that re-ranking has can be observed by looking at the results for a sample query that could be asked for our patient A. We provide such a query in Section 8, where we show how re-ranking transforms the results of the query, and how summarization improves

[3] Not including the IDF feature and the value matching described above.

Parameter(s):
LVEF [C0428772],
calcium channel blockers [C0006684],
hypertension [C0020538],
diabetes mellitus [C0011849],
cardiopulmonary bypass time [C0007202]
Relation: not predict
Dependence: independent
Finding: atrial fibrillation [C0004238]

Figure 4. Template Example.

when presented with the re-ranked results compared to the originally retrieved articles.

7. Generating Tailored Summaries of Search Results

In the requirements gathering phase of the project, we observed that physicians do not read a study from beginning to end to determine if an article is relevant. Rather, they quickly glance through the *Methods* section describing the patient population, and then focus on the *Results* section. If a clinical study is found to be relevant to the specific patient, then the physician will read the article in more detail. TAS (Technical Article Summarizer) aims to facilitate this process by summarizing the results that are relevant to the patient from the input articles. It also provides links to the original articles, so that the physician can at any time read the whole clinical study.

Given the articles returned by the re-ranking component, we know whether, on the whole, an article is relevant to the patient; this is dependent on the matching weight it was assigned. However, even if a high-ranking article pertains to the input patient, not all the results reported in the article are relevant to the patient. TAS is responsible for finding pieces within the input articles that match with the patient and for including them in the summary in a coherent way.

TAS takes as input the top k clinical studies returned by re-ranking, along with the user model and the query passed to the Search component. Information is included in the summary only if it pertains to the patient represented in the user model. In addition, TAS handles repetitions or contradictions across articles by dynamically merging and ordering all the results from the different input articles and generating a coherent, fluent summary.

TAS follows a pipeline architecture.[4] First, results are extracted from the input articles. We analyzed a corpus of clinical studies to formally determine a definition for what constitutes a result. So far, we have focused on result sentences reported in the articles that relate disease, patient characteristics, or therapies with outcomes. A result is formally defined as a template of the form {<*parameter(s)*>, <*relation*>, <*dependence*>, <*finding*>}, where *finding* is the outcome, *parameter* is typically a condition, or a body part, and *relation* is the type of relation that holds between the two. For instance, the fact that having hypertension and in addition having the habit of smoking increase the risk for heart failure, can be represented as {*(hypertension, smoking), risk, dependent, heart failure*}. This result is encoded as statistically dependent, because the combination of the two *parameters* (hypertension and smoking) represent a risk for the *finding* (heart failure). Based on our corpus, we identified six types of *relations*: risk, association (or statistical correlation), prediction, and their corresponding negations. To extract such templates, we used traditional information extraction technology: sentences are parsed using a shallow syntactic parser, and they are checked against a set of patterns. Since parameters and findings are typically medical terms, we take advantage of the preprocessing of articles which identifies medical terms and tags them with their corresponding UMLS CUI. Given the example sentence "*Atrial fibrillation was not predicted by left ventricular ejection fraction, the use of calcium channel blockers, history of hypertension, diabetes mellitus, or cardiopulmonary bypass time.*", and the pattern "<*finding*> *was not predicted by* <*parameters*>", the template shown in Figure 4 is extracted. The result is statistically independent, which means that each parameter, by itself, does not predict atrial fibrillation (as opposed to the combination of them).

Parameter(s):
LVEF [C0428772],
hypertension [C0020538],
diabetes mellitus [C0011849],
Relation: not predict
Dependence: independent
Finding: atrial fibrillation [C0004238]

Figure 5. Matched Template Example.

Templates constitute a good representation of the information in the input articles. They augment the raw text with semantic information while selecting only the concepts relevant to the summarization task (such as parameters, relations, and findings). Based on the semantic information and the primitive matching operation described in Section 4, we are able to implement complex personalization strategies.

We established two strategies for deciding whether a template matches with a patient record. First, matching should not be performed on the *finding* field of the template representing the article sentence. For instance, in the template {*heart attack, predict, death*}, there is obviously no point in trying to match "*death*" with the patient profile. Only *parameters* are used to determine relevance. This is

[4] The full architecture of TAS is described in [3].

consistent with our definition of a result: parameters can be considered as the current condition of the patient, while the findings represent current possible outcomes. Second, among the parameters, different matching policies should be applied according to the degree of dependence of the parameters: for each parameter in the template, we check whether it matches the patient record. If the result reports independence on the parameters, we perform a logical *or* of all matching parameters. The matching parameters are kept as input for the next component in the system, while the non-matching parameters are discarded. In contrast, if the parameters are dependent, i.e., their combination relates to the finding, we perform a logical *and* of the matched parameters. If one parameter does not match, the whole template is discarded; if they all match, the whole template is passed as it is to the next component.[5] In our example, the template contains independent results, hence we apply a logical *or*. After matching each parameter with the patient profile using our primitive match operation, we obtain the matched template shown in Figure 5.

This process of matching is made easier by the representation of article and patient records as profiles of medical terms. We can use the same basic matching formula defined above, but we restrict it to parameters of a template only, matching against the full patient record. Merging, described below, also relies on basic matching, but in this case for matching all terms within a template against terms from another template, to see if the two templates represent repetitions in the text.

Matching at the template level is performed for each extracted result, for each input article. The next task is to assemble these independent pieces of information into a coherent set. As a first sub-step we split templates that can be split (that is, the "independent template", such as our example template) without changing its meaning. The template is turned into three separate templates: {*LVEF [C0428772], not predict, independent, atrial fibrillation [C0004238]*}, {*hypertension [C0020538], not predict, independent, atrial fibrillation [C0004238]*}, and {*diabetes mellitus [C0011849], not predict, independent, atrial fibrillation [C0004238]*}. This sub-step seems counterproductive with our goal of assembling templates, but in fact it helps us by simplifying the data whenever possible. The templates are then clustered in a hierarchical fashion. We refer to this step as *Merging*. The similarity function between two templates is computed as a combination of the primitive match between the parameters and the match between the findings, and a manually assigned weight to the type of relation. The clustering achieves two purposes: it identifies strictly identical templates (that is, repetitions across or inside articles) and it dynamically groups together templates that are semantically related to each other.

Clustering of templates is equivalent to dynamic document paragraph planning, where each cluster represents a paragraph. In the general content planning phase of the summarizer, the last task left is to decide in which order to present the paragraphs. This ordering is also done in a dynamic fashion. Each cluster gets an ordering weight based on several features: the number of templates it contains, the number of repetitions, the number of contradictions, and the number of different input articles that contributed to the cluster. The rationale behind this is based on user studies we conducted in the initial phase of the TAS design: as a general policy, physicians want to see the important pieces of information first. For instance, a paragraph which reports on a contradiction between two results is considered important and therefore its corresponding cluster should have a higher weight. Another feature is whether the cluster contains any template related to the input query. We use again the primitive match function to decide this.

The content planning phase of the summarizer comprises the above three steps: (1) personalized extraction of results, (2) merging, and (3) ordering. The two first steps make heavy use of the semantics associated with medical terms (CUIs) and the primitive for matching two given terms, while ordering uses them to compute the ordering weight. The second phase of the summarizer is the Content Realization. In our current implementation, we generate English text by combining extracted phrases with canned pre-written slotted sentences [8].

An example summary is given in Figure 9. It provides examples of the three main contributions of TAS:

- Personalization — The summary contains only the results relevant to the patient or the question asked. For instance, the second and third sentences are directly tailored to the input patient: she has coronary artery disease, as well as diabetes, hypertension, a low ejection fraction, and a history of smoking. The third and fourth paragraphs report results on amiodarone and sotalol which pertain to what the user asked.

- Merging and cohesion — The summary does not contain repetitions since identical results are merged. For instance, the fact "*a left atrial with diameter $>$ 60 mm predicts atrial fibrillation* " appears in two input articles (articles 6 and 7), but it is mentioned only once in the summary. Merging also allows semantically related results to be presented in a cohesive manner, as in the first sentence of the summary: results such as "*patient age is associated with atrial fibrillation*" and "*hospital stay is associated with atrial fibrillation*" are aggregated into the first sentence of the summary, even though they come from different input articles.

- Ordering and coherence — The dynamic ordering algorithm allows the summary to present the most im-

[5] Our ongoing work is investigating cases of partial matches.

1. “*Maintenance of sinus rhythm with oral d,l-sotalol therapy in patients with symptomatic atrial fibrillation and/or atrial flutter*”. The American Journal of Cardiology.
2. “*Oral amiodarone reduces incidence of postoperative atrial fibrillation*”. The Annals of Thoracic Surgery.
3. “***Efficacy and Safety of Sotalol in Patients with Refractory Atrial Fibrillation or Flutter***”. American Heart Journal.
4. “*Low-Dose Amiodarone Versus Sotalol for Suppression of Recurrent Symptomatic Atrial Fibrillation*”. The American Journal of Cardiology.
5. “***Efficacy of amiodarone for the termination of persistent atrial fibrillation***”. The American Journal of Cardiology.
6. “*Intraoperative amiodarone as prophylaxis against atrial fibrillation after coronary operations*”. The Annals of Thoracic Surgery.
7. “*Efficacy, Safety, and Determinants of Conversion of Atrial Fibrillation and Flutter With Oral Amiodarone*”. The American Journal of Cardiology.
8. “*Amiodarone versus propafenone for conversion of chronic atrial fibrillation: results of a randomized, controlled study*”. Journal of the American College of Cardiology.
9. “*Intravenous amiodarone for the prevention of atrial fibrillation after open heart surgery*”. Journal of the American College of Cardiology.
10. “*Intravenous amiodarone for prevention of atrial fibrillation after coronary artery bypass grafting*”. The Annals of Thoracic Surgery.

Figure 6. The first ten clinical studies returned from the search before re-ranking. Relevant articles are in bold.

1. “***Prophylactic Oral Amiodarone Compared With Placebo for Prevention of Atrial Fibrillation After Coronary Artery Bypass Surgery***”. American Heart Journal.
2. “*Intravenous amiodarone for prevention of atrial fibrillation after coronary artery bypass grafting*”. The Annals of Thoracic Surgery.
3. “*Intravenous sotalol decreases transthoracic cardioversion energy requirement for chronic atrial fibrillation in humans*”. Journal of the American College of Cardiology.
4. “*Intraoperative amiodarone as prophylaxis against atrial fibrillation after coronary operations*”. The Annals of Thoracic Surgery
5. “***Oral d,l sotalol reduces the incidence of postoperative atrial fibrillation in coronary artery bypass surgery patients***”. Journal of the American College of Cardiology.
6. “***Patient Characteristics and Underlying Heart Disease as Predictors of Recurrent Atrial Fibrillation After Internal and External Cardioversion in Patients Treated with Oral Sotalol***”. American Heart Journal.
7. “***Spontaneous Conversion and Maintenance of Sinus Rhythm by Amiodarone in Patients With Heart Failure and Atrial Fibrillation***”. Circulation.
8. “***Efficacy and safety of sotalol versus quinidine for the maintenance of sinus rhythm after conversion of atrial fibrillation***”. The American Journal of Cardiology.
9. “***Efficacy of amiodarone for the termination of persistent atrial fibrillation***”. The American Journal of Cardiology.
10. “***Prospective Comparison of Flecainide Versus Sotalol for Immediate Cardioversion of Atrial Fibrillation***”. American Journal of Cardiology.

Figure 7. The first ten articles after re-ranking the original search results. Relevant articles are in bold.

portant results first. The first paragraph of the summary is repeated in five articles, and therefore, is considered highly important to report to the user, whereas the fact that “*age predict sinus rhythm maintenance*” is reported in only one article, so it can be included at the end of the summary.

8. The Combined Effect of Personalization

In this section we present a full example and show how personalization can be relevant both at the Re-ranking and Summarization levels.

In our scenario with Patient A, we bypassed the stages of Context Selection and the Query Formulation phases of PERSIVAL, assuming the physician was looking at a patient record and wanted to ask a question about atrial fibrillation and possible treatments. Bypassing these stages allowed us to test just the re-ranking and summarization components alone. Since the physician wants to know about treatment of atrial fibrillation, we performed a search on a collection of technical documents with the following boolean query: get the documents whose titles contain `atrial fibrillation AND (sotalol OR amiodarone)`. The search returned 34 articles.[6] Following the PERSIVAL architecture,

[6] The number of hits is small considering that the collection contains more than 35,000 documents. However, this makes sense given that the

Left atrial diameter and arrhythmia duration predict conversion [5,7]. Left atrial size and atrial fibrillation duration are associated with conversion to sinus rhythm [7,8]. However, sex, gender, age, left ventricular ejection fraction, and heart rate were not found to be associated with conversion [5,8].
In a multivariate analysis, age and ejection fraction predict sotalol efficacy [3,4].
Amiodarone and left atrial size and are associated with conversion rate [5,7].
Atrial fibrillation is associated with hospital stay and increased cost [9,10].

Figure 8. Summary I, for the top 10 search results without re-ranking. The numbers in parenthesis refer to articles from Figure 6.

the Categorization module was invoked first, which filtered out documents that are not clinical studies. In this scenario, this resulted in 27 clinical studies. Figure 6 shows the first ten clinical studies retrieved by the search engine, while Figure 7 shows the first ten clinical studies after re-ranking was performed on the 27 articles.

Looking at the two sets of articles, we observe that the top 10 re-ranked articles constitute a better match with the patient record than the top 10 search articles. Manual examination reveals that only two of the top 10 search results (Figure 6) are relevant to patient A (articles 3 and 5). In contrast, the top 10 re-ranked results (Figure 7) contain seven fully relevant articles for Patient A. By providing personalization at the article level, the physician can get to the relevant articles faster by looking at the re-ranked results, than by looking at the search results alone.

In order to compare the combined effect of personalization at both levels, we produced two summaries. In our input to summarization, we used the same patient record and question from the physician (as described above), but we varied the set of articles being summarized. The first one (summary I, shown in Figure 8) was generated using the first ten clinical studies returned by the search engine, without any re-ranking involved (that is, the articles in Figure 6). The second summary (summary II, shown in Figure 9) was generated using the first ten clinical studies returned by the re-ranking component (that is, the articles in Figure 7).

Summarization answers the physicians' needs better than a list of articles. It is easier and quicker for the physician to read Summary I or II than to access all the articles in Figure 6 or Figure 7 and read them to determine which parts are relevant to the patient. In other words, summarization provides a more fine-grained tailoring of the information, complementary to that offered by the re-ranking.

Atrial fibrillation is associated with patient age, hospital stay, increased cost, and mortality rate [1,2,6,7,9].
Multivariate analysis identified coronary artery disease to predict atrial fibrillation [6,7]. Left ventricular ejection fraction, hypertension, diabetes mellitus, smoking were not found to predict atrial fibrillation [1].
Left atrial diameter $<$ 4.0 cm is a predictor for conversion [8,9]. Left atrial size $>$ 60mm predicts atrial fibrillation [6,7].
Amiodarone and conversion to sinus rhythm are associated [7]. Sex, age, and baseline heart rate are not associated with conversion [9]. Heart failure does not predict conversion to sinus rhythm [7].
Sotalol was associated with decreasing the incidence of atrial fibrillation, and tolerated recurrences [5,8]. There were no differences of mortality between sotalol and placebo [5].
In a univariate analysis, coronary artery disease and age predict recurrence [6,7].
Age and atrial fibrillation predict sinus rhythm maintenance [8].
In a multivariate analysis, there were no differences of relapsing into atrial fibrillation between the modes of treatment [6,7].

Figure 9. Summary II, for the top 10 re-ranked search results. The numbers in parenthesis refer to articles from Figure 7.

Finally, combining re-ranking with summarization strongly boosts performance: re-ranking provides a "better" input to summarization than search alone. Summary II was generated using globally relevant articles, and therefore, more relevant results were selected to be presented to the user: In Summary I, while the summarizer extracted 40 findings from the ten input articles, only 26 were considered matching with the input patient and ended up in the generated summary (yielding a matching rate of 65%). In contrast, for the re-ranked set of articles, the summarizer extracted 39 findings, and considered 35 of them to be relevant to the patient (matching rate of 89%). This trend was verified when manually going through the intermediate results of the summarizer. In addition, the summarizer was able to pick up many repetitions in articles from the re-ranked results, but not from the unmodified search results. The summarizer detected only 4 pairs of repetitive results when extracting information from the former set of articles, while it detected 8 pairs of repetitive results in the latter set. Thus, more of the information in Summary II is based on multiple sources, thus increasing its reliability. Presence of repetitive results across the input articles confirms the

search looks only at the titles of the documents, and the query terms are drug names, and therefore, fairly specific.

validity of the article selection process as well.

9. Conclusions and Future Work

Our research demonstrates how information about the patient, available in the online patient record, can be used to provide a personalized response to a physician's search query. Re-ranking and summarization leverage a common representation of articles and patient record to make their tasks easier. Through construction of article and patient profiles consisting of extracted medical terms, we enable the use of a relatively straightforward matching procedure based on comparison of term CUIs to determine relevance. Information from our primitive matching function is used as a common building block and combined in a variety of ways to improve determination of document relevance for re-ranking or sentence relevance for summarization. Taken together, re-ranking and summarization provide an increase in personalization that would not be possible with either one alone.

There are many directions that we are currently exploring. Feedback from our evaluation with re-ranking indicates that we could improve relevance if we could find better measures to weight importance of terms. Physicians in our group indicate that the more specific terms of any given type (e.g., diseases such as *atrial fibrillation*) are better indicators of a match. In looking at the profiles that were built for an article and a patient record, it is clear that the specific terms are not distinguished from the more general (e.g., age, heart rate). We are exploring methods to exploit depth in the UMLS hierarchy to help us make this determination.

For both re-ranking and summarization, better handling of value matching would also increase personalization. Currently, we only handle matches between numeric values—yet, often a numeric value (e.g., "ejection fraction of 35%") can be matched to a qualitative description (e.g., "low ejection fraction") if we have medical knowledge about what "low" means for "ejection fraction" in the current clinical context. We will investigate text mining techniques for automatically learning ranges of values for common medical attributes, which will allow us to map qualitative descriptions to a part of the range of possible values.

Acknowledgments

This paper is based upon work supported by the National Science Foundation under Digital Library Initiative Phase II Grant No. IIS-98-17434. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

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

[2] S. Ebadollahi, S.-F. Chang, H. Wu, and S. Takoma. Indexing and summarization of echocardiogram videos. In *American College of Cardiology*, 2001.

[3] N. Elhadad and K. McKeown. Towards generating patient specific summaries of medical articles. In *Proc. of NAACL Workshop on Automatic Summarization*, 2001.

[4] N. Green, P. Ipeirotis, and L. Gravano. SDLIP + STARTS = SDARTS: A protocol and toolkit for metasearching. In *Proc. of JCDL*, 2001.

[5] V. Hatzivassiloglou, S. Teufel, K. McKeown, D. Jordan, and S. Sigelman. Personalized search of the medical literature: An evaluation. Technical Report CUCS-003-03, Columbia University, 2003.

[6] G. Hripcsak, J. Cimino, and S. Sengupta. WebCIS: Large scale deployment of a web-based clinical information system. In *Proc. of the AMIA Symposium*, 1999.

[7] B. Humphreys, D. Lindberg, H. Schoolman, and G. Barnett. The Unified Medical Language System: an informatics research collaboration. *JAMIA*, **5**:1–11, 1998.

[8] P. Jacobs. PHRED: A generator for natural language interfaces. *Computational Linguistics*, 11(4):219–242, 1985.

[9] R. Kass and T. Finin. Modeling the user in natural language systems. *Computation Linguistics*, 14(3):5–22, 1988.

[10] S. Lok and S. Feiner. The AIL automated interface layout system. In *Proc. of Intelligent User Interfaces*, 2002.

[11] K. McKeown, S.-F. Chang, J. Cimino, S. Feiner, C. Friedman, L. Gravano, V. Hatzivassiloglou, S. Johnson, D. Jordan, J. Klavans, A. Kushniruk, V. Patel, and S. Teufel. PERSIVAL, a system for personalized search and summarization over multimedia healtcare information. In *Proc. of the Joint Conf. on Digital Libraries*, 2001.

[12] K. McKeown, D. Jordan, and V. Hatzivassiloglou. Generating patient-specific summaries of online literature. In *Proc. of Intelligent Text Summarization, AAAI Spring Symposium*, 1998.

[13] E. Mendonca, J. Cimino, S. Johnson, and Y. Seol. Accessing heterogeneous sources of evidence to answer clinical questions. *Journal of Biomedical Informatics*, 34, 2001.

[14] D. L. Sackett, R. B. Haynes, G. H. Guyatt, and P. Tugwell. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Little, Brown and Company, Boston and Toronto, second edition, 1991.

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

[16] A. Spink, S. Milchak, M. Sollenberger, and A. Hurson. Elicitation queries to the Excite web search engine. In *Proc. of CIKM*, 2000.

[17] S. Teufel, V. Hatzivassiloglou, K. McKeown, K. Dunn, D. Jordan, S. Sigelman, and A. Kushniruk. Personalized medical article selection using patient record information. In *Proc. of AMIA*, 2001.