PERSIVAL, a System for Personalized Search and Summarization over Multimedia Healthcare Information

Kathleen R. McKeown[*], Shih-Fu Chang[†], James Cimino[‡], Steven K. Feiner[*], Carol Friedman[‡], Luis Gravano[*], Vasileios Hatzivassiloglou[*], Steven Johnson[‡], Desmond A. Jordan[**], Judith L. Klavans[††], André Kushniruk[‡‡], Vimla Patel[‡], and Simone Teufel[*]

[*]Department of Computer Science, Columbia University, New York, NY 10027, USA
[†]Department of Electrical Engineering, Columbia University, New York, NY 10027, USA
[‡]Department of Medical Informatics, Columbia University, New York, NY 10032, USA
[**]Department of Anesthesiology and Department of Medical Informatics, Columbia University, New York, NY 10032, USA
[††]Center for Research on Information Access, Columbia University, NY 10027, USA
[‡‡]Department of Mathematics and Statistics, York University, Toronto, Ontario, CANADA

Contact email: kathy@cs.columbia.edu

ABSTRACT

In healthcare settings, patients need access to online information that can help them understand their medical situation. Physicians need information that is clinically relevant to an individual patient. In this paper, we present our progress on developing a system, PERSIVAL, that is designed to provide personalized access to a distributed patient care digital library. Using the secure, online patient records at New York Presbyterian Hospital as a user model, PERSIVAL's components tailor search, presentation and summarization of online multimedia information to both patients and healthcare providers.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing; H.5.2 [**HCI**]: User Interfaces; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval —*Information filtering, Query Formulation*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-base services*; H.3.5 [**Information Storage and Retrieval**]: Digital Libraries—*Dissemination, Systems issues, User issues*

Keywords

Medical digital library, personalization, search, query interface, multimedia, summarization, natural language

1. INTRODUCTION

In healthcare settings, both consumers and their providers need quick and easy access to a wide range of online resources. Patients and their family members need information that can educate them about their personal situations, while physicians need information that is clinically relevant to an individual patient. But unless online search and presentation of results is personalized to the individual patient, end users can be overloaded with far more information than is useful. Providing patient-specific information can have enormous benefits. When the latest medical information is provided at the point of patient care, it can help practicing clinicians and physicians in training to avoid missed diagnoses, choose only effective interventions and diagnostic tests, and minimize impending complications. The latest medical information, when expressed in understandable terms, can empower patients to take charge of their health, take appropriate preventive measures and make more informed choices regarding their treatment.

In this paper, we report on the ongoing development of PERSIVAL (PErsonalized Retrieval and Summarization of Image, Video And Language), a system designed to provide personalized access to a distributed digital library of medical literature and consumer health information. Our goal for PERSIVAL is to tailor search, presentation, and summarization of online multimedia information to the end user, whether patient or healthcare provider. PERSIVAL utilizes the online patient records at New York Presbyterian Hospital [7] as a sophisticated, pre-existing user model that can aid in predicting user interests. For the healthcare provider, our approach facilitates finding literature relevant to the specific patient under her care; for the healthcare consumer, our approach facilitates finding and understanding information relevant to his medical situation.

In the remainder of this paper, we first present the system's architecture and then discuss the components that we are developing. These include a user query component, which allows inference of meaningful questions given the

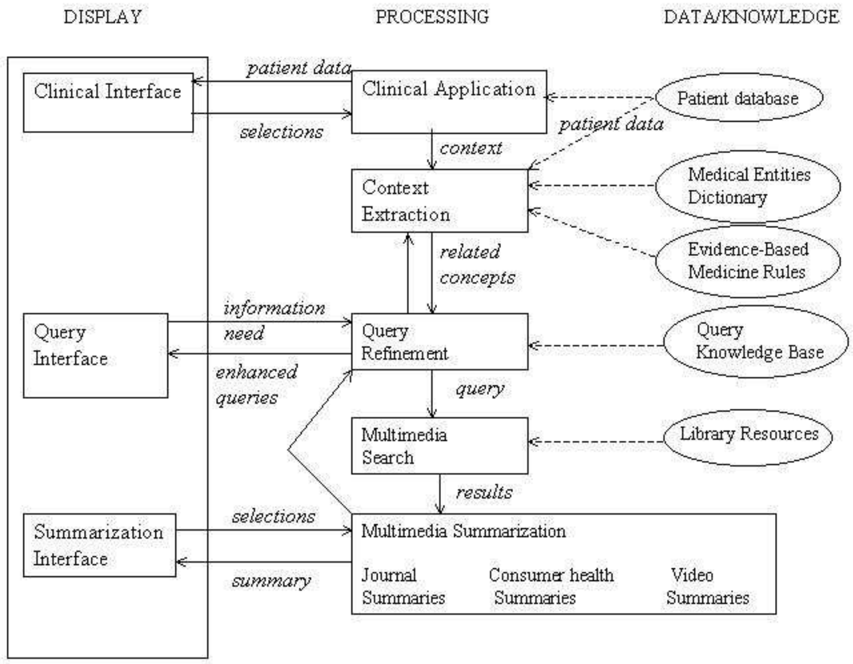

Figure 1: PERSIVAL's system architecture

clinical context, providing relevant information from the context for search; search over heterogeneous, distributed sources, re-ranking results by matching against the patient record; presentation of results to highlight information that is relevant to the patient, including summarization of textual and video resources along with definitions of unfamiliar terms; and interaction with the results through a highly interactive, multimodal, web-based thin-client architecture. Finally, we have begun preliminary experiments with end users that can provide specific information on how to personalize search and summarization.

2. SYSTEM ARCHITECTURE

The online clinical information infrastructure at New York Presbyterian Hospital provides many applications for viewing patient data. One of these, WebCIS [16], allows physicians to view patient records and another, PatCIS [3], allows patients to view information from their record. Patients may have questions about the meaning of different concepts (e.g., "What is unstable angina?") while physicians may have questions about possible treatments and diagnoses (e.g., "What are the risk factors for unstable angina?"). Thus, one way for users to access PERSIVAL is from within one of the above clinical applications, which trigger PERSIVAL to extract content from the patient record. Patients and physicians may also access PERSIVAL when questions arise at other times. For example, a physician, resident or medical student may have a question about a patient during the course of normal tasks (e.g., during rounds), in which case she must first provide the patient's identifying information and go through a secure clinical interface before accessing PERSIVAL, thus again giving the system the opportunity to associate a question with patient data.

The system architecture for PERSIVAL is shown in Figure 1. On the basis of data extracted from the patient record, PERSIVAL will generate questions that are meaningful in the clinical context and determine which concepts from the patient record are relevant for the search. The query, augmented with concepts from the patient record, is passed to the search component, triggering both textual and multimedia search. Textual search requires determining which of various distributed resources (some of which may only be accessible through a remote, online interface) may contain relevant documents. The results of this initial search are then fed through a component which more closely examines the retrieved documents by extracting and matching terms from the patient record, and by re-ranking the documents which are more relevant to this patient's case. Multimedia search may initially be triggered by a concept from the patient's record (e.g., "valve regurgitation") which matches diagnostic videos (e.g., echocardiograms), but PERSIVAL also allows follow-up search using image features representing important diagnostic information. The results of the search are passed on to a summarization component. For text documents, different approaches are used depending on whether the end user is a clinician or a patient. For the clinician, an informative summary indicates which results in the retrieved journal articles are relevant to the patient, while for the patient, a mixture of informative and indicative information is used to allow the patient to browse retrieved consumer health information. For video documents, PERSIVAL segments the video and creates a storyboard summary. A layout component will arrange and present the results to the user using a thin client server which allows for the intensive computation to be done at the server, incrementally passing results to the client as they become available. After results are presented, the user ultimately will have the option of refining his query or issuing a new search, though we do not report on this here. We are currently developing an XML interface to connect the components, which at this point in time have been developed separately.

For most of the above mentioned modules, we have identified major challenges, developed initial solutions to some of these challenges, and performed preliminary evaluations. We report in detail our approach to each of these subproblems and the status of current implementation in the sections that follow. Some of the modules, such as the layout planning module and the manager of interactions and feedback between modules are in an early prototype or design stage, and for those we report on our planned approach.

One contribution of our work is a remote application execution infrastructure that we call *Remote JAVA Foundation Classes.* Our infrastructure is built upon standard JAVA technologies (Java Foundation Classes (JFC) and Remote Method Invocation (RMI)) and supports the delivery of a highly interactive user interface over a low-bandwidth network connection. We take a hybrid approach between fat and thin clients that provides the benefits of a thin-client approach (e.g., low-cost terminals and centralized management and security), while offering functionality typically only found in fat clients (e.g., asynchronous server events and multimodal interaction). This is accomplished by delivering a user-interface-toolkit–aware client via HTTP to the web browser, which uses the RMI registry to obtain a remote reference to the application server. Once the client is registered, the server uses RMI to transmit JFC user interface toolkit commands to, and receive events from, the client. By manipulating user interface toolkit components and handling events remotely, our infrastructure provides functionality typically found in remote framebuffer systems (e.g., Tridia VNC [30], Symantec PCAnywhere (`http://www.symantec.com`), and Citrix Metaframe (`http://www.citrix.com`)), while consuming a fraction of the bandwidth.

3. ASKING QUESTIONS

Our approach features the ability to ask questions in context, where the context is drawn from the patient record. The patient record, however, is quite large, often consisting of hundreds of individual text, graph or video documents. It is critical that we extract information that represents information that the patient or physician is interested in and that is relevant to the query and the current clinical situation. Our approach will both allow the user to pose queries and will also automatically generate queries based on data that the end user is currently viewing, if the user indicates the need for information. The user also will ultimately have the option of entering a full question using a speech interface. In that case, we will use evidence-based medicine techniques [31] to determine which information from the record is relevant.

Previous work has examined technological solutions for linking medical records to online information resources. The MedLine Button [6] was the first system to use clinical data as input to real-time searches against a bibliographic database. In that system, patients' diagnoses and procedures were translated into Medical Subject Headings (MeSH) terms, assembled into search statements, and passed on to the MedLine search engine. That work has continued, in the form of *infobuttons* [4] that use a variety of clinical data and a variety of information resources to attempt to address the information needs of clinical information system users. A key part of providing automated searches is an understanding of the information needs that are most likely to arise in a given context. The MedLine Button generated questions that were drawn from a set of *generic queries* [5] which were, in turn, created from a database of actual questions posed to medical librarians. The questions were analyzed to determine recurring semantic patterns of concepts, such that specific concepts could be replaced by general semantic types. For example, "Do chest x-rays cause cancer?" becomes "Does <procedure> cause <disease>?" When a person reviewing a medical record selects, for example, a procedure and disease of interest, the system can then select the generic query that has such semantic types, and then instantiate it with the specific terms of interest. This prior work established a direct link between the data that the user was examining in the clinical application and potential questions that could be posed.

Our work on PERSIVAL makes significant enhancements in several areas: leveraging what is known about the individual patient whose record the clinician is reviewing; matching to an improved model of user information needs; and enabling the user to refine the query. The patient record contains a wealth of information that can be used to suggest and refine user information needs. For example, simply knowing the age or sex of the patient may help in focusing on pediatric or gynecologic searches. We are currently developing a set of rules founded on principles of evidence-based medicine is used to extract relevant facts from the patient record. These patient data are then matched against a new knowledge base of generic queries, which are based on questions asked by clinicians regarding patient care [11]. An enhanced interface enables the user to indicate the focus of the search, select an appropriate query, and refine the query as needed, by replacing or adding terms. A prototype of this enhanced interface and the component for matching patient data has been developed, while we are currently working on evidence-based medicine rules. A prototype link between the New York Presbyterian Hospital's clinical information system (WebCIS) is being created, specifically for a set of cardiology patients whose records have been sanitized. If the user is looking at a cardiology procedure, such as the 12-Lead Electrocardiogram, (which includes a description of abnormalities, such as lateral ischemia, unstable angina or left ventricular hypertrophy) and clicks on a link to PERSIVAL, the system extracts all of the findings from the report and displays them, in tabular fashion, to the user. The user can then select any number of these concepts for submission to PERSIVAL. For example, the user may select *lateral ischemia* and *unstable angina*. PERSIVAL also examines the patient's records for concepts relevant to lateral ischemia and if it finds, for example, that the patient is on the medication *aminophylline*, it will generate questions such as "What is unstable angina?", "What are the risk factors for lateral ischemia?", and "Can aminophylline cause unstable angina?". If the user selects the last of these questions, the system returns and issues the search query

```
User Question Analysis Summary
Type: pharmacologic-agent causes pathologic-finding
Clinical concepts: aminophylline, unstable angina
Search: (((aminophylline[mh]+OR+albuterol[tw])+AND+
```

```
(unstable angina[mh]+OR+unstable angina[tw]))
+AND+(adverse+[tw]+AND+effect[tw]))
```

When the user is examining a narrative report in the patient's record, information is extracted using a natural language information extraction system called MedLEE [12] that was initially developed for the domain of radiographic reports, and was subsequently extended to other domains. Primary findings are extracted as well as modifier relations. Modifier relations, such as negation, time, and uncertainty are frequently expressed in the reports, and are critical to obtain because they change the underlying meaning of the information. MedLEE contains several processing components: i) the preprocessor uses a lexicon containing semantic categories and target forms to segment the report into sections, sentences, words, and atomic phrases; ii) the parser structures the sentences by using a semantics-based phrase structure grammar to specify well-formed structures and their target forms; iii) a phrase regularization component composes phrases when appropriate; and iv) an encoding component maps the target terms to codes associated with a standard vocabulary. MedLEE was independently evaluated a number of times, and was shown to perform effectively for realistic clinical applications; in fact, it was not significantly different from medical experts in detecting specific conditions [17].

For PERSIVAL, MedLEE was extended to handle electrocardiogram and echocardiogram reports; we plan on further extensions in the cardiology domain. The extensions consisted of adding new entries to the lexicon, and of adjusting certain grammar rules. The encoding portion of MedLEE was also fine-tuned in order to generate output encoded into UMLS (Unified Medical Language System [18]) codes. These codes are critical because they form the basis for interoperability of two heterogeneous Natural Language Processing systems in this project: the patient record extraction system and the reranking of search results based on patient matching (Section 4.2). Following extraction of concepts by MedLEE, evidence-based rules are used to determine which of the many concepts are most important.

4. SEARCH

After a patient's or physician's query has been augmented with important information from the patient record, the augmented query is processed by a multimedia search component. PERSIVAL provides search tools over distributed collections of both text and echocardiogram video. For textual documents, search occurs in two stages. In the first stage, the system searches over multiple distributed text collections using a uniform metasearcher. In the second stage, results from the first are reranked by examining characteristics of patients studied in the article and matching those characteristics against concepts found in the patient record. For videos, the system searches relevant video segments from large collections by using automatically extracted annotation labels as well as clinically important multimedia features.

4.1 Searching and Browsing over Distributed Text Collections

Distributed resources available on the Internet do not present uniform searching capabilities, which complicates query processing. Furthermore, these resources differ widely in their topic and along every conceivable dimension. As a central component of PERSIVAL, we are deploying infrastructure for searching over distributed collections. Also, we have developed techniques for automatically classifying document collections into a topic categorization scheme that users can then browse to find the collections of interest.

PERSIVAL's query interface will offer the illusion of a single collection; the *metasearcher* is the component that enables a virtual integrated view of heterogeneous sources. To build our metasearcher we are merging two complementary existing search protocols that have been developed within digital library projects in the United States, namely SDLIP and STARTS, into a combined protocol, which we refer to as SDARTS [14].

SDLIP (Simple Digital Library Interoperability Protocol) [27], jointly developed by Stanford University, the University of California at Berkeley and at Santa Barbara, the San Diego Supercomputer Center, the California Digital Library, and others, defines a layered, uniform interface to search over each collection. SDLIP is a flexible, extensible protocol, and can "host" different query languages and metadata specifications for the sources to export. In particular, the requirements for the metadata interface are minimal, but extensions are allowed and encouraged. As a result, a perfect complement for SDLIP is STARTS (Stanford Protocol Proposal for Internet Retrieval and Search) [13]. STARTS is a high-level protocol that defines, among other things, the specific metadata that sources should export to facilitate metasearching.

SDLIP and STARTS complement each other naturally. At Columbia University, we have combined them into a protocol named SDARTS, which extends SDLIP by instantiating its query language and by defining a rich metadata interface according to what STARTS dictates. The result is a simple, expressive protocol that facilitates the construction of metasearchers [14]. In addition, we have developed a software toolkit to simplify the indexing of "local" document collections so that they are SDARTS compliant, as well as to simplify the construction of "wrappers" around external collections over which we do not have any control. SDARTS and its associated software toolkit provide the necessary infrastructure to incorporate collections into our project with minimal effort.

Metasearchers let users query over distributed collections. An alternative mode of interaction is for users to *browse* Yahoo!-like directories to locate collections of interest and then submit queries to these databases. Recently, commercial web sites have started to *manually* organize web-accessible text collections into hierarchical classification schemes (e.g., see InvisibleWeb at `http://www.invisibleweb.com`). Automating this classification is challenging, since many times the contents of searchable collections on the web are not available other than by querying. For example, consider the PubMed medical database from the National Library of Medicine, which stores medical bibliographic information and links to full-text journals accessible through the web. This database is accessible through a query interface at `http://www.ncbi.nlm.nih.gov/PubMed/`. If we query PubMed for documents with keyword `angina`, PubMed returns 36,150 matches, corresponding to high-quality citations to medical articles. The abstracts and citations are stored locally at the PubMed site and are not distributed over the web. Unfortunately, the high-quality contents of

PubMed are not "crawlable" by traditional search engines. A query on AltaVista for all the pages in the PubMed site with keyword "angina" (i.e., `angina host:www.ncbi.nlm.nih.gov`) returns only three matches. This example illustrates that often we cannot classify a valuable text collection by extracting and analyzing all the documents that it contains.

To automate the classification of searchable collections like PubMed, we have developed a novel technique that learns a small number of *query probes* to issue off-line to the collections [19]. We start with a comprehensive, predefined topic hierarchy with an associated training set of preclassified documents. We then characterize these documents by selecting the best features (i.e., words) using an information theoretic feature selection algorithm that eliminates the words that have the least impact on the class distribution of documents [21]. This step eliminates the features that either do not have enough discriminating power (i.e., words that are not strongly associated with one specific category) or features that are redundant given the presence of another feature. After this feature selection step, we train a rule-based document classifier [8] to produce rules like the following:

IF ibm AND computer THEN Computers
IF diabetes THEN Health
IF cancer AND lung THEN Health

For example, a document having the word "diabetes" will be classified into category "Health" according to this classifier. The next step is to transform each of these rules into query probes, and to adaptively issue the queries to the collections that we want to classify, extracting only the number of matches for each query. The number of documents that match a specific query at a database (e.g., "cancer AND lung") represents the number of documents that would match the corresponding classifier rule if we could run it over every document in the collection. Finally, our method classifies the collections using simply the number of query matches, without retrieving any documents from the collections. As a result, our strategy efficiently produces an accurate collection classification using a small number of query probes (typically fewer than 200 queries of a few words each are needed to classify a collection). Users can then browse the hierarchy of categories to identify the collections that match their information need.

4.2 Reranking Search Results

The query and search modules of our system allow the user to specify and adapt questions according to the clinical context and retrieve a wide variety of relevant documents from multiple sources. However, these modules are primarily *query-oriented*; while some patient information is used to direct the search, the primary focus of these modules is to include documents in the results that match the entered query.

Yet many of the documents that are generally relevant to a query about "unstable angina" may not be of high priority to a specific patient. For example, an article describing complications in patients who have both angina and diabetes should be ranked lower when the patient is not diabetic. On the other hand, given the sample patient record and article sentences shown in Figure 2, we can assume that the given article is very likely to be relevant to the patient.

Patient Record:
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 [...] The patient was admitted to New York Presbyterian Hospital on 12/3/99 [sic] a worsening CHF and **unstable angina** for evaluation for heart transplant.

Medical Article:
This was a multicenter prospective study of consecutive patients admitted to coronary care units with **unstable angina**. Baseline characteristics were age 60.18 ± 16 years, history of **prior myocardial infarction** in 336 patients (32%) [...] In-hospital treatment consisted of [...], angioplasty, or **coronary artery bypass grafting (CABG)** in 25.1% ...

Figure 2: Term Matches between Patient Record and Medical Article

PERSIVAL takes advantage of the patient record information to filter out documents that match the query well but the patient record poorly by reranking the results of the search according to how well they match with key elements of the patient record. To determine the degree of this match, more computationally expensive, natural language processing techniques are applied to the small portion of the entire set of the documents that match the query in the first place.

We base our comparison between patient records and medical journal articles on a common representation of both as lists of important technical terms, with associated values when they occur. Terms, that is words and phrases that capture technical content and have a fixed meaning within a specific domain, contain a large part of the information present in an article or patient record; demographics, diseases, treatment procedures, and drugs are all likely to be represented in the text via terms. In some cases, the term is associated with a value (e.g., "base heart rate over 90", where *base heart rate* is the technical term and *over 90* is the value). Representing both the patient record and the documents as vectors of term-value pairs provides a basis for converting document information into a form that can be used for quantitative comparisons. In particular, we represent terms as UMLS unique identifiers which capture the *semantic concepts* conveyed by the terms.

Patient records are analyzed by MedLEE as described in section 3; scientific articles, which employ more general language and less structure than patient records, are analyzed by the procedure described here. To find terms within scientific articles, we use a variety of surface indicators for each candidate term:

- Its relative frequency in medical and general text; terms are expected to be far more frequent in medical texts.
- Its distributional characteristics across different documents; terms usually bunch together more than ordinary words [1].
- Measures of cohesion between adjacent words help identify multiword terms; the component words of multiword terms and collocations occur together much more frequently than would be expected from their individual marginal frequencies [2].

- Syntax places constraints on terms; usually, terms consist of nouns, possibly premodified by adjectives and postmodified by a single prepositional phrase.

To collect this information from the text, we process it with tokenization and part of speech tools, as well as with a finite state grammar that enforces syntactic constraints on terms, expands invisible term connections from conjunctions (e.g., "unstable and stable angina pectoris" is a variant of "unstable angina pectoris") and associates terms with values by capturing attributive and predicative modifications between terms and numeric or adjective phrases (e.g., "*acute* myocardial infarction, *severe* unstable angina, systolic blood pressure *of 113.6*"). The numeric information from our statistical criteria (the first three indicators of terms above) is combined in a log-linear model [25], a supervised learning technique. By training on a list of established terms (the large scale vocabulary test (LSVT, [24]) we obtain the weights for the variables, producing a measure of how likely a word or phrase is to be a term.

After terms are identified, another algorithm performs the actual matching, measuring the overall importance of a term within both patient record and article. Two factors come into play: the relative specificity of a term and its semantic category, since more specific terms and terms that refer to diseases, treatments, and drugs are more likely to influence the matching. We use the semantic hierarchy in the UMLS to retrieve the semantic category, and, along with term frequency, to measure term specificity [28].

Once terms, values, and semantic categories have been obtained from both the patient record and the document, we calculate their degree of matching as the cosine product of these two vectors, with each term weighed according to the importance value assigned to it. In the example of Figure 2, "unstable angina" and "myocardial infarction" provide a large part of the matching score, since as a matched symptom and a matched disease, respectively, they contribute more than a matched theurapeutic procedure like "CABG". When values are present, a further matching step is executed, which alters the sign and magnitude of the match at that term according to how well the values match. Currently, we detect incompatibilities between values and partial matches between quantitative ranges of values.

The term recognition and matching modules are also used to provide anchor points for the two text summarization modules described below, and "topics" of summaries for the video module, in order for that module to assign labels to and access textbook examples and actual patient ECG videos.

4.3 Search and Organization of Echocardiogram Video

PERSIVAL also provides efficient tools for automatic indexing of echocardiogram videos and searching over large echochardiogram video collections. Echocardiography is an important imaging technique, which assists the cardiologist in the diagnosis of heart abnormalities. Because this method is non-invasive and cheap it is usually available in almost every major healthcare center. For example, at New York Presbyterian hospital there are thousands of echo videos taken and archived each year. However, very few tools and computer facilities are available for indexing and accessing such large video collections. Most echo videos are still stored on analog tapes with limited annotations.

In PERSIVAL, we envision that users will be able to access, search, and interact with digital echo videos efficiently and effectively. Video data will be integrated with other modalities of information and presented to the right users in the right context. For example, doctors, clinicians or medical students may retrieve echo videos of related cases with certain abnormalities from the library in order to compare with prior findings. Such facilities will be very useful in diagnosis, surgical planning, or teaching processes.

Echo video presents unique research challenges and opportunities for video indexing and summarization. Unlike other types of video addressed by existing work, echo video does not include speech, audio, or transcript information that can be used to index the video content. Information is predominantly contained in visual form. On the other hand, there are usually predictable structures in the production of echo videos. Sonographers usually follow a recommended sequence of transducer positions for capturing the two-dimensional echocardiograms. In addition, there is associated information from other modalities, such as ECG and diagnosis reports, which can be used in analyzing the video content or providing useful annotations.

To achieve these goals, our current research involves the following objectives and approaches:

Index video at the syntactic and semantic levels. Working with the domain specialists, we identified the syntactic structures and semantics of echo video. In particular, we developed a view transition model to represent rules used in the echo video scanning procedure. Characterized by the unique position and angle of the transducer, each view captures information about specific anatomical structures of the heart from a specific orientation. One of our objectives is to develop automated algorithms and tools for segmenting and recognizing constituent views in the video. To do this, we analyze unique spatio-temporal visual patterns of anatomic parts and apply a domain-specific view transition model. Results of the automatic tools that we have developed will allow users to randomly access views of interest, interactively browse constituent views at an intuitive level, and selectively transmit important views over networks.

Annotate and organize large collections of video. The video segments and summaries described above can be annotated by labels from view recognition and information contained in the diagnosis reports associated with each video. For example, descriptions of abnormalities related to certain parts (e.g., valves and muscles) can be linked to views in which such abnormalities are most visible. In addition, we are developing a taxonomy for classifying and organizing large collections of representative cases that can be used for research and teaching purposes.

Develop intuitive content-based video search tools. A query to the echo video library may be based on concepts in textual form or multimedia form. For example, terms describing specific abnormalities can be first used to retrieve specific segments of videos and their associated diagnostic findings. After seeing the returned videos (entirely or in a summary form), users may use graphic tools to select regions in the video and ask the system to find other videos showing similar visual patterns that are related to important clinical concepts (e.g., speed and volume of blood flows shown in the video). By combining such search tools using visual features with clinically meaningful categorization, users will be able to find more efficiently specific videos in large collections.

Currently, we have achieved promising results in view segmentation, recognition, and key frame extraction by using automatic image analysis algorithms and view transition models [9]. We are in the process of constructing a video library containing several hundreds of cases with important abnormalities, and evaluating our current tools.

5. PRESENTATION

PERSIVAL must formulate a concise and effective presentation that enables the user to understand the main points of the retrieved documents without having to examine them directly. It must also maintain links to the documents from the summary, allowing users to easily select the documents they judge most relevant. Our approach involves summarization of both text and video, including the ability to provide definitions for unknown terms. A layout component will integrate and cross-link search results, ultimately presenting summaries along with original documents for easy viewing and manipulation. Since PERSIVAL will be made available to end users on a variety of platforms, including low-end PCs, it is important that processing is efficient. We use a thin client server for this purpose.

5.1 Textual Summarization

A textual summary is generated to describe important information across the set of textual documents returned in a search. The method used to generate this multi-document summary depends in part on the document genre. Clinicians are more likely to be interested in seeing medical journal articles or textbooks that are relevant to the patient's case while patients will be more interested in consumer health information. These genres are quite different in how they are structured and thus, we use different techniques to produce a summary in each case. Furthermore, patients and clinicians are likely to be interested in different kinds of information, also requiring different processing. For both types of summaries, our approach involves a unique integration of statistical processing to select relevant phrases with symbolic processing to edit and weave phrases together to form the summary.

Summaries for Clinicians. Medical journal articles, of interest to clinicians, are written in a relatively rigid form, with sections (e.g., results) coming in a more or less predetermined order. Information extraction techniques [32] can take advantage of this structure to locate particular pieces of information. Experimental research articles typically present the outcome of a clinical study for several groups of patients (at least one test group and one control group). In a user study on summarization [26], we found that physicians can determine relevance of an article by quickly skimming the results and recommendations pertaining to the patient under their care. Thus, this is the information that should be included in a summary.

Our summarization module for clinicials [10] extracts this patient-specific information from a medical article. Structure is exploited to find the "results" section, then text categorization techniques are used to separate out sentences within the section that actually describe results. PERSIVAL has been trained to find words and phrases that are good indicators of results (e.g., "predictor"). This stage selects on average one third of the Results section, which is in turn only a portion of the full article. In the final stage of this component, we use pattern matching techniques to

---

In a univariate analysis, NYHA class, pulmonary artery systolic, and atrial fib were associated with a decreased event free survival ([ajc]). But only NYHA class was considered as associated in a multivariate analysis ([ajc]). Prior angina was considered in both univariate and multivariate analysis a predictor of in-hospital morbidity ([1]). Atrial fib was not significant in multivariate analysis ([5]). The occurrence of angina after admission showed a strong univariate relation with the incidence of in-hospital acute MI or death ([5]).

Figure 3: Extracted summary phrases from results reported in journal articles

find particular types of results (e.g., "Multivariate analysis showed ... ") and select only those portions of the sentence that match the patient. The phrases which match information that is currently extracted for the query "What are the risk factors for unstable angina?" are shown in Figure 3; for example, the first sentence in that summary matches a patient who has atrial fibrillation. We are currently working on generating these target phrases from the extracted information. Following each sentence is a pointer to the article in which it was found. Ultimately, this information will be used by the layout component to directly link pieces of the summary to specific documents, enabling selective access to the related documents.

Summaries for patients. For patients, we summarize the set of consumer health documents that are determined to be relevant. In this case, we cannot assume that all patients will be interested in a specific type of information such as results. Instead, we provide information that is commonly repeated across documents and thus provides a synopsis of the set of documents. We follow this with "indicative" descriptions which characterize the kind of information contained within the documents, indicating which documents provide more detail on what topics and which documents are different from others in either content or form.

For the synopsis, we use a similarity tool that we developed for summarization of news [15]. Using statistical measures of pairwise similarity between sentences followed by clustering, it identifies sets of sentences across articles where each set describes similar information. From each set of similar sentences, it extracts one representative sentence to form part of the summary.

For the indicative part of the summary, the system uses a hierarchical representation of common topics found across all documents retrieved in the search. From this tree structure, it can determine the portion of the tree common to all documents, the different formats used, when a document presents more detail than all other documents, and when a document provides information that is not related to information presented in other documents. The summary that PERSIVAL generates in response to the search query "What is unstable angina?" is shown in Figure 4.

5.2 Explanations for Technical Terms

Currently, our summarization module uses terms and phrases that are found in the documents being summarized. However, some of these terms may not be familiar to patients. Ultimately, we want to be able to provide definitions for unfamiliar terms in the summary. To do this, we have

Treatment is designed to prevent or reduce ischemia and minimize symptoms. Angina that cannot be controlled by drugs and lifestyle changes may require surgery. Angina attacks usually last for only a few minutes and most can be relieved by rest. Most often, the discomfort occurs after strenuous physical activity or an emotional upset. A doctor diagnoses angina largely by a person's description of the symptoms. The underlying cause of angina requires careful medical treatment to prevent a heart attack. Not everyone with ischemia experiences angina. If you experience angina, try to stop the activity that precipitated the attack.

Highlighted differences between the documents include:

- This file (5 minute emergency medicine consult) is close in content to the summary
- The Merck manual of medical information contains extensive information on the topic.

Figure 4: Generated summary of documents retrieved for the query "What is unstable angina?"

developed a component of PERSIVAL that can identify and extract medical terminology, along with their definitions and modifiers, from reliable online resources such as the Heart Information Network (HIN, `www.heartinfo.org/reviews`).

In our study, we automatically analyze these resources in order to explore structural and linguistic methods for the identification and extraction of definitions and the terms they define, complementing our work on term extraction (see Section 4.2). This component of PERSIVAL, called DEFINDER (Definition Finder), uses rule-based techniques on text along with the Universal Medical Language System (UMLS) knowledge base. For the definition extraction, DEFINDER uses both shallow text processing (based on cue phrases, structural, and linguistic indicators) and a rich, dependency-oriented lexicalist grammar, the English Slot Grammar [23], for analyzing more complex linguistic phenomena. For example, our analyzer identifies that verapamil and diltiazem should both be included in the category of calcium channel blockers, from text such as: "Nifedipine is one of the three widely prescribed calcium channel blockers. The others are verapamil and diltiazem." In this example, the referring anaphoric phrase "the others" indicates that these three items belong in one category.

Our results show that medical texts for the popular audience, when of high quality, provide a valuable source of medical terminology and definitions. We performed two evaluations: 1) for the definition extraction method and 2) for the quality of defined terms. In the first case our system obtained 84% precision and 83% recall. For the second evaluation we choose a set of 93 terms and their definitions from our corpus and compare them with 3 other online dictionaries (including UMLS). The results presented in [20] show that the dictionaries appear to be incomplete (e.g only 60% of our term set are present and defined in UMLS; 24% of the terms are present but undefined; and 16% were absent altogether). A recent comparison between definitions in the UMLS and those automatically produced by DEFINDER shows that the latter are more useful and readable to lay people. Examples of our results include: (1) angina—the chest pain that occurs when the heart is deprived of oxygen due to diminished blood flow; (2) atrial fibrillation—improper contraction of the upper left chamber of the four-chambered heart; and (3) hypertension—high blood pressure. The output of our system can be used in the creation and enhancement of online terminological resources and in summarization.

5.3 Presentation and Summarization of Echo Video

For the video data, PERSIVAL provides efficient tools to present and summarize the retrieved videos. For example, from a patient's record, we know the patient was diagnosed to possibly have mitral valve regurgitation. Using this term as a query input, we retrieve video segments related to this abnormality from the patient's echo video or other videos in the library. After seeing the displayed videos, users may also use the content-based search tools described earlier to find more specific videos showing visual characteristics revealing important clinical information.

Depending on the context and needs, users may want to view returned videos at different lengths. Based on their inherent structures and semantics, echo videos can be summarized at different levels with different lengths. The first level includes presentation of key frames and associated data showing the most informative view of the heart in each segment. At the second level, each segment is represented by one complete heart cycle (clinical summary) that also shows the dynamics of the heart. At the third level, a highlight version of video can be produced by concatenating several short video clips each of which represents one single heart cycle from selected views. Such video highlights will be very useful for accessing the video library through bandwidth limited networks, such as wireless networks. Figure 5 shows a window including a key frame summary of selected video segments related to mitral valve.

5.4 Automated Layout

To help create a high quality user interface, we are developing methods for laying out automatically the material being presented, so that coherent, understandable transitions are employed as the presentation changes (e.g., by adding or deleting a display, or changing a display's contents). We are building on our previous research on automated generation and layout, which uses hierarchical decomposition planning techniques [33]. Unlike that earlier work, which generates all components on a single display, one challenge in PERSIVAL is to manage a set of displays, including some that are externally generated (e.g., the existing WebCIS and PatCIS systems). To determine a suitable approach, we surveyed previous work on automated user interface layout [22], and have begun to apply some of the best existing ideas to PERSIVAL. New directions include employing evaluation techniques to resolve inconsistencies in automatically generated constraints, and adding zoomable user interface components [29] to the set of rendering possibilities available to the automated layout system.

6. COGNITIVE STUDIES

At the same time as we develop the system, we are also carrying out formative evaluation that can help us determine how best to implement personalization. Evaluative work to date has focused on identifying how physicians assess relevancy of information in the context of particular patients, for tasks involving both searching (i.e. finding articles relevant to a patient) and summarization (i.e. extracting from an article information relevant to a patient). The objective

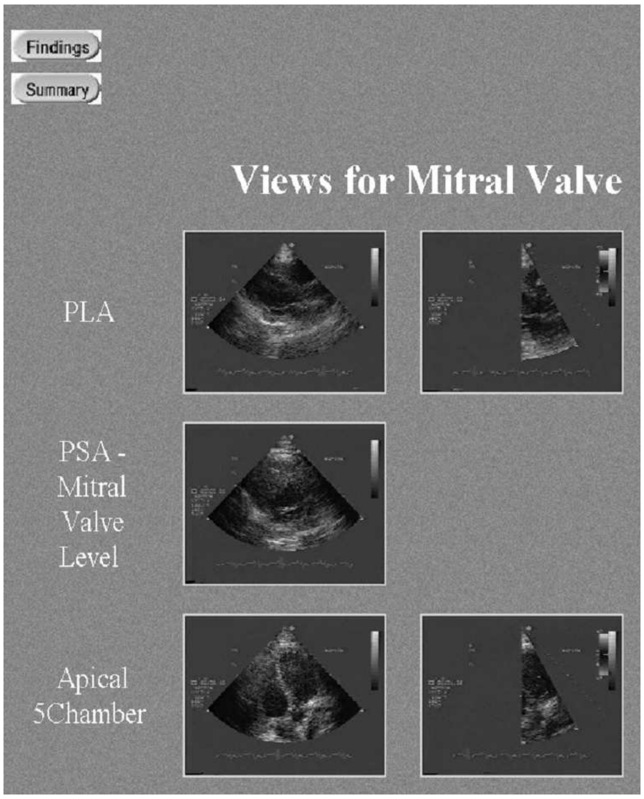

Figure 5: Echocardiogram Visual Summary

of this work has been to provide input to the design of system components, based on the empirical data, and to lay the groundwork for subsequent formative and summative evaluation of the corresponding PERSIVAL components.

In our first studies a test set of articles in the area of cardiology were collected. Other study materials included medical records from three cardiac patients, including electrocardiogram and echogram reports, as well as a written description of each patent. The study task involved having ten physicians review the information about the patients (one at a time) and then indicate whether each of the articles were relevant to care of the patient. Subjects were asked to "'think aloud" while doing this task and the audio recorded sessions were analyzed for strategies used in assessing relevance. Physicians were also asked to indicate which statements in the articles should be included in a summary for that patient. Results to date indicate that a number of different strategies were applied by subjects in scanning the articles and in deciding on their relevance. Differences in ratings were found to be related to a number of factors, including individual interests and level of physician expertise.

We are currently extending our study of relevance rating to a design where physician subjects are asked to rate relevancy of articles in the context of specific medical situations. The approach will later be extended to evaluation of PERSIVAL components in order to compare processing of information by physicians with that of automated system components. Plans are also being made for assessing the usability of other PERSIVAL components as prototype implementations become available, including visual summarization and presentation. Related ongoing evaluations involve audio recording and analysis of actual questions asked during intensive care rounds and assessment of information needs in naturalistic health care settings.

7. CONCLUSIONS AND CURRENT DIRECTIONS

We have shown how information from the patient record can be used to personalize the processes of search and summarization across multimedia information. To date, our work has focused on developing the components of PERSIVAL; we have developed a user query component that can use clinical context to help the user formulate meaningful queries and extract important information from the record, a distributed online multimedia search component that uses machine learning to find relevant sources and patient information to rerank articles, and a multimedia presentation component that uses patient information to determine relevance, automatically finds explanations of terms, uses segmentation and domain knowledge to summarize echocardiogram video, and ultimately will integrate this information in automated layout. Our next steps will be to develop interfaces between the currently separated components and automatically feed information from one stage to the next. We will also be focusing on further formative evaluation that can be used to improve personalization in PERSIVAL.

ACKNOWLEDGEMENTS

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.

The authors would like to acknowledge the support of students involved in the project who have had many of the ideas and done most of the programming, including Kathy Dunn, Noemie Elhadad, Shahram Ebadollahi, Yun Ho, Panagiotis Ipeirotis, Simon Lok, Min-Yen Kan, Eneida Mendonça, Smaranda Muresan, and Sergey Sigelman.

REFERENCES

[1] K. W. Church. Empirical estimates of adaptation: The chance of two Noriegas is closer to $p/2$ than p^2. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, volume 1, pages 180–186, Saarbrücken, Germany, August 2000.

[2] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**:22–29, 1990.

[3] J. Cimino, J. Li, E. Mendonça, S. Sengupta, V. Patel, and A. Kushniruk. An evaluation of patient access to their electronic medical records via the world wide web. In *Proceedings of the Annual Fall Symposium of the American Medical Informatics Association*, pages 151–5, Los Angeles, CA, November 2000.

[4] J. J. Cimino. From data to knowledge through concept-oriented terminologies: Experience with the Medical Entities Dictionary. *Journal of the American Medical Informatics Association*, **7**(3):288–297, 2000.

[5] J. J. Cimino, A. Aguirre, S. B. Johnson, and P. Peng. Generic queries for meeting clinical needs. *Bulletin of the Medical Library Association*, **81**(2):195–206, 1993.

[6] J. J. Cimino, S. B. Johnson, A. Aguirre, N. Roderer, and P. D. Clayton. The MedLine button. In *Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care*, pages 81–85, Baltimore, Maryland, November 1992.

[7] 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.

[8] W. W. Cohen. Learning trees and rules with set-valued features. In *Proceedings of AAAI'96*, 1996.

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

[10] N. Elhadad and K. R. McKeown. Towards generating patient specific summaries of medical articles. In *Proceedings of the NAACL Workshop on Automatic Summarization*, June 2001.

[11] J. W. Ely, J. A. Osheroff, M. H. Ebell, G. R. Bergus, B. T. Levy, M. L. Chambliss, and E. R. Evans. Analysis of questions asked by family doctors regarding patient care. *BMJ*, **319**(7206):358–361, 1999.

[12] C. Friedman, G. Hripcsak, W. DuMouchel, S. B. Johnson, and P. D. Clayton. Natural language processing in an operational clinical information system. *Natural Language Engineering*, **1**(1):83–108, 1995.

[13] L. Gravano, C.-C. K. Chang, H. Garcia-Molina, and A. Paepcke. STARTS: Stanford proposal for Internet meta-searching. In *Proceedings of the 1997 ACM International Conference on Management of Data (SIGMOD-97)*, May 1997.

[14] N. Green, P. G. Ipeirotis, and L. Gravano. SDLIP + STARTS = SDARTS: A protocol and toolkit for metasearching. In *Proceedings of the First ACM and IEEE Joint Conference on Digital Libraries (JCDL 2001)*, June 2001.

[15] V. Hatzivassiloglou, J. L. Klavans, and E. Eskin. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 203–212, College Park, Maryland, June 1999.

[16] G. Hripcsak, J. Cimino, and S. Sengupta. WebCIS: Large scale deployment of a Web-based clinical information system. *Journal of the American Medical Informatics Association*, **6**:804–8, 1999.

[17] G. Hripcsak, C. Friedman, P. I. Alderson, W. DuMouchel, S. B. Johnson, and P. D. Clayton. Unlocking clinical data from narrative reports: A study of natural language processing. *Annals of Internal Medicine*, **122**(9):681–688, May 1995.

[18] B. Humphreys, D. Lindberg, H. Schoolman, and G. Barnett. The Unified Medical Language System: An informatics research collaboration. *Journal of the American Medical Informatics Association*, **5**, 1998.

[19] P. G. Ipeirotis, L. Gravano, and M. Sahami. Probe, count, and classify: Categorizing hidden Web databases. In *Proceedings of the 2001 ACM International Conference on Management of Data (SIGMOD 2001)*, May 2001.

[20] J. Klavans and S. Muresan. DEFINDER: Rule-based methods for the extraction of medical terminology and their associated definitions from on-line text. In *Proceedings of the Annual Fall Symposium of the American Medical Informatics Association*, 2000.

[21] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML-97)*, pages 170–178, 1997.

[22] S. Lok and S. Feiner. A survey of automated layout techniques for information presentations. In *Proceedings of Smart Graphics 2001 (Int. Symp. on Smart Graphics)*, Hawthorne, NY, March 2001.

[23] M. McCord. English slot grammar. Technical report, IBM, 1990.

[24] A. T. McCray, M. L. Cheh, A. K. Bangalore, K. Rafei, A. M. Razi, G. Divita, and P. Z. Stavri. Conducting the NLM/AHCPR large scale vocabulary test: A distributed Internet-based experiment. In *Proceedings of the Annual AMIA Fall Symposium*, 1997.

[25] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 2nd edition, 1989.

[26] K. R. McKeown, D. A. Jordan, and V. Hatzivassiloglou. Generating patient-specific summaries of online literature. In *Proceedings of the 1998 AAAI Spring Symposium on Intelligent Text Summarization*, pages 34–43, Stanford, California, March 1998.

[27] A. Paepcke, R. Brandriff, G. Janee, R. Larson, B. Ludaescher, S. Melnik, and S. Raghavan. Search middleware and the Simple Digital Library Interoperability Protocol. *D-Lib Magazine*, **6**(3), 2000.

[28] R. J. Passonneau, K. K. Kukich, J. Robin, V. Hatzivassiloglou, L. Lefkowitz, and H. Jing. Generating summaries of work flow diagrams. In *Proceedings of the International Conference on Natural Language Processing and Industrial Applications*, pages 204–210, New Brunswick, Canada, June 1996.

[29] K. Perlin and D. Fox. PAD: An alternative approach to the computer interface. In *Proc. SIGGRAPH '93*, pages 57–64, Anaheim, California, August 1993.

[30] T. Richardson, Q. Stafford-Fraser, K. Wood, and A. Hopper. Virtual network computing. *IEEE Internet Computing*, **2**:33–38, 1998.

[31] 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, 2nd edition, 1991.

[32] B. M. Sundheim, editor. *Proceedings of the Sixth Message Understanding Conference (MUC-6).* Morgan Kaufmann, San Mateo, California, 1996.

[33] M. Zhou and S. Feiner. Top-down hierarchical planning of coherent visual discourse. In *Proc. IUI '97 (1997 Int. Conf. on Intelligent User Interfaces)*, pages 129–136, Orlando, Florida, January 1997.