A Medical Digital Library to Support Scenario and User-Tailored Information Retrieval*

Wesley W. Chu†, IEEE Fellow, David B. Johnson†, Hooshang Kangarloo‡

Abstract— **Current large scale information sources are designed to support general queries and lack the ability to support scenario specific information navigation, gathering, and presentation. As a result, users are often unable to obtain desired specific information within a well defined subject area. Today's information systems do not provide efficient content navigation, incremental appropriate matching, or content correlation. We are developing the following innovative technologies to remedy these problems: (1) Scenario-based proxies, enabling the gathering and filtering of information customized for users within a pre-defined domain; (2) Context-sensitive navigation and matching, providing approximate matching and similarity links when an exact match to a user's request is unavailable; (3) Content correlation of documents, creating semantic links between documents and information sources; and (4) User models for customization of retrieved information and result presentation. A digital medical library is currently being constructed using these technologies to provide customized information for the user. The technologies are general in nature and can provide custom and scenario-specific information in many other domains (e.g. crisis management).**

I. Introduction

Information access on the WWW is designed to support general queries for any possible subject. Navigating the web to find *specific* information, however, is frustrating, labor intensive and time consuming.

Consider the following example of a young Hispanic woman with a high school education, who complains of chronic headache and has been under treatment for hypothyroidism. Using WWW search engines, she found over 10,000 references on headache, but the information was too confusing and did not reflect her specific condition. Many of the references were testimonials, advertisements, as well as general medical information about headaches, but not specific to her symptoms. Her physician also attempted to find specific information on the WWW related to chronic headache in a young Hispanic woman suffering from hypothyroidism, but failed to find any appropriate information which could enhance management of this patient. The main reason is due to the fact that current search engines do not provide effective content navigation or appropriate matching to user specified queries.

Therefore, we are developing the following innovative technologies to remedy these problems: (1) *Scenario-based proxies*, a framework enabling the gathering and filtering of information customized for users within a pre-defined subject domain; (2) *Context-sensitive navigation and matching*, providing approximate matching and similarity links when an exact match to user's request is unavailable; (3) *Content correlation of documents*, creating semantic links between documents and information sources; and (4) *User models* for customization of retrieved information and result presentation. Using these technologies, we are developing a digital medical library providing customized information for users (Figure 1).

We hope the resulting scenario-based digital medical library will empower active patient participation, facilitate practice of evidence-based medicine by health care providers, and enhance the evaluation and understanding of medical intervention and their cost-effectiveness ratios. Since health care remains among the highest public priorities and as traditional methods of "managed care" have raised concerns related to quality, a unique solution based on information technology and education can significantly improve quality and cost-effectiveness of medical intervention. Further, these technologies are general in nature and can provide custom and scenario-specific information in many other domains (e.g. crisis management).

II. Scenario-Based Proxies

To provide focused information search and navigation, we will express the information needs of a specific domain as a *scenario*. For example, health care provider and patient information needs can be summarized and abstracted in a *medical scenario*. Based on the scenario, relevant information can be obtained by direct access, similar-to search or content correlation links of the information sources expressed via the scenario-based information web (Figures 1 and 2).

A *scenario-based proxy* (Figure 1) is a framework consisting of: (1) A *scenario information web* which provides access paths to information sources; and (2) User model guided knowledge-based intelligent access to information sources by approximate matching of resources, content navigation, and content correlation. Each proxy is tailored for the information needs of a specific domain and represented by a common ontology modeling its objects and concepts.

An *ontology* [1], [2] is an explicit specification of all terms and information that need to be communicated. The ontology enables us to provide information matching, correlation, and access in the proxy. For example, a patient with chronic headaches in the frontal region includes the following features or ontological terms: **condition:** *headache* = head pain = ache and pains in head, **duration:** *chronic* = longtime = several months; and **location:** *frontal* = forehead = front of head. Thus, a query using a patient specified chief complaint, "forehead pain for a long time",

* *This work is supported by grants from the NSF (IRI-9619345) and NIH (CA51198-07).* †Department of Computer Science, ‡Department of Radiology, University of California, Los Angeles (email: wwc@cs.ucla.edu, djohnson@cs.ucla.edu)

the proxy matches this information to chronic headache. Utilizing the proxy the user can retrieve scenario specific information from the information web.

A. Scenario-based Information Web

The scenario-based information web embodies three fundamental principles: (1) it offers information as objects in a model of specific domain concepts; (2) it populates the concepts from *only* those sources that need to be accessed for relevant content; and (3) it interlinks concepts to traverse the relevant portion of the scenario-based web in an incremental manner. These principles allow users (e.g., medical personnel) to access information that is relevant to their current specific needs. The contextual focus provided by scenario-based proxies is critical to the success of information access.

All users in a specific domain will access information through the same proxy. The proxy must provide guidance to relevant information related to known objects. To provide this functionality, the scenario-based information web enables access to a network of information, and also:

- interlinks information based on *similar-to* links, and
- interlinks information based on *content correlation* via patterns (Section IV)

Figure 2 shows the scenario-based information web for a medical application proxy. The web links the objects in the domain and their inter-relationships among information sources. For example, given a 24 year old female patient with hypothyroidism, chronic headaches and brain lesion located in cerebellum with high signal intensity, a request for information regarding this patient will correlate *patient* information with *medical literature* based on the patient's *conditions* and *symptoms*. Furthermore, using information specified in the user query, the scenario-based information web can provide more focused searching to access information and match relevant documents by content. The resulting information taken from the medical literature will therefore be more specific to the patient's medical condition. Likewise, when the physician presents the symptoms (e.g. headache, abdominal pain) and conditions (e.g. osteoporosis) of the patient, the proxy will correlate the symptoms and conditions with medical literature and also with other patients having similar symptoms, conditions and medical history. The patient-specific information can be used to determine the work-up and treatment protocols used by other physicians.

B. Proxy Implementation

Implementing the proxies involves (1) accessing information sources, and (2) identifying when a source applies to a specific context. Accessing the source requires that the proxy interface with the underlying intelligent information mediation and retrieval service. The scenario-based proxies and the mediation and retrieval service must use the same ontology, allowing easy creation of the scenario-based information web by domain experts, who only need to follow the ontology when linking a domain concept to a specific set of data sources. For example, if a domain expert creates a *patient* object, it is automatically mapped to a set of data sources in the mediation and retrieval service. These sources could be comprised of the patient's medical records, images, family history reports, and medical insurance records, which can be translated, referenced and communicated by ontological terms. Thus, the domain expert can use ontological terms *chronic headache*, and the mediation and retrieval service will provide the relevant information sources.

Selection of a relevant source is achieved through the use of the scenario-based information web contained within the proxy. The web represents the objects in the domain and their inter-relationships. This network of objects and relationships can determine what information sources are relevant to a specific query.

C. Related Work

Current resource discovery facilities focus on simple browsing, indexing, and searching facilities [3], [4], [5]. These facilities provide access to data in the information space, but leave the task of intelligently selecting resources and traversing through the network to the user. Work at the University of Colorado seeks to expand these facilities by more automated search routines (Harvest) and more flexible forms of searching (i.e., search by kinds of resources, search based on degree of shared interest, etc.) [6]. The Harvest system, though, only supports an "atomic query" mode where each information access is taken as a new request, independent of previous requests, and returns flat lists of ranked resources.

The "intelligent agent" community has introduced methods for assisting users with maintaining focus in large information spaces [7], [8], [9]. These techniques revolve around agents that model users and tailor the users' perception of the information space. The tailoring attempts to provide contextual focus and relevancy rankings of information. The agents maintain a persistent (and evolving) model of the user to base information focusing. The scenario-based proxies we have designed will use user model guided content navigation, content correlation, and intelligent resource mediation, facilitating information retrieval for users in a specific domain.

III. Content-based Navigation and Matching

Content-based navigation provides navigation in an information space based on content similarity. When a user requests the traversal of a *similar-to* link, the proxy transforms this request into a query for objects similar to the target instances based on a set of user specified attributes or by the user's model and context. Thus, the *similar-to* link is a mechanism for incrementally traversing an information web and constructing neighborhoods of relevant information. The similarity measure can be based on semantic distance, the context, and the instance values and their frequency of occurrences. We incorporate such multi-faceted similarity in the information web by extending our past success in approximate query answering (CoBase) so as to provide context-sensitive and controlled *similar-to* links.

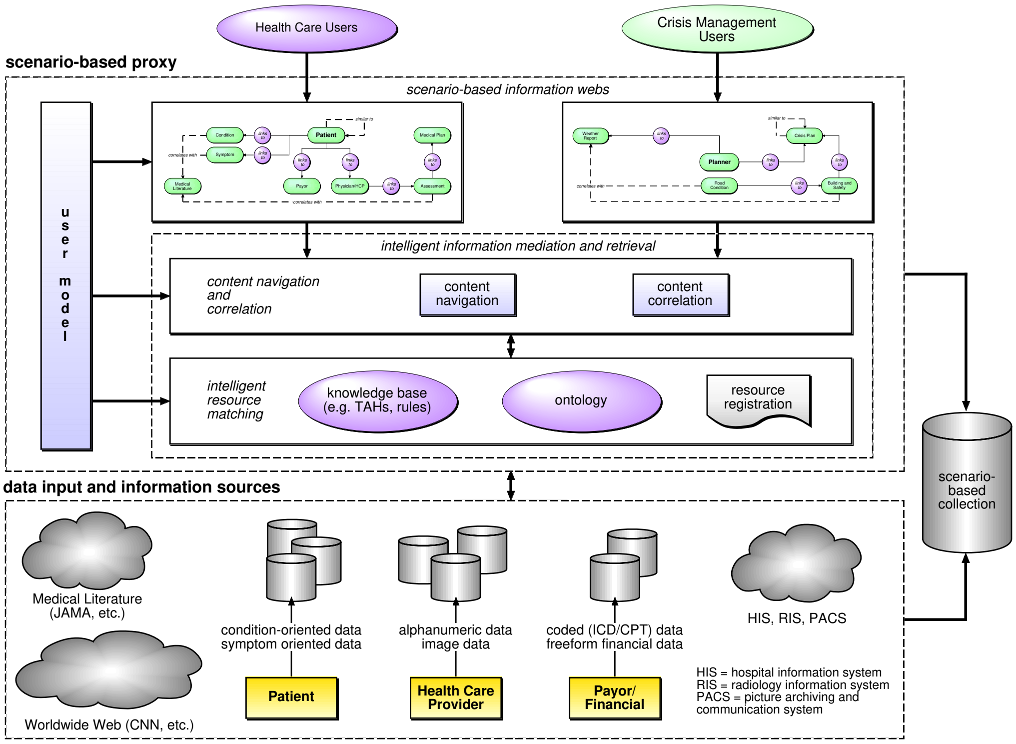


Fig. 1. Intelligent information retrieval guided by scenario-based proxies.

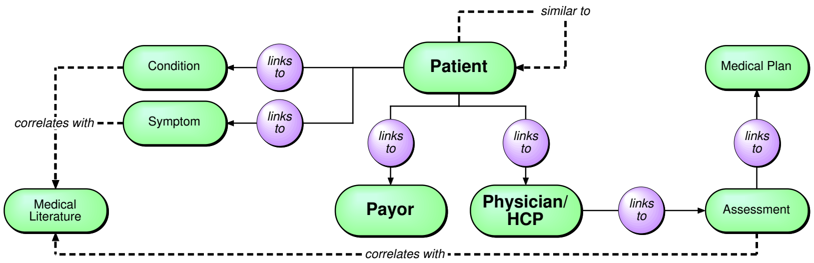


Fig. 2. Scenario-based information web for a medical application

A. Technical Basis

The cooperative database system, CoBase [10], provides *approximate and conceptual query answers* that `approximately` match query conditions when exact answers are unavailable. The knowledge base used by CoBase is a tree structure hierarchy called the type abstraction hierarchy (TAH) [11]. Using such a representation, CoBase is able to provide a structured, efficient, and scalable technique to relax attribute values incrementally. Such relaxation can be controlled by the user or through defaults provided by the user model.

Figure 3 depicts the data flow of query relaxation. When a query is presented to CoBase, the pre-processor first relaxes any explicit cooperative operators (e.g. approximate, near-to, similar-to) in the query. The modified query is then presented to the underlying database system for execution. If no answers are returned, then the *relaxation manager* uses the type abstraction hierarchies to iteratively perform query relaxation. The policy for relaxation control depends on many factors, including the user's profile and query context. The list of relaxation control operators includes: attribute relaxation order (e.g. *size* first, then *location*), non-relaxable attribute(s), the number of answers required, etc. The relaxation manager combines these control policies to generate approximate answers that best match the user's original query. Finally, a post processing module filters and ranks answers according to a given error measure (e.g. mean-square error).

A.1 Type Abstraction Hierarchies (TAHs)

Type Abstraction Hierarchies (TAH) provide multi-level classification of instances within a specific domain [11]. TAH nodes can be labeled by conceptual terms that describe the value ranges covered by their children nodes (See Figure 4). Unmatched attribute values or concepts can be relaxed and approximately matched by moving up the type

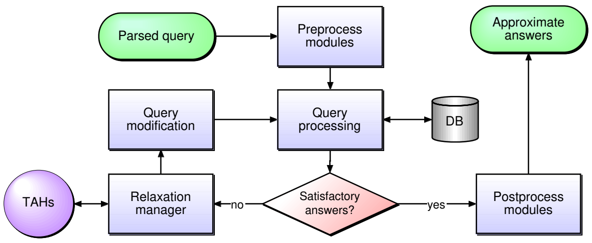


Fig. 3. Dataflow of Query Relaxation.

abstraction hierarchy (generalization) which broadens the search scope, and then moving down the type abstraction hierarchy (specialization) to find an approximate match. The quality of the approximations to the exact match, known as the relaxation error, can be computed and will be given [11]. Many objects require multi–dimensional representation. For example, the location of an object on a 2-D plane, is represented by two attributes, x and y. Therefore we have extended TAHs to classify multiple attributes (i.e., database tuples) capturing generalization and specialization between multi-dimensional data instances [12].

TAHs can capture different types of knowledge, such as database attribute values, image features and text pattern features. Thus, TAHs can be generated from different information sources, such as structured databases, image feature databases and ontological terms used in resource registration (Figure 6). For example, Figure 4(b) is a multi-attribute TAH representing features of brain tumors. The features (intensity, location, size) are selected by domain experts and extracted from image databases via segmentation [13], [14] (outlined contours of a tumor) and feature extraction (compute features) algorithms [15]. Thus a feature database is created and each feature tuple corresponds to a tumor and also links to the image that contains this tumor. These feature tuples can be classified by clustering algorithms and the resulting classification is represented as TAHs.

A.2 Automatic Construction of TAHs from Information Sources

TAHs can be generated and maintained automatically based on the data instances. For numerical attribute values, our Distribution Sensitive Clustering algorithm (DISC) method considers both *frequency* and *value* distributions of data, making the discovered concepts more content sensitive [11]. To optimize the quality of generated hierarchies, DISC partitions the data set of one or more attributes selected by the user into clusters such that the average *relaxation error* (the value difference between the instance value and the target value) is minimized [12]. For non-numerical attribute values, our Pattern Based Knowledge Induction (PBKI) method first derives IF-THEN rules for all pairs of attributes. Based on the derived rules, PBKI then computes co-occurrences between all pairs of values for a single attribute [16]. Starting with each value as a cluster, PBKI recursively merges clusters with highest co-occurrence to form larger clusters. The merging process is terminated when there is only one cluster left.

Both DISC and PBKI have polynomial time complexity, making them scalable to large application domains. To demonstrate the effectiveness of DISC and PBKI, we have used them to generate TAHs for a large transportation database which consists of 94 relations, the largest one of which has 12 attributes and 195,598 tuples. The results show that our methods are efficient, taking less than a minute on a Sun SPARC 10 Workstation to generate TAHs for all but a few cases with large number of tuples, in which case the TAHs can be generated in less than 20 minutes.

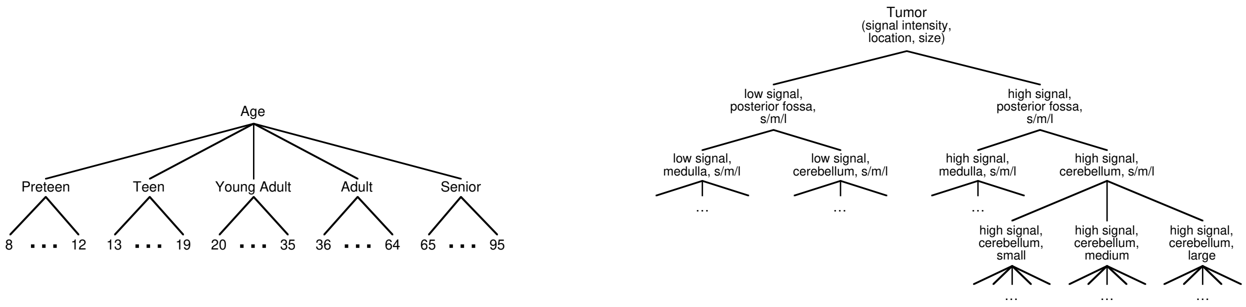
B. Intelligent Resource Registration

Intelligent Resource Registration is comprised of *meta-data resource registration* with *approximate resource matching*. Meta-data resource registration organizes the vast amount of information available to users on the WWW or other data sources to facilitate the location of specific information. The registration is viewed as a semantic labeling of information sources through content summarization. Basically, a semantic tree structure of the information sources will be built, such that each branch of the tree will represent a more conceptually specific set of data sources. These semantic labels and summarizations will follow the common ontology used throughout the system, facilitating the interface with the content-sensitive registration level.

In Figure 5(a) we depict an unstructured information network. The network consists of a set of uninterpreted information sources on various topics. Figure 5(b) depicts the semantic classification required to make information sources (e.g., web pages) available to the proxy. First, each source is annotated with *meta-data*, or the type of information it contains. For example, sources S_1, S_2 and S_5, of Figure 5, contain `medical report data` information. Each report has attributes describing the patient's `treatment` and outcome, and the results of their `physical` in the report. Despite the addition of meta-data, without further categorization, sources S_1, S_2 and S_5 look identical to the system. Therefore, whenever a doctor wants `patient reports`, S_1, S_2 and S_5 must be accessed and searched for appropriate instances. To decrease the number of sources accessed and searched, *content summarization* is needed to further classify the sources within their scope. For example, when a physician wants to evaluate a patient's headache, then all data related to the headache, including a brain MRI, are provided; if evaluation pertains to the patient's neck pain, then only relevant information to neck pain, including bone images, are provided.

B.1 Approximate Resource Matching

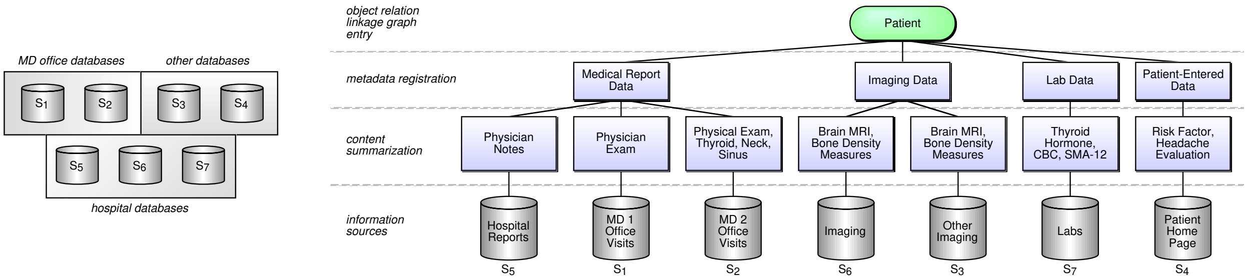
Approximate resource matching allows the system to perform inexact matching of requests to information sources. When a proxy requests information from the intelligent resource registration, the request will be specified in domain ontological terms. If an exact match yields no answers, the intelligent resource registration can *relax* information requests to retrieve approximate and useful in-



(a) Age Classification and concepts defined by domain experts

(b) A multi-attribute brain tumor TAH

Fig. 4. Example of Type Abstraction Hierarchies



(a) Uninterpreted sources

(b) Ontological spaces and content spaces for a patient

Fig. 5. Information spaces to meta-data spaces in the medical domain.

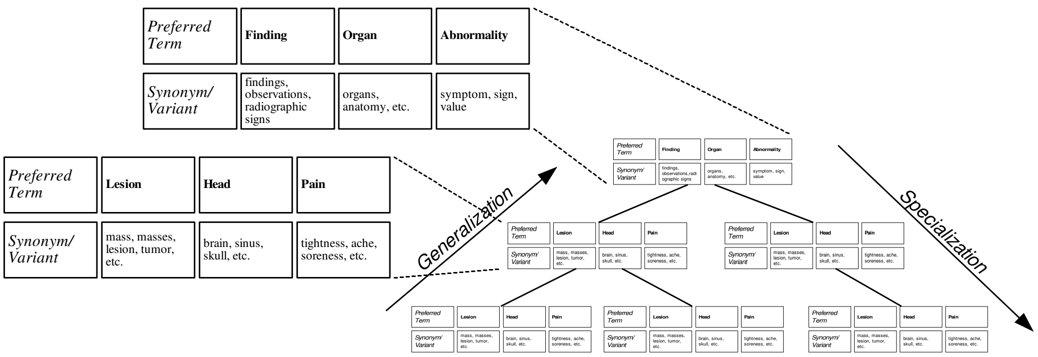


Fig. 6. Type Abstraction Hierarchy for approximate matching in resource registration

formation.

Approximate resource matching by relaxation of ontological terms aids in resource matching. TAHs constructed from ontological terms can be used to relax resource requests. Approximate resource matching can enhance the system in two situations: if a request matches some resources but the information does not satisfy the user, or if the request does not match any resources at all. For instance, a request for information on "sinus" may match a few sources, but these sources may not meet the user's needs. The system can relax "sinus" to "skull" and "head and neck" to enlarge the search scope (Figure 6).

C. Related Work

Previously proposed cooperative database systems use a rule based approach [17], [18] which lacks the structure to provide query relaxation and is hard to scale and maintain. In our approach, knowledge is structurally represented at different levels of abstraction by the TAHs, thus providing a structured way of performing query condition relaxation. Further, TAHs can be automatically generated from feature-based data sets for both numerical and non-numerical data.

Our knowledge acquisition methods (DISC [11] and PBKI [16]) derive hierarchical nearness structures sensitive to the distribution of attribute values in the database. Comparable automatic acquisition techniques do not combine both the hierarchical and distribution sensitive concepts. For instance, COBWEB [19] and CLASSIT [20] use *category utility* as a similarity measure and do not adjust to the distribution (or population) of data values. The similarity measure used in WITT [21] is based on pair-wise *attribute* correlation rather than the nearness of attribute *values* and thus it is not suitable for deriving approximate answers.

IV. Content Correlation

As new documents are added to an information web, either by agents searching the web or through an information resource registration system, it is useful to provide links between related and relevant documents. Current techniques use well-known similarity measures to link textual documents together. We are extending these techniques by indexing on *semantic content* (i.e., what a document is about or what a picture shows) based on an ontology, with patterns of *document features* (i.e., word frequencies, color histograms and textures) extracted from data. Using such semantic indexing, the system can infer the content of unstructured documents (text and images) by correlating document features with previously indexed content.

A. Content Indexing

Content indexing links documents based on their content. Documents are clustered with common content specified by a set of patterns. Each pattern consists of a set of features and their co-occurring frequency (confidence).

For example, given several patient documents that describe the symptoms and conditions associated with hypothyroidism, the common document features related to the symptoms and conditions of that disease, treatment protocols, patient demography, etc. can be used to create a set of patterns. The patterns capture the key document features to infer the meaning of the document. For a set of documents (structured or unstructured text, or images), D, specific semantic features, D_f, that describe the documents are derived through data mining. Based on these set of features, patterns that consists of a set of co-occurring features can be formed by domain experts or automatically, based on their frequency of occurrence. Each pattern is also linked with the corresponding document cluster that contains the pattern.

To support pattern relaxation, the patterns can be organized into a hierarchy pattern organization (Figure 7). Parent nodes $P_1(D_1)$ for document set D_1 represent more general patterns (e.g. headache) with its children nodes ($P_{11}(D_{11})$ – headache, young adult; $P_{12}(D_{12})$; and $P_{13}(D_{13})$) representing specialized patterns. Document clusters corresponding to the patterns can be organized accordingly. This provides a structured way to enlarge the correlation scope and obtain a larger set of correlated documents (e.g. D_{11}, D_{12}, and $D_{13} \subseteq D_1$).

The target document, T, can be translated into a target feature representation, T_f (i.e., finding, severity, and location features, patient demography). A content index, CI, can then be formed by selecting the set of the patterns that match with the target pattern with high confidence (k_i). For example:

$$\begin{aligned} CI &= (Pattern_1, Pattern_2, Pattern_3)\ where \\ Pattern_1 &= (k_1, headache, hypothyroidism, young, D_1) \\ Pattern_2 &= (k_2, lesion_1, location_1, size_1, texture_1, D_2) \\ Pattern_3 &= (k_3, lesion_2, severity_1, T2\ level_1, D_3) \end{aligned}$$

$Pattern_1$, from a textual patient record, states that headache co-occurs with hypothyroidism in young adults with $confidence_1$ ($0 < confidence_1 < 1$) and D_1 links a document set that contains the pattern. $Pattern_2$ represents a pattern from an image describing a lesion's location and size. $Pattern_3$ is also an image-based pattern, but based on the relationship of the severity of a headache with signal intensity of a brain lesion.

B. Content Linking

Correlated-with is a semantic link joining related documents with different content types represented by patterns. Consider the existence of two data sets: one target set consisting of patient records, radiology images and reports, and another set consisting of web-based medical literature. Further, the features of the patient reports have been extracted [22], [23]. Thus the content index can be formed by selecting the patterns of the medical literature that match the target pattern with high confidence. Using the content index, for example, feature descriptions in radiology reports can be correlated with medical information found on the WWW or a radiology report can be *correlated-with* a pathology report. Like a *similar-to* link, a *correlated-with* link provides a fundamental mechanism for linking and incrementally traversing an information web based on patterns in the content index.

Let us consider the process of creating a *correlated-with* link between patient records based on a patient specific target pattern T, "24 year old female patient with hypothyroidism, chronic headaches and brain lesion located in cerebellum with high signal intensity." Based on this target pattern, the content index for correlated medical documents is formed. By correlating the target pattern with pattern sets in the content index, a set of correlated documents is returned.

If the resulting correlated document set yields too few results, the system can enlarge the correlation scope by relaxing terms in the pattern (e.g. "24-year-old female" to "young adult", Figure 4). Furthermore, we can relax the lesion location from "cerebellum" to "posterior fossa" (Figure 4b). Using the relaxed pattern, content correlation can match documents with features like "treatment protocol for hypothyroidism in young adult patients." Further, through associative query answering [24], relevant information such as the success rate, and side effects associated with the treatment procedure can be provided to the user.

C. Feature Extraction for Correlation

A *pattern* can be defined from a data set as a set of unique features (i.e., words or word stems) that characterize some element of the data set. For example, text data mining can be considered the process of discovering patterns which occur in a set of documents. In text, patterns can take the form of one term or sets of terms which co-occur together [25]. For all types of data, we are interested in searching for patterns of features which co-occur frequently and together have some semantical significance.

Several methods can be used to extract patterns from text documents. The frequency of word or word n-gram occurrence can be used by experts to find useful informa-

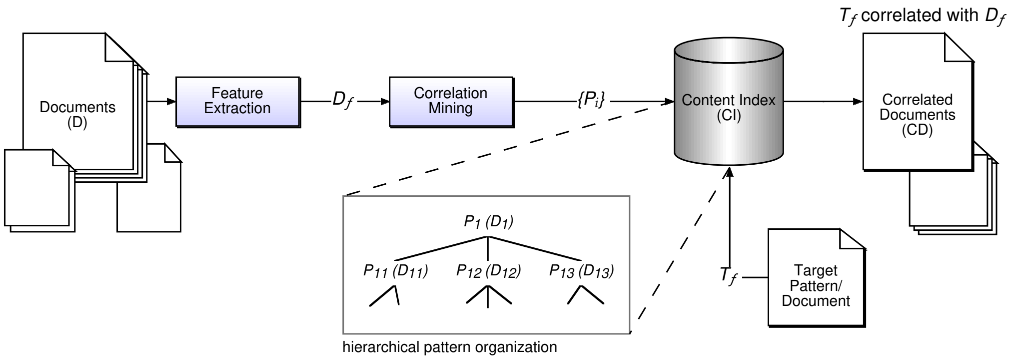


Fig. 7. Information flow for content correlation.

tion as well as anomalies in data sets [26]. As patterns are formed from several co-occuring terms, pattern selection depends on the process of term selection. In information retrieval, terms are selected based on their ability to differentiate documents. Typically, the indexing terms are selected based on their frequency of use in the corpus. Intuitively, words which are used frequently (or infrequently) do not differentiate documents well, as they appear in a large (or small) subset of the corpus. Experimental results have shown that neither high nor low frequency words work well for indexing documents [27]. The appendix of this paper provides details for focusing text pattern extraction to those patterns with higher semantical significance.

For medical imaging, semi-automatic segmentation methods can be used to extract image features [13], [14]. Lesions characteristics are also defined based on algorithms developed to highlight canonical parameters such as T1, T2, and proton density (ρ) [28]. The lesion characteristics and features such as size, location, and shape can be computed and automatically clustered into TAHs for content-based matching [12], [29].

V. User Model Guided Information Retrieval

User models contain a representation of characteristic information about the user as well as a description of the user's query intention and goals. These models help interpret the content of a user query and effectively customize results by guiding the query facilities in deriving the answer.

Previous work on user modeling typically follows one of three user classification methods: *Non-domain specific* which characterize users by the extent of their knowledge [30], *Domain specific* which use stereotypes to describe general groups of users [31], [32], and *Multi-dimensional approaches* that combine non-domain and domain specific techniques [33]. However, these approaches do not fully exploit domain specific knowledge, and are often limited to a single application domain and/or a small set of user classes [32], [34], [35]. We remedy these problems by constructing such domain specific knowledge. Thus in addition to a user model, a constraint model is added to customize the information retrieval process. The *constraint selection module* consists of a set of pre-defined rules that select a set of attributes based on the user profile to add additional constraints to the original user query. The constraint rules are different for each query function (i.e. service).

In the following, we discuss the use of the user model in different query functions:

Resource selection and filtering: The data sources for answering queries are selected on the basis of the user's profile. As data sources are added to the intelligent resource registry and scenario proxies, additional attributes are used to classify the data sources, such as the goal of the data source (e.g. educational, reference, etc.), the source's expertise level (novice, expert), and the source's target audience (children, adults). The constraint selection module specifies what set of user attributes (e.g. age, gender, education level, race) are used to choose the data sources. Constraints for resource selection are thus generated based on the mapping of the user model and data source attributes.

Content correlation: To provide relevant information to the user, additional constraints are added to the query prior to content correlation. The constraint selection module maps a user model to a list of attribute fields that specify which attributes and values are used as constraints (e.g. for images, correlation of tumor intensity values vs. tumor location). Secondly, the user model stores information on how the data types should be correlated (e.g. intensity vs. size for images); this information is supplied by a domain expert to the user model.

Approximate matching: The approximate matching process references the user model to guide relaxation. A set of relaxation operators such as *relaxation level*, *relaxation order*, *non-relaxable* attributes, and *use TAH* can be specified in the user model or query to control the relaxation of the query objects/conditions (e.g. relax intensity before location, use a TAH based on intensity, size, and location). If no approximate matching methodology is specified by the user in a query, then the system uses the defaults defined within the user model [36].

Result ranking: Query results are ranked using a nearness measure, such as weighted relaxation error for approximate matching, appropriateness of resource, user preferences (e.g. diagrams are preferred over text), etc.

Visualization: The interface to the digital library takes into account the user to determine an effective means of data presentation. Default user interfaces are stored in the user model, but are easily adapted and customized at the individual modeling level.

A. Using the library collection

There are two modes for using the digital library:

Querying Mode: In this mode the user will seek information about a specific condition or disease, a treatment protocol, outcome, process of care, or costs.

Browsing Mode: This mode functions very similar to reviewing a teaching file or a traditional library. When the user is not seeking specific information related to a condition or disease, but instead is interested in reviewing what is contained in the library, this mode is invoked. The user will be provided with a table of contents based on ICD [37] and CPT [38] codes and can then choose to review further specific areas of interest.

VI. Conclusion

A new medical digital library for providing scenario- and user-tailored information retrieval is proposed in this paper. The library infrastructure consists of several innovative technologies:

1. Scenario-based proxies for focused information searches and navigation;
2. Facilities for knowledge-based approximate content (e.g. image features and textual patterns) navigation and resource matching. The knowledge bases are automatically generated from data sources and are thus scalable to large-scaled multimedia information systems;
3. Content correlation mechanism for linking relevant documents via key patterns. The patterns may be obtained by extracting frequently occurring terms in the textual documents. The key patterns can then be derived from the pattern sets via certainty measures of the clustered documents; and
4. Retrieval of user-tailored information via constraints and requirements specified in the user model.

These technologies provide the system with the capability to retrieve scenario-specific and user-tailored information. The library can be used in query mode as well as in browse mode. We are currently working jointly with the UCLA Medical School in implementing such a medical digital library, which will play an integral part of the proposed "paper-less" hospital.

Appendix

Algorithms to Derive Key Patterns

A collection of documents is defined as a set $D = \{d_1, ..., d_n\}$ with size $|D|$, where d_i is an individual document. A term or a pattern p which appears in the collection D can be considered as a random variable. The probability of p, denoted by $h(p)$, can be estimated from the ?? frequency in the collection. Let $D(p)$ be the set of documents that contains the pattern p and $|D(p)|$ be the size of $D(p)$ then

$$h(p) = \frac{|D(p)|}{|D|} \tag{1}$$

Pattern t	$p(t)$	Entropy E
lobe	150/178	0.6278
lobe right	111/178	0.9555
lobe left	86/178	0.9992
lobe right upper	82/178	0.9955
lobe left upper	63/178	0.9375
lobe mass upper	71/178	0.9703
lobe mass upper right	35/178	0.7151
lobe mass upper left	36/178	0.7264

TABLE I

Frequency & entropy of patterns in a thoracic radiology collection

The entropy of a random variable measures the amount of information provided by that variable [39]. Thus, the entropy of a pattern p can be used to measure how well p differentiates documents. The entropy of a pattern p with probability $h(p)$ in a collection of documents is:

$$E[h(p)] = -h(p)\log_2 h(p) - (1 - h(p))\log_2(1 - h(p)) \tag{2}$$

A collection of 178 thoracic radiology reports was used to calculated the entropy of terms and patterns found in the collection [40]. Table I shows the probability and entropy of a selection of terms and patterns taken from a collection of thoracic radiology reports. The data shows how adding terms to a frequent pattern (*lobe* $\rightarrow$ *lobe right*) reduces the frequency, while increasing the entropy of the pattern. The data also show that adding terms to reletively infrequent patterns (e.g., *lobe mass upper* $\rightarrow$ *lobe mass upper right*) only further reduces the pattern's frequency and entropy, thus decreasing the pattern's indexing power.

From Equation 2, we know $E[h(p)]$ is maximum at $h(p) = 0.5$, that is patterns that occur in 50% of the documents in the collection, have the highest entropy. Those patterns that occur most frequently as well as very seldomly in the documents have a low entropy. Patterns with high entropy differentiate documents better than those patterns with low entropy. Thus, the entropy value of a pattern may indicate its indexing capability within the document collection.

Entropy returns similar values for both frequent and infrequent probabilities. We will assume that infrequent patterns are in general not useful for classifying the content of the documents in a cluster. This motivates us to define a metric, *positive certainty* $C[h(p)]$, related to entropy which scores infrequent patterns negatively and frequent patterns positively:

$$\begin{aligned} C[h(p)] &= \quad 1 - E[h(p)], & \mathit{if}\, h(p) \geq 0.5 \\ C[h(p)] &= -(1 - E[h(p)]), & \mathit{if}\, h(p) < 0.5 \end{aligned} \tag{3}$$

Figure 8 shows a graph of positive entropy versus the probability of a binary random variable. The positive certainty measure indicates that only frequent patterns yield a large and positive C values.

To support content correlation, we are interested in extracting patterns which capture the content of documents

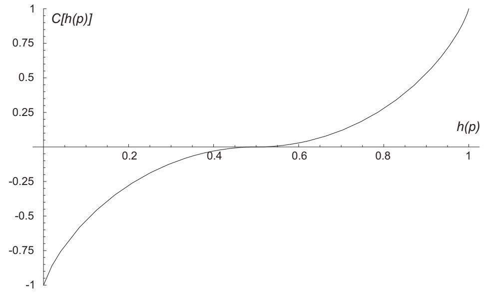


Fig. 8. Positive certainty (C)

in the collection. To focus the search for patterns, we partition the collection into groups of documents related by their content.

Documents can be grouped or clustered based on their similarity using a variety of methods [41]. The *cluster hypothesis* [?] states that a collection of documents can be grouped using a similarity measure and that those documents within the cluster are more likely to be related by content (i.e., relevant to the same query and user) than documents found outside of the cluster. Experimental systems have been developed which used document clustering to expand retrieval results to other relevant documents [42], [43], [44]. Evaluation and user testing have supported the conclusions of the cluster hypothesis [45]. Based on these results, patterns that co-occur in documents within the cluster may be more meaningful (i.e., related to the content of the cluster) than those patterns derived from the entire collection. Further, partitioning the document set into smaller groups, we can more efficently search for patterns which hold across several *related* document sets.

Let the *key patterns* of a cluster be those patterns which occur frequently in the cluster and do not occur as frequently outside of the cluster. Thus, *key patterns* can be used to describe the content of the cluster.

We define two metrics to guide the process of extracting key patterns from a cluster of documents. The first is used to measure the *exclusivity* or isolation of a pattern to a particular cluster. The second metric measures the *support* for a pattern in a given cluster. The two metrics can be used to extract those patterns that best capture the content of the cluster.

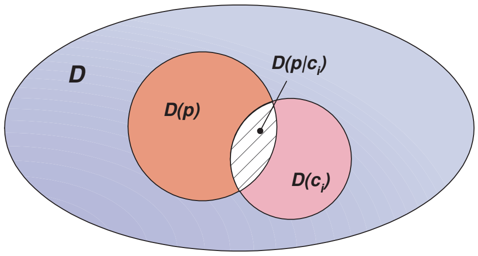


Fig. 9. Sub-sets of the document collection

Given a cluster c_i and a pattern p we define the following sub-sets of the document collection, as seen in Figure 9:

$D(p) =$ the set of documents in the collection with pattern p
$D(c_i) =$ the set of documents in the cluster c_i
$D(p|c_i) =$ the set of documents in c_i with pattern p

Let the *global cluster certainty* of pattern p in the cluster c_i be $C_g(p|c_i)$. The global cluster certainty is the positive certainty C_p of the percentage of the document set $D(p)$ which are members of the cluster c_i.

$$C_g(p|c_i) = C_p(h_g(p|c_i)) \quad (4)$$
$$where\ h_g(t|c_i) = \frac{|D(p|c_i)|}{|D(p)|}$$

$C_g(p|c_i)$ measures the exclusivity of the pattern p to the cluster. A pattern with a very high C_g (near 1) implies it is used almost exclusively in the cluster. A pattern with a C_g near 0 means the pattern is used just as frequently outside of the cluster as that of inside the cluster. A pattern with low C_g (near -1) is used infrequently within the cluster compared to its use outside of the cluster.

Let the *local cluster certainty* $C_l(p|c_i)$ be the positive certainty of the percentage of documents in the cluster c_i that contain the pattern p.

$$C_l(p|c_i) = C_p(h_l(p|c_i)) \quad (5)$$
$$where\ h_l(p|c_i) = \frac{|D(p|c_i)|}{|D(c_i)|}$$

C_l measures how frequent a pattern appears within documents in the cluster $D(c_i)$. A pattern with high C_l (near 1) means it is used in most of the documents in the given cluster c_i. A pattern with C_l near 0 implies the pattern appears in documents with equal probability: that is, nearly half of the documents use the pattern and the other half do not. A pattern with low C_l (near -1) implies the pattern t is used infrequently in the document cluster $D(c_i)$.

It should be noted that neither high C_g or high C_l alone is capable to derive a key pattern. C_g measures the pattern's exclusivity to the cluster, a low C_g implies that the pattern occurs frequently outside of the cluster, thus does not capture content specific to the cluster. C_l measures the pattern's distribution within the cluster, a low C_l implies that the pattern does not occur in a majority of the documents of the cluster, and does not represent content that permeates the cluster.

C_g and C_l can be used together to extract patterns from text data. For example, patterns found in a cluster that have high C_l (i.e., used frequently in the cluster) and low C_g (i.e., appear frequently outside of the cluster) may be good candidates for adding additional terms to make the pattern more specific to the cluster, thus providing more indexing power.

We have developed an algorithm to automatically derive key patterns from clusters found in the collection. Figure 10 is a flow chart of the algorithm. The algorithm starts with a single term pattern. Terms are added to the pattern one at a time using a greedy algorithm. The key pattern is complete when adding any new term will decrease the sum of the C_g and C_l for the pattern. The algorithm maximizes the key pattern's length while insuring the sum of C_g and C_l does not decrease.

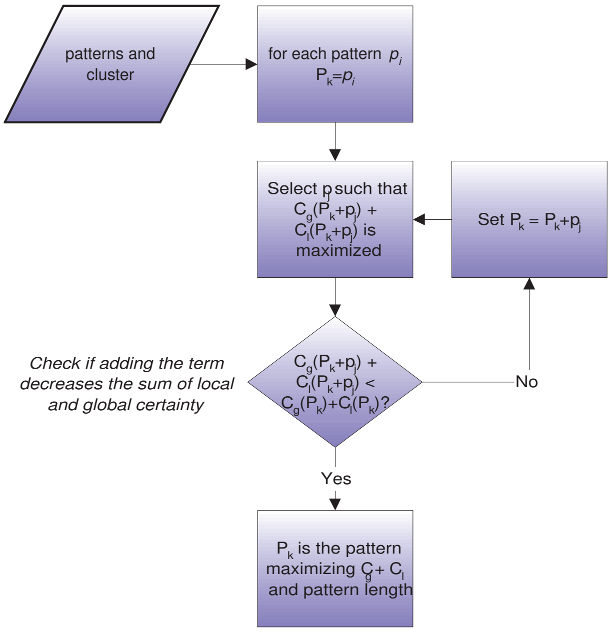


Fig. 10. Algorithm to extract patterns from text documents

Acknowledgements

The authors would like to thank John David Dionisio and Alex Bui of UCLA for their stimulating conversations and comments during the development of this research.

References

[1] T. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," *International Journal of Human-Computer Studies*, vol. 43, no. 5-6, pp. 907–28, 1995.

[2] U. Hahn, S. Schulz, and M. Romacker, "Part-whole reasoning: a case study in medical ontology engineering," *IEEE Intelligent Systems*, vol. 14, no. 5, pp. 59–67, 1999.

[3] E. Selberg and O. Etzioni, "The metacrawler architecture for resource aggregation on the web," *IEEE Expert*, 1997.

[4] K. Obraczka, P. B. Danzig, and S. H. Li, "Internet resource discovery services," *Computer*, vol. 26, no. 9, pp. 8–24, September 1993.

[5] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *7th International World Wide Web Conference*, Brisbane, Qld., Australia, 1998.

[6] Michael F. Schwartz, "Internet resource discovery at the University of Colorado," *Computer*, vol. 26, no. 9, pp. 25–36, September 1993.

[7] Y. Lashkari, M. Mentral, and P. Maes, "Collaborative interface agents," in *Twelfth National Conference on Artificial Intelligence*, 1994.

[8] B.A. Nardi, Miller J.R., and D.J. Wright, "Collaborative, programmable intelligent agents," *Communications of the ACM*, vol. 41, no. 3, 1998.

[9] D. Mladenic, "Text-learning and related intelligent agents: a survey," *IEEE Intelligent Systems*, vol. 14, no. 4, pp. 44–54, 1999.

[10] W. W. Chu, H. Yang, K. Chiang, M. Minock, G. C., and C. Larson, "CoBase: A scalable and extensible cooperative information system," *Journal of Intelligent Information Systems*, vol. 6, no. 11, 1996.

[11] W.W. Chu and K. Chiang, "Abstraction of high level concepts from numerical values in databases," in *Proceedings of the AAAI Workshop on Knowledge Discovery in Databases*, July 1994.

[12] W. W. Chu, K. Chiang, C. C. Hsu, and H. Yau, "An error-based conceptual clustering method for providing approximate query answers," *Communications of ACM, Virtual Extension Edition (http://www.acm.org/cacm/extension)*, vol. 39, no. 12, pp. 216–230, Dec. 1996.

[13] J.S. Jin, "Xra image segmentation using regression," in *SPIE's International Symposium on Medical Imaging*, 1996, vol. 2710, pp. 864–868.

[14] M. Brown, M. McNitt-Gray, N. Manknovich, J. Goldin, J. Hiller, L. Wilson, and D. Aberle, "Method for segmenting chest CT image data using an anatomical model: preliminary results," *IEEE Transactions on Medical Imaging*, vol. 16, no. 6, pp. 828–838, 1997.

[15] C.C. Hsu, W. W. Chu, and R. K. Taira, "A knowledge-based approach for retrieving images by content," *IEEE Transaction on Knowledge and Data Engineering*, Aug. 1996.

[16] M. Merzbacher and W. W. Chu, "Pattern-based clustering for database attribute values," in *Proceedings of AAAI Workshop on Knowledge Discovery*, 1993.

[17] F. Cuppens and R. Demolombe, "How to recognize interesting topics to provide cooperative answering," *Information Systems*, vol. 14, no. 2, pp. 163–173, 1989.

[18] A.S. Hemerly, M.A. Casanova, and A.L. Furtado, "Cooperative behaviour through request modification," in *Proc of the Tenth International Conference of the Entity-Relationship Approach*, San Mateo, CA, USA, 1991.

[19] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Machine Learning*, vol. 2, no. 2, pp. 139–172, 1987.

[20] J. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," *Artificial Intelligence*, vol. 40, pp. 11–62, 1989.

[21] S. J. Hanson and M. Bauer, "Machine learning, clustering and polymorphy," in *Uncertainty in Artificial Intelligence*, L. N. Kanal and J. F. Lemmer, Eds. Elsevier Science Publishers B. V., North-Holland, 1986.

[22] D.B. Johnson, R.K. Taira, A.F. Cardenas, and D.R. Aberle, "Extracting information from free text radiology reports," *International J. on Digital Libraries*, vol. 1, pp. 297–308, 1997.

[23] W. W. Chu, A. F. Cardenas, and R. K. Taira, "KMeD: A Knowledge-Based Multimedia Medical Distributed Database System," *Information Systems*, vol. 20, no. 2, pp. 75–96, 1995.

[24] W. W. Chu and G. Zhang, "Associative query answering via query feature similarity," in *Proceedings International Conference on Intelligent Information Systems*. Nov. 1997, IEEE Press.

[25] C. Clifton and R. Steinheiser, "Data mining on text," in *Proceedings. The Twenty-Second Annual International Computer Software and Applications Conference (Compsac '98)*. 1998, IEEE Comput. Soc.

[26] J.A. Goldman, W.W. Chu, D.S. Parker, and R.M. Goldman, "TDDA, a Data Mining Tool for Text Databases: A Case History in a Lung Cancer Text Database," in *1st International Conference on Discovery Science*, 1998.

[27] G. Salton and M.J. McGill, *Introduction to modern information retrieval*, McGraw-Hill, 1983.

[28] S Sinha, U Sinha, H Kangarloo, and HK Huang, "Magnetic resonance image synthesis from analytic solutions of spin-echo and radio frequency-spoiled gradient-echo images," *Investigative Radiology*, vol. 27, pp. 856–864, 1992.

[29] W. W. Chu, C. C. Hsu, A. F. Cardenas, and R. K. Taira, "Knowledge-based image retrieval with spatial and temporal constructs," *IEEE Transactions on Knowledge and Data Engineering*, 1998.

[30] K. Waldhor and H. Anschutz, "User centered explanations in knowledge based systems," in *Human-Computer Interaction INTERACTION '90: Proceedings of the IFIPTC13 Third International Conference*, Cambridge, UK, 1990.

[31] A. Ballim and Y. Wilks, "Beliefs, stereotypes and dynamic agent modeling," *User Modeling and User-Adapted Interaction*, vol. 1, pp. 33–65, 1991.

[32] A.J. Kok, "A review and synthesis of user modeling in intelligent systems," *The Knowledge Engineering Review*, vol. 1, no. 1, pp. 21–47, 1991.

[33] X. Huang, G.I. McCalla, J.E. Greer, and E. Neufield, "Revising deductive knowledge and stereotypical knowledge in a student model," *User Modeling and User-Adapted Interaction*, vol. 1, pp. 87–115, 1991.

[34] J.A. Bateman and C.L. Paris, "Phrasing a text in terms the user can understand," in *Proceedings of the Eleventh International*

Joint Conference on Artificial Intelligence, Detroit, Michican, USA, 1989.

[35] E. Shifroni and B. Shanon, "Interactive user modeling: An integrative explicit-implicit approach," *User Modeling and User-Adapted Interaction*, vol. 2, pp. 331–365, 1992.

[36] W.W. Chu, C.C. Hsu, I.T. Ieong, and R.K. Taira, "Content-based image retrieval using metadata and relaxation techniques," in *Managing Multimedia Data*, A Sheth and W Klas, Eds. McGraw Hill, 1998.

[37] *The International Classification of Diseases, 9th revision*, McGraw-Hill, New York, 1996.

[38] *CPT '98: Physicians' Current Procedural Terminology*, American Medical Association, Chicago, IL, 1998.

[39] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1991.

[40] D.B. Johnson and W.W. Chu, "Domain specific document retrieval using n-word combination index terms," in *In Proceedings: Fusion '99*, 1999.

[41] P. Willett, "Recent trends in hierarchic document clustering: a critical review," *Information Processing & Management*, vol. 24, no. 5, 1988.

[42] P.G. Anick and S. Vaithyanathan, "Exploiting clustering and phrases for context-based information retrieval," in *Proceedings of the Ann Int'l SIGIR Conference*, Philadelphia PA, 1997.

[43] D.R. Cutting, D.R. Karger, J.O. Pedersen, and J.W. Tukey, "Scatter/gather: A cluster-based approach to browsing large document collections," in *Proceedings of the 15th Ann Int'l SIGIR*, Denmark, 1992.

[44] H. Schütze and C. Silverstein, "Projections for efficient document clustering," in *Proceedings of the Ann Int'l SIGIR Conference*, Philadelphia PA, 1997.

[45] M.A. Hearst and J.O. Pedersen, "Reexamining the cluster hypothesis: Scatter/gather on retrieval results," in *Proceedings of the 19th Annual International SIGIR Conference*, Zurich, Switzerland, 1996.