Hierarchical Concept Indexing of Full-text Documents in the UMLS® Information Sources Map[1]

Lawrence W. Wright[2]
Holly K. Grossetta Nardini[3]
Alan R. Aronson[4]
Thomas C. Rindflesch[5]

1. This research was supported in part by an appointment of Holly K. Grossetta Nardini to the NLM Associate Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education.

2. National Library of Medicine (under an Intergovernmental Personnel Agreement with Yale University). Now at International Cancer Information Center, National Cancer Institute, Bethesda, MD 20892. Tel: (301) 496 9096, Email: lwright@exchange.nih.gov [to receive correspondence and proofs]

3. Johns Hopkins University SAIS Bologna Center, via Belmeloro, 11, Bologna 40126, Italy. Tel: (39-51) 232 185, Fax: (39-51) 288 505, Email: holly@jhubc.it

4. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894. Tel: (301) 435 3162, Fax: (301) 496 0673, Email: alan@nlm.nih.gov

5. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894. Tel: (301) 435 3191, Fax: (301) 496 0673, Email: tcr@nlm.nih.gov

ABSTRACT: Full-text documents are a vital and rapidly growing part of online biomedical information. A single large document can contain as much information as a small database, but normally lacks the tight structure and consistent indexing of a database. Retrieval systems will often miss highly relevant parts of a document if the document as a whole appears irrelevant. Access to full-text information is further complicated by the need to search separately many disparate information resources. This research explores how these problems can be addressed by the combined use of two techniques: (1) natural language processing for automatic concept-based indexing of full text, and (2) methods for exploiting the structure and hierarchy of full-text documents. We describe methods for applying these techniques to a large collection of full-text documents drawn from the Health Services/Technology Assessment Text (HSTAT) database at the National Library of Medicine (NLM), and examine how this hierarchical concept indexing can assist both document- and source-level retrieval in the context of NLM's Information Sources Map project.

INTRODUCTION

When seeking computer-accessible information, biomedical researchers have traditionally faced several problems. Biomedical information is scattered among many information sources, each of which has to be searched separately, often using different search techniques. Furthermore, these sources are organized and indexed differently, and often use widely divergent medical vocabularies.

An added challenge is the recent availability of full text, as opposed to abstracts, in the online biomedical literature (Sievert, McKinin, & Johnson, 1995; Sievert, 1996). Full-text information can be critical to biomedical researchers, who may need detailed information for time-sensitive tasks (McKinin et al., 1991). Yet, as more full-text documents are made available electronically, it is increasingly difficult to retrieve specific information efficiently and accurately.

Information retrieval research is increasingly concerned with the issues associated with multiple, disparate information sources (e.g. Callan, Lu, & Croft, 1995; Cousins et al., 1997; Voorhees & Tong, 1997; Buckland & Plaunt, 1997) as well as full-text documents (e.g. Salton, Allan, & Buckley, 1993; Callan, 1994).

The Information Sources Map (ISM) (Masys & Humphreys, 1992), which is one component of the Unified Medical Language System® (UMLS) being developed by the National Library of Medicine (NLM), is particularly relevant to these concerns. The ISM is a database describing network-accessible biomedical information sources. It is intended to support systems which can identify and connect to sources likely to be relevant to a user's information need and, when possible, retrieve particular documents, or sections of large documents, that satisfy that need.

Research associated with the ISM (Miller et al., 1995; Rodgers, 1995) is ongoing, and many of the methodologies supporting the full realization of its potential are still under development. The work reported here explores how two techniques, used in tandem, can improve source and document selection in the context of the ISM: (1) natural language processing for automatic con-

cept-based indexing of full text, together with (2) methods for exploiting the structure and hierarchy of full-text documents.

BACKGROUND

The UMLS Information Sources Map

The ISM is one of four "knowledge sources" that form the core of the UMLS, a building block for integrated systems that help health and information professionals link varying information sources, including computer-based patient records, factual and bibliographic databases, and expert systems (Lindberg, Humphreys, & McCray, 1993; Humphreys et al., 1998). The UMLS is a fundamental tool for surmounting retrieval problems caused by varying medical terminology and by the diffusion of medical knowledge into many databases. Each of the knowledge sources contributed to this project, but we focused on identifying potential improvements to the ISM.

The first knowledge source, the Metathesaurus®, matches different names for the same concept found in nearly 40 biomedical vocabularies and classifications, including NLM's Medical Subject Headings (MeSH®), Systematized Nomenclature of Medicine (SNOMED), Diagnostic and Statistical Manual of Mental Disorders (DSM), and the International Classification of Diseases (ICD). For example, the Metathesaurus stipulates that "Malignant Tumor" is synonymous to the preferred term "Cancer."

The second source is the SPECIALIST™ Lexicon, which provides part of speech and other syntactic information for English words and phrases. From the Lexicon, one can determine that *cancer* is a noun and has regular variants.

Third, the Semantic Network contains information about the categories to which all Metathesaurus concepts have been assigned (e.g. 'Neoplastic Process,' 'Hazardous or Poisonous Substance,' 'Pathologic Function') and the permissible relationships between these types (e.g., 'Neoplastic Process' can be 'CAUSED BY' 'Hazardous or Poisonous Substance').

The Information Sources Map, the fourth UMLS knowledge source, is a database containing expert descriptions of various biomedical databases and other information sources. As of January 1998, it covers 49 major biomedical sources, 34 of which are produced partly or wholly by NLM. The types of sources are quite varied, including databases of bibliographic, full-text, factual, image and directory information, as well as some expert system and other resources (see full list with brief descriptions in the Appendix). Each source is described in terms of subject coverage (using both MeSH terms and UMLS semantic types), the likelihood of usefulness for such purposes as teaching and patient care, the types of content covered (e.g. journal articles) and provided (e.g. abstracts), as well as technical characteristics and access methods.

Research associated with the ISM is exploring ways to use and improve this database to automate information retrieval from multiple network-accessible sources. The goal is to connect users with sources relevant to a particular query and, where possible, to retrieve specific contents from those sources. In the current experimental ISM prototype, a user's natural language request for information on *breast cancer* and *tamoxifen*, for example, is processed to yield the MeSH terms "Breast Neoplasms" and "Tamoxifen," which are then compared with the MeSH terms assigned to each information source to determine that MEDLINE®, CANCERLIT® and HSTAT, among others, are likely to be useful. For most sources in the database, the system can then translate the user's request into search statements adjusted for each source, and retrieve matching documents from each.

Efforts to match user-specified queries directly to the high-level expert source descriptions have achieved some success. However, matching of this type has yielded acceptable recall of the most relevant sources only at considerable expense to precision, and selects many additional sources unlikely to be useful. Beyond issues of precision and recall, there have been serious concerns about problems maintaining such descriptions and extending them to a potentially much wider range of information sources. If source-level selection is to become an effective tool for distributed information retrieval in biomedicine, new methods will be required for subject indexing of sources and their contents (the central concern here), as well as for query expansion and matching queries to sources (which are being addressed in related research). Over and above effective retrieval, serious issues also arise in integrating and presenting the results of multi-source searches (Masys, 1992; Miller et al., 1995; Rodgers, 1995; Humphreys et al., 1998; Voorhees, Gupta, & Johnson-Laird, 1995; Cousins et al., 1997).

Full-text Documents

As noted above, the ISM covers several full-text databases. Such collections present special challenges to information retrieval systems, which have traditionally been designed for much shorter, tightly structured and carefully worded documents, such as abstracts. Full text, on the other hand, can present overwhelming amounts of information, spanning a range of topics, with abundant explicit and implicit internal structure which plays a key role in interpreting and navigating it effectively.

Several approaches are emerging to accommodate full-text retrieval: Sievert, McKinin, & Johnson (1995) suggest several heuristics to aid the searcher, while Purcell and Mar (1992) propose an expert system to help focus retrieval. Further support for dealing with the amount of information found in a large document can be obtained by using the structural organization of the document itself to identify topics and subtopics (Salton, Allan, & Buckley, 1993; Fuller et al., 1993).

Topic and subtopic are dependent on the logical organization of the content of a document, which correlates with the division of the document into sections and subsections. If the structure of a large document, including sections and subsections, can be identified, the indexing method of an information retrieval system can take advantage of this structure in order to provide specific sections, rather than the entire document, in response to a user's query.

Passage-based systems account for this structure and can improve search results for full-text databases. Passages in these systems can be based on arbitrary, fixed lengths (Kaszkiel & Zobel, 1997; Callan, 1994), can be determined automatically (Hearst & Plaunt, 1993), or can be based on explicit sectioning. Systems employing explicitly marked sections may either refer to Standard Generalized Markup Language (SGML) markup (Fuller et al., 1993; Wilkinson, 1994) or author-supplied headings and titles (Salton, Allan, & Buckley, 1993). Passage-based systems may retrieve passages only, or entire documents, based on how well particular sections within a document match the user's query. Some systems combine these two approaches and return either a section or an entire document, depending on which best satisfies the information need expressed by the query.

Indexing

Traditionally, indexing research has concentrated on representing individual documents or similar "retrievable" content items. Automatic methods use either words, phrases, or concepts as terms describing document content (see, for example, Evans et al., 1991; Wagner, 1991; Smeaton,

1992). These techniques can be combined with the passage-based approaches described above for accommodating full text retrieval.

More recently, research has been directed at representing the contents of entire databases of documents and other online information resources. This source-level indexing is prompted both by the explosive growth and diversity of such resources, and by the high cost and frequent impossibility of directly indexing their contents. Some source-level indexing efforts use full word-based indexes for each source (e.g. Callan, Lu, & Croft, 1995; Gravano, García-Molina, & Tomasic, 1994). Others, generally seeking more compact and stable methods for indexing highly diverse sources for which full, word-based indexes are often unavailable, have explored higher-level indexing methods including free-text and controlled-vocabulary metadata schemes (cf. Baldonado, Chang, & Gravano, 1997), semantic representations (Chakravarthy & Haase, 1995), and query-based indexing with training sets (Voorhees & Tong, 1997).

METHODS

In order to portray the content of both source databases and individual documents and document sections, we employed natural language processing software being developed at NLM for automatic concept-based indexing. After selecting a test database, we devised a method for using the explicit information regarding document sections. We then combined the information regarding document structure with our automatic, concept-based indexing to provide a hierarchical index exploitable during the information retrieval process. This hierarchical index was then used for exploratory testing of both source and document retrieval.

A Test Database

Health Services/Technology Assessment Text (HSTAT), a full-text electronic resource developed and maintained by NLM, was selected for testing our methodology. The number, format, and size of the documents in this database are well-suited to the tools being developed, while its topical diversity makes it an especially interesting case for enhanced source selection. HSTAT provides an array of health services information of interest to both professionals and consumers, including clinical practice guidelines, consensus development conference reports, health technology assessment reports, and treatment improvement protocols.

In order to reduce the number of documents processed for this study, a sampling plan was developed based on a single theme that occurs throughout HSTAT: breast cancer. A text-word search was issued using *breast cancer* and variants *breast malignancy, breast tumor, carcinoma of the breast,* and *breast neoplasms*. The resulting set of four HSTAT files, containing 66 distinct documents, was diverse in terms of structure and purpose, but was focused in content (and vocabulary). These were: (1) the AHCPR Clinical Practice Guideline entitled *Quality Determinants of Mammography,* (2) the NIH Consensus Development Conference Reports, (3) the AHCPR Health Technology Assessments, and (4) the Guide to Clinical Preventive Services provided by the U.S. Preventive Services Task Force. These four files collectively comprise almost 50% of the entire HSTAT database.

Sectioning Documents Based on SGML

Documents in HSTAT are coded in Standard Generalized Markup Language (SGML) (Prettyman, 1997), a system for encoding structural information in electronic documents. SGML indicates a document's structure by explicitly tagging components such as chapter, section, and

subsection. This characteristic is exploitable in addressing the challenges associated with document selection in full-text information retrieval (Fuller et al., 1993). The following schematic example, drawn from the AHCPR Clinical Practice Guideline entitled *Quality Determinants of Mammography,* illustrates the embedding of two sections (delimited by <SEC> and </SEC>) within a chapter (<CHP> and </CHP>). The chapter occurs as part of the complete document, which is bounded by the <BODY> and </BODY> tags.

<BODY>
[contents through end of chapter 6]
<CHP>
<h2>7. Referring Providers</h2>
[initial sections]
<SEC>
<h3>Adverse Consequences of Mammography</h3>
<p>Excessive biopsies are a possible adverse consequence...[text of this section]
</SEC>
<SEC>
<h3>Mammography Standards</h3>
<p>A 1990 study of mammography sites...[text of this section]
</SEC>
[final section]
</CHP>
[remaining chapters etc.]
</BODY>

In order to take advantage of SGML coding of document structure, we developed a software tool called SGML-Extractor. This program translates the original document into a series of document fragments which correspond to the hierarchical organization of the source document. For example, from the parts of the SGML-encoded document shown above, the program derives the following three document fragments. The first is the chapter entitled “Referring Providers,” while the second two fragments are two of the constituent sections of this chapter: “Adverse Consequences of Mammography” and “Mammography Standards.”

UI - 2.10:mamc.body.chp
TI - 7. Referring Providers
TX - Adverse Consequences of Mammography.
Excessive biopsies are a possible adverse consequence...[text of this section]
Mammography Standards
A 1990 study of mammography sites...[text of this section]

UI - 2.10.11:mamc.body.chp.sec
TI - Adverse Consequences of Mammography
TX - Excessive biopsies are a possible adverse consequence...[text of this section]

UI - 2.10.12:mamc.body.chp.sec
TI - Mammography Standards

TX - A 1990 study of mammography sites...[text of this section]

Each document fragment includes a unique identifier which encodes the name of the original document ("mamc" above for *Quality Determinants of Mammography*) and the hierarchical position of the fragment in the source document. "body.chp," for example, indicates that this fragment (chp) is an immediate component of the document as a whole (body). Both the title and text of each fragment are then given.

This representation of a large document provides an explicit and uniform method for referring to the text from any level of structural organization of a document, whether this is the entire document, a particular chapter, or just a section in that chapter. Before considering how we take advantage of this information to address the problems associated with document selection in full-text information retrieval, we discuss automatic concept-based indexing, as applied to any segment of medical text.

Automatic Concept-based Indexing

Our approach to automatic indexing is based on a natural language processing tool, called MetaMap (Aronson, Rindflesch, & Browne, 1994), which matches medical text to concepts in the UMLS Metathesaurus. For example, MetaMap determines that the text *pre-operative chemotherapy for stage IIIb breast cancer, followed by mastectomy* maps to the following Metathesaurus concepts: "Operative Procedures," "Drug Therapy," "stage IIIB breast cancer," and "Mastectomy."

Although some words (*mastectomy,* for example) map to very similar (or identical) Metathesaurus concepts, in general the concepts provided by MetaMap have a distinct advantage over text words for automatic indexing. Concepts which can be expressed in several synonymous ways, such as "Chemotherapy" and "Drug Therapy," are normalized to only one expression ("Drug Therapy"), while phrases can be treated as single concepts ("stage IIIB breast cancer," for example).

In processing text, MetaMap proceeds as follows: After delimiting sentences and other significant linguistic units in the input, the program extracts noun phrases and determines the Metathesaurus concept or concepts which best express the meaning of each phrase, taking into account morphological variation and synonymy. MetaMap uses the synonymy relations encoded in the Metathesaurus and draws on additional UMLS resources, including the SPECIALIST Lexicon and associated lexical programs (McCray, Srinivasan, & Browne, 1994) as well as a supplemental synonym knowledge base drawn from *Dorland's Illustrated Medical Dictionary.*

The Metathesaurus concepts culled by MetaMap serve as the basis for a program called MMI (MetaMap Indexing), which provides automatic concept-based indexing for medical text. MMI indexes a document by ranking the Metathesaurus concepts found by MetaMap. The ranking of concepts reflects their relative importance as an indication of what the document is about. Concepts which occur in the title are ranked higher than those which occur in the text. More frequently occurring concepts are ranked higher than those which occur less frequently. The notion of specific (versus general) is also addressed by the ranking algorithm, in that more specific terms (as determined by MeSH tree depth) are ranked higher.

As an example, MMI is applied to one of the document fragments introduced above. In addition to the title and full text of the section, the ten top-ranked Metathesaurus concepts are given along with their MMI ranking scores.

UI - 2.10.11:mamc.body.chp.sec

TI - Adverse Consequences of Mammography
TX - Recommendation: Referring health care providers should be aware of the possible adverse consequences of mammography, the likelihood of each, and procedures to lower their likelihood. (C) Excessive biopsies are a possible adverse consequence of mammography. There is also a low probability of breast cancer induction due to radiation exposure. Other problems associated with mammography include inadequate communication of results, the need to return for additional or repeat views, the inconvenience of scheduling mammography, pain or discomfort, false reassurance, delay in diagnosis and treatment, and cost. Adverse consequences and other possible problems are discussed in Chapter 8. Recommendation: Be aware that the risk of breast cancer induction from annual screening mammography beginning at age 40 or 50 is negligible. The estimated risk of breast cancer induction increases in women who are younger at the time of exposure. (A)

MMI concepts for 2.10.11:mamc.body.chp.sec:
34.5 Mammography
18.9 Risk
8.2 Probability
7.6 Pain
6.6 Biopsy
5.5 Recommendations
5.0 Health Personnel
5.0 Communication
4.8 Breast Cancer
4.4 Women

The list of ranked terms provided by MMI is considerably less accurate in representing the content of a document than the indexing terms assigned by human indexers. However, these MMI terms appear to have a distinct advantage over automatic indexing based solely on the words occurring in a text. Although MMI can provide indexing terms for any medical text, we apply it to all of the document fragments produced by SGML-Extractor. The resulting combination of indexing terms and structural information can be combined to form a hierarchical index.

Hierarchical Indexing

Hierarchical indexing is a method of indexing large documents at several levels of structure, so that a retrieval system can pinpoint the most relevant sections within each document: the whole document, a chapter, or a specific section. As an example of hierarchical indexing consider the schematic chapter given above, which contains two sections. MMI is applied to each of these three document fragments, the chapter as a whole and each of the constituent sections, to yield the following hierarchical index (only the top five MMI terms are given) for the chapter:

UI - 2.10:mamc.body.chp
TI - 7 Referring Providers
TX - [text of chapter]
MMI - 34.1 Mammography
19.4 Breast

17.0 Health Personnel
16.0 Women
10.6 Physicians

UI - 2.10.11:mamc.body.chp.sec
TI - Adverse Consequences of Mammography
TX - [text of section]
MMI - 34.5 Mammography
18.9 Risk
8.2 Probability
7.6 Pain
6.6 Biopsy

UI - 2.10.12:mamc.body.chp.sec
TI - Mammography Standards
TX - [text of section]
MMI - 33.9 Mammography
25.6 Accreditation
18.1 Certification
14.8 United States
12.6 Accounting

This index illustrates the way that the MMI terms, along with the structural information provided, serve as the basis for determining which section of a large document best satisfies the user's information need. Consider the concept "Risk," which is most fully discussed in section 2.10.11. The prominence of "Risk" in this section is indicated by the relatively high MMI score. Although by virtue of occurring in this section, the concept also occurs in the chapter containing this section (2.10), "Risk" does not occur frequently enough throughout the entire chapter to be included in the top five MMI terms for the chapter.

In order to construct a framework for testing the viability of hierarchical indexing as a basis for improving source and document selection, we first subjected the four sample HSTAT files mentioned above to SGML-Extractor. All resulting document fragments were then indexed automatically, using the MMI techniques. The resulting indexed document fragments were comparable to those just illustrated for the schematic chapter entitled "Referring Providers."

DISCUSSION

The accuracy of the MMI-based automatic indexing was initially reviewed by an expert indexer at NLM. Although numerous inaccuracies were noted, the terms assigned by MMI were deemed adequate to serve as the basis for further testing. (Previous research had demonstrated the advantage of concept-based indexing over word-based approaches (Aronson, Rindflesch, & Browne, 1994).) Our evaluation then focused on the two aspects of ISM-related research: source selection and document selection. We first assessed the effectiveness of automatically generated indexing terms as a type of source-level subject coverage coding aimed at guiding source selection. Secondly, we used a statistically based information retrieval system to test the value of our

automatic hierarchical concept indexing for retrieving relevant documents and pinpointing the most relevant sections of large documents.

Source Selection

One of the central goals of this project was to examine how automatic concept-based indexing could help address known problems with the source subject coverage indexing that is currently provided in the ISM. In addition to obviating the need for expert review and coding, methods based on automatic processing of the actual contents of the database could provide a more accurate basis for determining its likely relevance to specific queries. While formal evaluation was ruled out by both the partial coverage of HSTAT's contents and the lack of a test collection of evaluated queries, two highly suggestive approaches were possible: to compare the characteristics of indexing terms generated by the expert-profile and automatic-concept methods, and to examine how the differences might affect source selection for a small set of test queries.

Current ISM indexing of subject coverage at the source level is done primarily through human assignment of relatively broad MeSH terms, signifying coverage of the given term and all terms under it within the MeSH tree hierarchy ("child terms"). MeSH subheading qualifiers are applied where appropriate, indicating restriction to a specific aspect such as "therapy." A newly-added "extent" code indicates whether coverage is judged comprehensive or substantial. The coding for HSTAT is shown in Table 1.

Table 1. **ISM Database MeSH Coding for HSTAT**

MeSH Term	MeSH Tree Code	Child Codes	Extent
Diseases	C	7,271	substantial
Mass Screening	E1.563	4	comprehensive
Therapeutics	E2	368	substantial
Equipment and Supplies	E7	195	substantial
Preventive Medicine	G2.403.790.548	4	comprehensive
Health Services	N2.421	176	substantial
Economics	N3.219	142	substantial
Health Services Research	N3.349.380	2	substantial
Technology Assessment, Biomedical	N3.880	1	comprehensive
Practice Guidelines	N4.761.700.350.650	0	comprehensive
Quality of Health Care	N5.715	163	comprehensive

Since we had subjected nearly half of HSTAT's contents to our MMI-based, automatic indexing technique, it was possible to use the resulting terms to build some exploratory source-level indexes for preliminary comparison and testing. We had sectioned and indexed 66 HSTAT documents, yielding separate lists of indexing terms for all "document fragments" at each level of granularity, ranging from whole documents to their smallest subsections. As a possible basis for source-level indexing, we collected all of the MMI term assignments which scored highly as potential indexing terms (MMI scores of 5.0 or above). To examine both the broadest and most specific levels of indexing, we tested two subsets of these assignments: at document level and lowest subsection level. In addition to terms from NLM's MeSH vocabulary, the UMLS Metathesaurus contains terms from nearly 40 other vocabularies; the MMI-generated terms in our exploratory source-level indexes reflect this diversity, despite a strong scoring bias towards using MeSH

terms. Table 2 provides a summary of the resulting source indexing terms, grouped by indexing level and vocabulary.

Table 2. **MMI Index Terms at HSTAT Source Level**

	MeSH Tree Codes		MeSH Terms		Non-MeSH Terms	
Indexing Level	**total**	**unique**	**total**	**unique**	**total**	**unique**
Document	2,988	1,161	1,484	619	330	141
Subsection	69,578	6,324	35,963	3,551	8,795	1,025

In general characteristics, the MMI indexes look attractive as an alternative compact representation of source subject coverage. The number of MeSH tree codes covered in the ISM database is 8,327 -- comparable to the count for subsection-level MMI indexing, and several times that for document-level MMI indexing. However, the actual composition of the MMI and ISM codes is radically different. The MMI MeSH codes are widely distributed and normally fairly specific, providing a detailed representation of subjects covered. At both document and subsection levels, over 60% of MMI-generated MeSH term assignments are not covered by the current ISM MeSH indexing of HSTAT, and would be missed in any simple matching of query terms to source descriptions. Further, the MMI-generated non-MeSH concepts represent a rich added source of indexing terms, potentially very useful in matching to queries but not readily mapped into the ISM's MeSH-based scheme.

A sample query evaluation illustrates these and other differences between the two indexing methods. In the query *adjuvant therapy of breast cancer and tamoxifen*, three concepts are identified by MMI: "Tamoxifen," "Drug Therapy," and "Breast Cancer." "Tamoxifen," a MeSH term, is not covered by the ISM MeSH index, but is covered by both document- and subsection-level MMI indexes. "Drug Therapy" is also a MeSH term, covered by "E2 Therapeutics" in the ISM and occurring directly, with high frequency, in both MMI-based indexes. "Breast Cancer" is found directly in the non-MeSH MMI indexes; as a MeSH non-print entry term for "Breast Neoplasms," it also would likely be identified as covered by "C Diseases."

In addition to missing at least one of the three sample query terms for which HSTAT is clearly relevant, the ISM index also produces many inappropriate matches where HSTAT is not useful. "Diseases," for example, would likely have matched to "Breast Cancer," but would equally match "Hodgkin's Disease" and "Bovine Spongiform Encephalopathy" (mad cow disease), neither of which is currently covered in HSTAT. In fact, only 1,540 out of the possible 7,272 MeSH disease tree codes were identified by MMI in the half of the database which was processed.

The use of a few broad categories, as seen in the existing, human-assigned, source-level indexes in the ISM, is necessarily hit-and-miss for matching to user queries. It seems likely that MMI-based source indexing of HSTAT could help dramatically improve recall and precision at the level of source selection, although confirmation of this must await further testing.

Document Selection

In addition to scrutinizing the value of MMI-based indexing for source selection, a further goal was to determine whether the insights provided by current research could be applied to the specific problems associated with directing users to the most relevant information satisfying their information need. This information could be an entire document, when that is appropriate, or a partial document when that can be determined to be more relevant to the query than the entire, larger document. In order to assess the effectiveness of our methods, we provided the indexed,

hierarchical document fragments for our four HSTAT test files to Inquery (Callan, Croft, & Harding, 1992), a probabilistic, inference net-based retrieval system. We then submitted to Inquery nine queries covering various diagnostic, etiologic, and therapeutic aspects of breast cancer. These ranged from brief queries (*adjuvant therapy of breast cancer and tamoxifen*) to more lengthy ones (*relationships between dietary fat and breast cancer. qualitative effects of dietary lipids and breast cancer. unsaturated vs. saturated fats. including epidemiological studies, clinical trials, and animal models*).

Our evaluation was necessarily informal, since we do not have a test collection with relevancy judgments based on partial documents. In order to provide some indication of the usefulness of our techniques, we compared the results of our searches using Inquery to the results produced by the HSTAT search facility.

HSTAT uses a sophisticated word-based search facility which identifies the smallest sections - including paragraphs, tables, and references - which match one or more terms of a query (cf. Fuller et al., 1993). The sets of matched terms are presented in a ranked list, and for each set of terms, document subcollections are ranked by the number of matching items they contain. For each user-selected subcollection, the links to the low-level content sections are presented in a similar fashion, first in ranked groups according to which query terms each matches, and within each group by order of occurrence in the text. The textual context surrounding the relevant passage is also provided. Hierarchical context includes (in a separate window) the table of contents which refers to the current passage.

We assume that comparing our results with the results from the HSTAT search facility gives some indication of how well our methods would work on a full-text database which did not have the sophisticated search facility provided by HSTAT. The results of using our techniques were very similar to those produced by the HSTAT search facility, with one important exception. The HSTAT facility seeks to identify and retrieve the lowest level passages within the document subcollections containing them. Our approach to hierarchical indexing allowed all the levels of a document to compete in the Inquery database, and could thus retrieve both small sections which satisfied the stated information need and larger sections or whole documents when these were better. Two examples illustrate this.

The top-ranked document returned for the query *chemotherapy for advanced metastatic breast cancer* is the same from the HSTAT search facility and from Inquery, and initially appears to be disappointing. It is a table from an AHCPR Health Technology Assessment Report entitled *Autologous Peripheral Stem-Cell Transplantation* (APSCT), and might not seem relevant. Nevertheless, because the technique being described is associated with high-dose chemotherapy, examination of this table ("Clinical experience involving APSCT") reveals a very pertinent reference to a journal article involving metastatic breast cancer. The hierarchical indexing in this instance allows the retrieval system to pinpoint the most relevant information in a large, full-text document. Since HSTAT always returns the smallest section, and since in this case the smallest section is the most relevant, there is no notable advantage to our approach. However, the opposite situation is seen in response to the following query.

For *adjuvant therapy of breast cancer and tamoxifen*, the top four documents returned by Inquery and the HSTAT search facility were sections from a single larger text, the NIH Consensus Development Conference Statement entitled *Treatment of Early-Stage Breast Cancer.* This is a twenty-page document divided into several sections, including a prefatory abstract and final conclusions and recommendations. According to both search methods, the top-ranked document was the section entitled "What is the role of adjuvant therapy for patients with node negative breast

cancer?" This is a short (two page) section containing a focused discussion particularly relevant to all concepts in the query. The other documents ranked highly by both systems were the abstract, the conclusions, and recommendations. According to Inquery, however, the fifth most relevant document for this query was the entire Consensus Development Conference Statement, suggesting that this single document had so many relevant parts that the entire document might be worthwhile in satisfying the information need expressed. By contrast, proposing the entire document is not within the scope of the HSTAT search facility. The added component of our hierarchical indexing can help indicate when smaller sections are most relevant to a query and when a larger section, or the whole document is more useful.

CONCLUSIONS

Natural language techniques for automatic concept-based full-text indexing, taking advantage of the explicit structural information often available in full-text documents, can make important contributions to integrated retrieval of biomedical information in distributed, network-accessible information sources. We have explored this potential in two main areas of information retrieval research: identifying appropriate information sources in a distributed environment, and improving retrieval of specific documents and document fragments from full-text sources.

As a basis for indexing the subject coverage of full-text information sources, we found several clear advantages to the use of these techniques. They provide relatively specific indexing terms, reflect the full scope of actual content, and use concepts from multiple structured biomedical vocabularies, without the need for expert coding and maintenance. At the same time, new issues arise, including the likely need for more frequent updates to reflect changes in content, and the sometimes egregious errors inevitable with automated indexing techniques.

For document retrieval, hierarchical concept-based indexing and document sectioning show promise for improving on word indexing alone. While previous work led us to expect there would be some improvement in retrieval from our conceptual mapping techniques (Aronson, Rindflesch, & Browne, 1994), we were not able to formally test this hypothesis. The use of structured, hierarchical indexing presented more clear-cut advantages, but will also require rigorous evaluation.

Standard formal testing (Sparck Jones, 1981) of the techniques being explored here would provide results in a metric such as 11-point average precision combining recall and precision measures. Such evaluation calls for a test collection with relevancy judgments which address both source and document selection, and which compare retrieval at different levels of document hierarchy. While several test collections for information retrieval exist in the biomedical domain (Schuyler, McCray, & Schoolman, 1989; Hersh, Buckley, Leone, & Hickam, 1994; Hersh, Hickam, Haynes, & McKibbon, 1994), they are constructed around abstracts for evaluating document selection and do not support questions involving either source selection or full-text databases. We are investigating how we might be able to conduct such formal testing as a future extension of our work.

Current research in information retrieval addresses a cognitive paradigm of evaluation (Harter & Hert, 1997) which goes beyond the standard model and is concerned with multiple dimensions of the system being evaluated, the inclusion of the user in the evaluation process, as well as interaction between the system and the user. The HSTAT full-text retrieval system supports a large user community and in principle provides an ideal environment for pursuing user-centered evaluation of source selection and hierarchical indexing of full-text documents.

In addition to HSTAT, the ISM covers several full-text sources, with significant differences in document structure, subject focus, update frequency and other characteristics. Transporting the methodology to other databases would help explore issues presented here, and doubtless raise new ones. NLM's Hazardous Substances Data Bank (HSDB) is one likely database for trial deployment and testing. This database has extensive full-text information in a fielded, hierarchical record structure, and will likely require changes in the methods used to handle the hierarchical aspects of content indexing. It will also present an added challenge in its specialized chemical vocabulary.

Future work could also address many unexplored parts of the current project. Document navigation, perhaps through a structured index, merits further study. For selection of information sources, term co-occurrence and frequency data seem likely to be important.

The results of this work show promise for richer ways to identify, search and navigate full-text documents. Research is accelerating within the digital library and information retrieval communities to develop new methods for both source- and document-level retrieval of full-text (and other) contents. NLM's PubMed™ retrieval system (National Center for Biotechnology Information, 1998), and its relationships with publishers to provide full text of medical journals, points to a time when most biomedical information might be available completely online. Technologies such as SGML and natural language processing tools are likely to play an important role in making the new wave of full-text information resources readily accessible to those who need them.

ACKNOWLEDGMENTS

We gratefully acknowledge the support and help of Maureen Prettyman and the HSTAT development team. We are also indebted to Susanne Humphrey for her diligent and insightful analysis of our automatic, concept-based indexing, and to Betsy Humphreys and Anna Harbourt for their helpful suggestions on the draft text of this article.

REFERENCES

Aronson, A., Rindflesch, T., & Browne, A. (1994). Exploiting a large thesaurus for information retrieval. In *RIAO 94 Conference Proceedings* (pp. 197-216). New York: Center for the Advanced Study of Information Systems, Inc. (CASIS) and Centre de Hautes Etudes Internationales d'Informatique Documentaire (CID).

Baldonado, M., Chang, C. & Gravano, L. (1997). Metadata for Digital Libraries: Architecture and Design Rationale. In R. Allen & E. Rasmussen (Eds.), *Proceedings of the 2nd ACM International Conference on Digital Libraries* (pp. 47-56).

Buckland, M., & Plaunt, C. (1997). Selecting Libraries, Selecting Documents, Selecting Data. In *Proceedings of the International Symposium on Research, Development and Practice in Digital Libraries: ISDL '97* [http://www.DL.ulis.ac.jp/ISDL97/proceedings/chris/chris.html]; print proceedings not yet available.

Callan, J. (1994). Passage-level evidence in document retrieval. In W. Croft & C. van Rijsbergen (Eds.), *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 302-309).

Callan, J., Croft, W., & Harding, S. (1992). The INQUERY retrieval system. In *Proceedings of the 3rd International Conference on Database and Expert System Applications* (pp. 347-356). Valencia, Spain: Springer-Verlag.

Callan, J., Lu, L., & Croft, W. (1995). Searching distributed collections with inference networks. In E. Fox, P. Ingwersen, & R. Fidel (Eds.), *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 21-28).

Chakravarthy, A. & Haase, K. (1995). NetSerf: Using semantic knowledge to find internet information archives. In E. Fox, P.Ingwersen, & R. Fidel (Eds.), *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 4-11).

Cousins, S., Paepcke, A., Winograd, T., Bier, E., & Pier, K. (1997). The digital library integrated task environment (DLITE). In R. Allen & E. Rasmussen (Eds.), *Proceedings of the 2nd ACM International Conference on Digital Libraries* (pp. 142-151).

Evans, D., Hersh, W., Monarch, I., Lefferts, R., & Handerson, S. (1991). Automatic indexing of abstracts via natural language processing using a simple thesaurus. *Medical Decision Making, 11(suppl),* s108-s115.

Fuller, M., Mackie, E., Sacks-Davis, R., & Wilkinson, R. (1993). Structured answers for a large structured document collection. In R. Korfhage, E. Rasmussen, & P. Willett (Eds.), *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 204-213).

Gravano, L., García-Molina, H., & Tomasic, A. (1994). The Effectiveness of GlOSS for the Text-Database Discovery Problem. In *Proceedings of the 1994 ACM SIGMOD Conference* (pp. 126-137).

Harter, S. & Hert, C. (1997). Evaluation of information retrieval systems: Approaches, issues, and methods. In M. Williams (Ed.) *Annual Review of Information Science and Technology,* Volume 32, 3-94.

Hearst, M. & Plaunt, C. (1993). Subtopic structuring for full-length document access. In R. Korfhage, E. Rasmussen, & P. Willett (Eds.), *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 59-68).

Hersh, W., Buckley, C., Leone, T., & Hickam, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In W. Croft & C. van Rijsbergen (Eds.), *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 192-201).

Hersh, W., Hickam, D., Haynes, R. B., & McKibbon, K. A. (1994). A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association,* 1, 51-60.

Humphreys, B., Lindberg, D., Schoolman, H., & Barnett, G. (1998). The Unified Medical Language System: An informatics research collaboration. *Journal of the American Medical Informatics Association, 5,* 1-11.

Kaszkiel, M. & Zobel, J. (1997). Passage retrieval revisited. In N. Belkin, A. Narasimhalu, & P. Willett (eds.) *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 178-185).

Lindberg, D., Humphreys, B., & McCray, A. (1993). The Unified medical Language System. *Methods of Information in Medicine, 32,* 281-291.

Masys, D. (1992). An Evaluation of the Source Selection Elements of the Prototype UMLS Information Sources Map. In M. Frisse (Ed.) *Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care* (pp. 295-298).

Masys, D. & Humphreys, B. (1992). Structure and function of the UMLS information sources map. *Proceedings of MEDINFO 92* (pp. 1518-21).

McCray, A., Srinivasan, S., & Browne, A. (1994). Lexical methods for managing variation in biomedical terminologies. In J. Ozbolt (Ed.), *Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care* (pp. 235-239).

McKinin, E., Sievert, M., Johnson, E., & Mitchell, J. (1991). The MEDLINE/full-text research project. *Journal of the American Society for Information Science, 42*, 297-307.

Miller, P., Frawley, S., Wright, L., Roderer, N., Powsner, S., (1995). Lessons learned from a pilot implementation of the UMLS information sources map. *Journal of the American Medical Informatics Association, 2*, 102-115.

National Center for Biotechnology Information (1998). The NLM PubMed Project (refer to the URL http://www.ncbi.nlm.nih.gov/PubMed/overview.html).

Prettyman, M. (1997). SGML as a Navigational Tool for Accessing Information. In *SGML/XML'97 Conference Proceedings* (pp. 65-70). Washington, D.C.: GCA.

Purcell, G. & Mar, D. (1992). SCOUT: Information retrieval from full-text medical literature. Mark E. Frisse (ed.) *Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care* (pp. 91-95).

Rodgers, R. (1995). Automated retrieval from multiple disparate information sources: The world wide web and the NLM's sourcerer project. *Journal of the American Society for Information Science*, 46, 755-764.

Salton, G., Allan, J., & Buckley, C. (1993). Approaches to passage retrieval in full text information systems. In R. Korfhage, E. Rasmussen, & P. Willet (Eds.), *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 49-58).

Schuyler, P., McCray, A., & Schoolman, H. (1989). A test collection for experimentation in bibliographic retrieval. In B. Barber, D. Cao, D. Qin, & G. Wagner (Eds.) *Proceedings of MEDINFO 89* (pp. 810-912).

Sievert, M. (1996). Full-text information retrieval: Introduction. *Journal of the American Society for Information Science, 47,* 261-262.

Sievert, M., McKinin, E., & Johnson, E. (1995). Full-text databases in medicine. *Journal of the American Society for Information Science, 46,* 748-754.

Smeaton, Alan F. (1992). Progress in the application of natural language processing to information retrieval tasks. *The Computer Journal, 35,* 268-278.

Sparck Jones, K. (Ed.) (1981). *Information Retrieval Experiment*. London: Butterworths.

Voorhees, E., Gupta, N., & Johnson-Laird, B (1995). The Collection Fusion Problem. In D. Harman (Ed.), *Proceedings of the Third Text REtrieval Conference (TREC-3)* (pp. 95-104). (Gaithersburg, MD: National Institute of Standards and Technology).

Voorhees, E. & Tong, R. (1997). Multiple search engines in database merging. In R. Allen & E. Rasmussen (Eds.), *Proceedings of the 2nd ACM International Conference on Digital Libraries* (pp. 93-102).

Wagner, M. (1991). An automatic indexing method for medical documents. In P. Clayton (Ed.), *Proceedings of he Fifteenth Annual Symposium on computer Applications in Medical Care (*pp. 1011-17).

Wilkinson, R. (1994). Effective retrieval of structured documents. In W. Croft & C. van Rijsbergen (Eds.), *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 311-317).

Appendix. **Information Sources Map: Alphabetical Summary of Sources Covered**

AI/RHEUM	NLM expert system: diagnostic consultant system for rheumatology
AIDSDRUGS	NLM database: substances being tested in AIDSTRIALS clinical trials
AIDSLINE®	NLM bibliographic database: AIDS-related items, most with abstracts
AIDSTRIALS	NLM database: AIDS-related clinical trials of AIDSDRUGS substances
AVLINE®	NLM bibliographic database: biomedical audiovisual materials/computer software
BIOETHICSLINE®	NLM bibliographic database: biomedical ethics & related issues
CANCERLIT®	NLM bibliographic database: cancer-related items, 83% w/ abstracts
CATLINE®	NLM bibliographic database: biomedical monographs and serials titles
CCRIS	NCI database: cancer-related chemicals
ChemID®	NLM vocabulary: compounds of biomedical & regulatory interest
DART®	EPA/NIEHS/FDA/NLM bibliographic database: teratology, developmental/reproductive toxicology
DIRLINE®	NLM directory: biomedical information resources
DNA Data Bank of Japan	National Institute of Genetics database: nucleotide sequence & other data produced in Japan
DOCUSER®	NLM directory: biomedical libraries & related entities
DXplain™	Mass. Gen. Hospital LCS expert system: disease-focused electronic textbook & reference system
EMBL	EMBL database: nucleotide sequence & related data, incl. citations
EMIC	ORNL/NLM/EPA/NIEHS bibliographic database: mutagenicity and genotoxicity
Envirofate	EPA/SRC database: released chemicals' properties & environmental fate
GenBank®	NLM database: all public genetic sequence data, & related information
GENE-TOX	EPA database: chemicals tested for mutagenicity
Ground Water On-Line	NGWA bibliographic database: all aspects of ground (well) water
Health Devices Alerts	ECRI expert synthesis: health devices problems, technology assessments etc.
Healthcare Product Comparison System	ECRI database: comparisons of major healthcare capital equipment
HealthSTAR™	NLM bibliographic database: health services technology, administration & research
HISTLINE®	NLM bibliographic database: history of medicine and related sciences
HSDB®	NLM expert synthesis: hazardous chemicals' effects, environmental fate, handling, etc.
HSRPROJ	NLM directory: ongoing health services research grants & contracts
HSTAT	NLM full-text database: clinical practice guidelines & related documents
ILIAD©	Univ. Utah/A.D.A.M. Software, Inc. expert system: medical diagnostic & treatment system
Images from the History of Medicine	NLM image database: prints and photographs on the history of medicine
IRIS	EPA database: risk & regulatory data on potentially toxic chemicals
MEDLINE®	NLM bibliographic database: article citations & abstracts on all biomedicine
MeSH®	NLM vocabulary: biomedical and related indexing terms
NEDRES	NOAA directory: environmental data sources
NIOSHTIC	NIOSH bibliographic database: occupational safety & health literature
NPIRS	Purdue Univ. CERIS database: pesticide descriptions, warnings and other data
Online Mendelian Inheritance in Man (OMIM™)	Johns Hopkins Univ./NLM expert synthesis: essays on human genetic traits, with references
Physician Data Query (PDQ®)	NCI expert synthesis: cancer descriptions, treatment and related contacts
POPLINE®	NLM bibliographic database: populations, demographics and family planning
Protein Identification Resource (PIR)	NBRF/JIPID/MIPS database: protein sequences & related data
Quick Medical Reference (QMR)	Camdat Corp. expert system: internal medicine information & diagnostic tool
RTECS®	NIOSH database: effects of potentially toxic chemicals
SERLINE®	NLM bibliographic database: biomedical serials titles & related publications
SPACELINE™	NLM/NASA bibliographic database: space life sciences
Technology Transfer Network	EPA information services: air quality databases, BBSs, file transfer, email

Appendix. **Information Sources Map: Alphabetical Summary of Sources Covered**

Toxic Chemical Release Inventory (TRI)	EPA database: yearly data on releases of toxic chemicals
TOXLINE®	NLM bibliographic database: effects of chemical, biological & physical agents
TRIFACTS	EPA database: safety, handling & effects of chemicals in TRI files
Visible Human Project™	NLM image database: complete, anatomical, 3-D data of male & female humans