Humphrey, S., Rindflesch, T., & Aronson, A. (2000). Automatic indexing by discipline and high-level categories: Methodology and potential applications. . 11th ASIS SIG/CR Classification Research Workshop, 103-116. doi:10.7152/acro.v11i1.12773

Automatic Indexing by Discipline and High-Level Categories: Methodology and Potential Applications

Susanne M. Humphrey*, Thomas C. Rindflesch, and Alan R. Aronson*****
Lister Hill National Center for Biomedical Communications
National Library of Medicine,
Bethesda, MD 20894, USA

* humphrey@nlm.nih.gov, 301-435-3187, 301-496-0673 (Fax)
** tcr@nlm.nih.gov, 301-435-3191, 301-496-0673 (Fax)
*** alan@nlm.nih.gov, 301-435-3162, 301-496-0673 (Fax)

This paper first describes the methodology of journal descriptor (JD) indexing, based on human indexing at the journal level using only 127 descriptors, and applying statistical methods that associate this journal indexing with text words in a training set of MEDLINE® citations. These associations form the basis for automatic indexing of documents outside the training set. The paper then presents the new technique of semantic type (ST) indexing, based on JD indexing associated with each of 134 ST's, and applying the standard cosine coefficient measure to compare the similarity between the JD indexing of a document and the JD indexing of each ST. The ST indexing of the document is the list of ST's ranked in decreasing order of similarity between the JD indexing of the document and the JD indexing of the ST's. Discussion of the potential usefulness and application of the very general indexing provided by JD's and ST's comprises the remainder of the paper. JD's have been used for more than thirty years to search MEDLINE by discipline, and discipline-based indexing is in evidence on the Web. It is suggested, with several examples, that ST's may convey a unique slant of a document's content not normally represented in standard indexing vocabularies. Use of ST indexing to rank retrieved output is mentioned as a possible application. Notwithstanding the importance of methodology and performance issues, the intent of this paper is to explore questions of the potential utility and applicability of JD and ST indexing.

1. INTRODUCTION

As part of the National Library of Medicine's Indexing Initiative (Aronson, Bodenreider, Chang, Humphrey, Mork, Nelson, Rindflesch & Wilbur, 2000), research has been performed on the automatic indexing of documents using journal descriptors (JD's). A new aspect to this JD indexing research is extending it to UMLS® (Unified Medical Language System®) semantic type (ST) indexing using the cosine measure for comparing documents.

Both JD and ST indexing are fully automated approaches to indexing text using a controlled vocabulary. However, unlike studies which attempt to index text automatically using the full set of descriptors in a large thesaurus, such as NLM's Medical Subject Headings (MeSH®), the aim of this research is to index text with a very small set of controlled descriptors.

ISSN: 2324-9773

Humphrey, S., Rindflesch, T., & Aronson, A. (2000). Automatic indexing by discipline and high-level categories: Methodology and potential applications. . 11th ASIS SIG/CR Classification Research Workshop, 103-116. doi:10.7152/acro.v11i1.12773

JD indexing uses only 127 descriptors extracted from MeSH, corresponding to large areas of biomedical knowledge, most of them biomedical disciplines, e.g., 'Anatomy', 'Biochemistry', 'Cardiology', used for describing journals indexed in NLM's MEDLINE database. ST indexing uses only 134 descriptors, which are the semantic types in NLM's UMLS Semantic Network, e.g., 'Disease or Syndrome', 'Diagnostic Procedure', 'Spatial Concept'. All concepts in the UMLS, including those from MeSH, are assigned one or more ST's linked to the concepts in an is-a relationship, for example, 'Angiography' is-a 'Diagnostic Procedure', 'Round shape' is-a 'Spatial Concept', 'Cholesterol' is-a 'Biologically Active Substance', 'Cholesterol' is-a 'Steroid'.

The basis for both JD indexing and ST indexing is assignment of JD's as descriptors of journals by humans. This can be considered minimal human effort compared to human indexing of text found in these journals, in books, on the Web, and so forth. The actual implementation of JD and ST indexing is statistical and uses accepted techniques described widely in the Information Retrieval (IR) literature. In a sense, JD's and ST's are used as labels of sets of words that bear strong statistical correlations with particular JD's and ST's. In contrast to approaches that cluster documents based on word co-occurrences without automatically labeling the clusters, there is the added benefit of having the JD's and ST's themselves as labels which can be used either directly by humans as search terms or in further computations to enhance other techniques used for retrieval.

Given the very general nature of such indexing, i.e., the fact that there are so few descriptors, and the situation that the emphasis in research is normally much more detailed indexing, the question of the utility of such indexing arises. Does general indexing improve even conventional retrieval where detailed indexing is available? There is good anecdotal evidence that it does. Experienced searchers have been using JD parameters (e.g., limiting the retrieval to Cardiology journals) for years in searching MEDLINE, although retrieval systems do not facilitate this technique. This particular research question has not been addressed to our knowledge.

JD indexing is being explored as an adjunct to other Indexing Initiative systems, in particular MetaMap, to determine if very general indexing can serve to improve the performance and results of other systems. Specifically, does describing the very general context of a document, or even a word in a document, assist the process and performance of more detailed automated indexing?

This paper will describe the JD indexing methodology as necessary background to understand the basis of the ST indexing technique, which is described subsequently. In the section on applications, it is emphasized that JD indexing has a long history of applicability to MEDLINE indexing, and that discipline-based indexing has naturally emerged at Web sites to further guide users in retrieving information. In discussing the utility of ST indexing, it is suggested that this indexing would often convey a general as well as rather unique sense of the context of a document. For example, the ST 'Laboratory Procedure' as an indexing term would convey a slant to the content of a document that complements conventional indexing terms such as 'Colony Count, Microbial', 'Sputum/MICROBIOLOGY', and 'Tuberculosis, Pulmonary/ MICROBIOLOGY'. Is 'Laboratory Procedure' automatically assigned to documents a useful descriptor for retrieving studies emphasizing the use of laboratory procedures? Furthermore, some ST's, such as 'Spatial Concept' express potential search parameters that normally are not expressed in conventional indexing vocabularies. Documents assigned this ST typically discuss surgery, anatomy or diagnostic imaging. Is this a useful descriptor to retrieve documents in which spatial relationships are emphasized? Like JD indexing, can ST indexing serve as a useful adjunct

ISSN: 2324-9773

Humphrey, S., Rindflesch, T., & Aronson, A. (2000). Automatic indexing by discipline and high-level categories: Methodology and potential applications. . 11th ASIS SIG/CR Classification Research Workshop, 103-116. doi:10.7152/acro.v11i1.12773

to other automated indexing approaches? The authors look forward to comments from other workshop participants on these sorts of question.

2. JOURNAL DESCRIPTOR INDEXING METHODOLOGY

JD indexing (Humphrey, 1998; Humphrey, 1999) uses a training set consisting of words in titles and abstracts of MEDLINE citations (also referred to as "documents" in this paper), and their statistical associations with 127 journal descriptors which index the approximately 4000 MEDLINE journals. Figure 1 shows a sample journal record for *Clinics in Chest Medicine* in NLM's journal (i.e., serial records) file, including the humanly-assigned JD 'Pulmonary Disease (Specialty)'.

TI - Clinics in Chest Medicine
TA - Clin Chest Med
JD - Pulmonary Disease (Specialty)

Figure 1. NLM's journal record for *Clinics in Chest Medicine* showing the JD 'Pulmonary Disease (Specialty)'.

Figure 2 shows a sample citation from the training set, including the JD 'Pulmonary Disease (Specialty)' mapped from the journal record. Thus, citations in the training set "inherit" JD's from journal records corresponding to the journals in which the documents are published. Each word in the sample title (Figure 2) from the training set (including *pulmonary,* which we emphasize) can be said to co-occur with the JD 'Pulmonary Disease (Specialty)' by virtue of this inheritance.

UI - 95385322
TI - Diagnosis and management of acute *pulmonary* embolism. Past, present, and future.
SO - Clin Chest Med 1995 Jun;16(2):229-33
*JD - Pulmonary Disease (Specialty)

*mapped from the journal record for *Clin Chest Med*

Figure 2. Sample document in the training set showing inheritance of JD from NLM's journal record.

The current training set contains 186,768 different words in 181,188 documents (MEDLINE citations from 2756 different journals). The word *pulmonary*, for example, occurs 9177 times in 3774 documents. Since each document inherits one or more JD's, an association between words and JD's can be represented as the number of co-occurrences of each word with each JD in the documents in the training set. Figure 3 shows the number of co-occurrences of *pulmonary* with the JD's. The top-ranked JD co-occurring with *pulmonary* is 'Pulmonary Disease (Specialty)'. Specifically, the word *pulmonary* occurs with the JD 'Pulmonary Disease (Specialty)' 2428 times in 922 documents.

The JD rankings for *pulmonary* can be expressed two different ways: 1) by the ratio of word count for *pulmonary* co-occurring with the JD divided by the total word count for

ISSN: 2324-9773

Humphrey, S., Rindflesch, T., & Aronson, A. (2000). Automatic indexing by discipline and high-level categories: Methodology and potential applications. . 11th ASIS SIG/CR Classification Research Workshop, 103-116. doi:10.7152/acro.v11i1.12773

JD	*pulmonary* word count	*pulmonary* citation count
Pulmonary Disease (Specialty)	2428	922
Cardiology	1326	401
Surgery	1215	460
Critical Care	499	199
Anesthesiology	368	151
etc.		
All JD's	9177	3774

Figure 3. Co-occurrences of the word *pulmonary* with JD's in the training set.

pulmonary in the training set, or 2) the ratio of citation count in which *pulmonary* co-occurs with the JD divided by the total citation count for *pulmonary* in the training set. In either case, the 127 JD rankings for *pulmonary*, ordered alphabetically, form a vector, as do the JD rankings for each word. For example, the JD vector for *pulmonary* based on word count is shown in Figure 4.

Acquired Immunodeficiency Syndrome	0.0005
Aerospace Medicine	0.0002
...	
Anesthesiology	0.0126
...	
Cardiology	0.0216
...	
Critical Care	0.0283
...	
Pulmonary Disease (Specialty)	0.0552
...	
Surgery	0.0118
...	
Vital Statistics	0.0001

Figure 4. JD vector for the word *pulmonary* based on word count in the training set (ratio of word count for *pulmonary* co-occurring with the JD divided by total word count for *pulmonary*).

If we wish to index a document (outside the training set), and we know only that it contains the word *pulmonary*, we can say, based on these associations, that this document is most likely to be in the speciality of pulmonary disease. The likelihood that a document is in this field increases the more words it has that co-occur with 'Pulmonary Disease (Specialty)' in the training set (assuming the application of normalization methods such as inverse document frequency). We therefore can assign JD's as indexing terms to a document based on the words in the document

ISSN: 2324-9773

Humphrey, S., Rindflesch, T., & Aronson, A. (2000). Automatic indexing by discipline and high-level categories: Methodology and potential applications. . 11th ASIS SIG/CR Classification Research Workshop, 103-116. doi:10.7152/acro.v11i1.12773

that also occur in the training set. We do this by computing a vector, which is the centroid of the JD vectors for all the words in the document to be indexed. The ranking for a JD in the centroid is the average of the rankings for this JD across all the words. We can also take into consideration the frequency of words in the document to be indexed (again, applying normalization methods). For example, using this technique, we index a MEDLINE title and abstract outside the training set, as shown by the ranked JD's in Figure 5.

TI - The early bactericidal activity of ciprofloxacin in patients with pulmonary tuberculosis.

AB - The early bactericidal activity (EBA) of ciprofloxacin (CIP) was measured in 80 patients with previously untreated, smear-positive pulmonary tuberculosis by counting viable bacilli in sputum collections during the first 2 d of treatment. Groups of about 10 patients were treated daily with graded doses of CIP or with 300 mg isoniazid or with no drug. The mean EBA, defined as the fall in log CFU/ ml sputum/d, increased from -0.011 in the no drug group to 0.046, 0.092, 0.121, and 0.205 in the groups receiving 250, 500, 1,000, or 1,500 mg CIP, respectively, a highly significant trend. These results demonstrate the antimycobacterial activity of CIP in high dosage, though the mean EBAs of 0.55 and 0.66 in two groups receiving isoniazid were much higher.

JD	Rank
Drug Therapy	0.0153
Antibiotics	0.0123
Pulmonary Disease (Specialty)	0.0113
Communicable Diseases	0.0108
Microbiology	0.0085
...	

Figure 5. JD indexing of sample MEDLINE title and abstract outside the training set.

3. SEMANTIC TYPE INDEXING METHODOLOGY

We are now investigating the automatic indexing of documents using the 134 semantic types in NLM's UMLS Knowledge Sources (National Library of Medicine, 2000). The underlying methodology in exploiting the UMLS semantic types for indexing is to compute a JD vector for each semantic type, just as a JD vector can be computed for a MEDLINE title and abstract. For example, the 127 JD's, ordered alphabetically, and their numerical rankings that index the above document (in Figure 5) form a vector, as shown in Figure 6. Similarly, a vector for each UMLS semantic type is created by considering the set of words making up all the concepts assigned to an ST as if they comprised a document, and then indexing this "semantic type document" with the JD methodology described above.

ISSN: 2324-9773

Humphrey, S., Rindflesch, T., & Aronson, A. (2000). Automatic indexing by discipline and high-level categories: Methodology and potential applications. . 11th ASIS SIG/CR Classification Research Workshop, 103-116. doi:10.7152/acro.v11i1.12773

Acquired Immunodeficiency Syndrome	0.0030
Aerospace Medicine	0.0003
...	
Antibiotics	0.0123
...	
Chemistry	0.0017
...	
Communicable Diseases	0.0108
...	
Drug Therapy	0.0153
...	
Microbiology	0.0085
...	
Pharmacology	0.0047
...	
Pulmonary Disease (Specialty)	0.0113
...	
Vital Statistics	0.0002

Figure 6. JD vector comprised of ordered numerical rankings of JD's indexing the sample MEDLINE title and abstract (Figure 5).

An important aspect of current work involves determining what text to use in computing the vector that represents each semantic type. The default criterion is to consider all UMLS phrasal strings that have been assigned a certain semantic type as the document for that semantic type. For example, Figure 7 illustrates the concepts assigned the semantic type 'Biomedical Occupation or Discipline' in the Metathesaurus:

...
"Adolescent Medicine"
"Adolescent Psychiatry"
"Adolescent Psychology"
"Aerospace Medicine"
"African Medicine, Traditional"
"Allergy and Immunology"
...

Figure 7. UMLS Metathesaurus concepts assigned semantic type 'Biomedical Occupation or Discipline'.

There are a number of ways in which this semantic type document can be modified in an attempt to compute the most effective JD vector for representing each of the UMLS semantic types. Many of the Metathesaurus strings have been assigned more than one semantic type, often

ISSN: 2324-9773

Humphrey, S., Rindflesch, T., & Aronson, A. (2000). Automatic indexing by discipline and high-level categories: Methodology and potential applications. . 11th ASIS SIG/CR Classification Research Workshop, 103-116. doi:10.7152/acro.v11i1.12773

because the string is ambiguous. For example, the string "discharge" has semantic type 'Body Substance' in addition to 'Health Care Activity' to reflect its ambiguous meaning.

One possible way to modify the semantic type document for each of the semantic types is to include only those Metathesaurus concepts that are unambiguous for each semantic type; that is, for each semantic type, consider only those concepts which have been assigned that semantic type exclusively. For example, "African Medicine, Traditional" in Figure 7 has been assigned the semantic type 'Health Care Activity' in addition to 'Biomedical Occupation or Discipline', and so would not be included in the list of concepts for 'Biomedical Occupation or Discipline' when the unambiguous concepts are considered as the semantic type document for this semantic type.

The Metathesaurus is a compilation of more than fifty vocabularies, including NLM's MeSH terminology. Another way to modify the semantic type document for each semantic type is to compute the JD vector for each semantic type based on only those strings in MeSH that have been assigned that semantic type. For example, Figure 8 represents all the Metathesaurus concepts with the semantic type 'Disease or Syndrome', while Figure 9 has MeSH terms only.

...
"Abdominal Abscesses"
"Abdominal Aortic Aneurysms"
"Abdominal Epilepsies"
"Abdominal Pregnancies"
"Abdominal actinomycosis"
"Abdominal actinomycotic infection"
...

Figure 8. Metathesaurus concepts for 'Disease or Syndrome'.

...
"Abdominal Abscesses"
"Abdominal Aortic Aneurysms"
"Abdominal Epilepsies"
"Abdominal Pregnancies"
...

Figure 9. MeSH concepts only for semantic type 'Disease or Syndrome'.

Currently, we consider the most effective semantic type document to be composed of unambiguous strings from all constituent vocabularies in the UMLS. For example, the 1999 Metathesaurus contains 1023 concepts assigned only to the ST 'Antibiotic' (e.g., Ampicillin+penicillinase-resistant penicillin, Cloxacillin, cloxacillin sulfone, Lactams, etc.). We

ISSN: 2324-9773

Humphrey, S., Rindflesch, T., & Aronson, A. (2000). Automatic indexing by discipline and high-level categories: Methodology and potential applications. . 11th ASIS SIG/CR Classification Research Workshop, 103-116. doi:10.7152/acro.v11i1.12773

form an ST document for 'Antibiotic' from the set of words that make up these UMLS concepts, and index them as if they comprised a title, as shown by the ranked JD's in Figure 10.

TI - ampicillin penicillinase resistant penicillin lactams cloxacillin sulfone, etc.

JD	Rank
Antibiotics	0.0550
Drug Therapy	0.0484
Pharmacology	0.0166
Communicable Diseases	0.0152
Chemistry	0.0147
...	

Figure 10. JD indexing of "semantic type document" for semantic type 'Antibiotics', comprised of the set of words from phrases corresponding to UMLS concepts assigned this semantic type.

The 127 ordered JD's and their numerical rankings which index this document form the vector shown in Figure 11.

Acquired Immunodeficiency Syndrome	0.0006
Aerospace Medicine	0.0001
...	
Antibiotics	0.0550
...	
Chemistry	0.0147
...	
Communicable Diseases	0.0152
...	
Drug Therapy	0.0484
...	
Microbiology	0.0128
...	
Pharmacology	0.0166
...	
Pulmonary Disease (Specialty)	0.0017
...	
Vital Statistics	0.0007

Figure 11. JD vector comprised of ordered numerical rankings of JD's indexing the 'Antibiotics' semantic type document.

Then, the standard vector cosine coefficient (Salton & McGill, 1983) is used to produce a result that measures on a scale from 0 to 1 the similarity between the above document to be

ISSN: 2324-9773

Humphrey, S., Rindflesch, T., & Aronson, A. (2000). Automatic indexing by discipline and high-level categories: Methodology and potential applications. . 11th ASIS SIG/CR Classification Research Workshop, 103-116. doi:10.7152/acro.v11i1.12773

indexed (Figure 5) and this ST document, the result being 0.7776. We measure the similarity between the document to be indexed and each of the 134 ST documents, and rank the ST's in decreasing order of similarity, resulting in a ranked list of ST indexing terms for this document (Figure 12). This result shows that the document to be indexed is most similar to the 'Antibiotic' ST document, followed by the 'Clinical Drug' ST document, the 'Laboratory Procedure' ST document, and so forth.

Antibiotic	0.7776
Clinical Drug	0.7321
Laboratory Procedure	0.7260
Pharmacologic Substance	0.7228
Disease or Syndrome	0.7000
...	

Figure 12. ST indexing of sample MEDLINE title and abstract (Figure 5) based on similarity between the JD vector for the sample document and the JD vectors for each of the ST documents.

4. APPLICATIONS: INDEXING BY DISCIPLINE AND BY HIGH-LEVEL CATEGORIES

Indexing by discipline has a long history of applicability for biomedical retrieval. Experienced searchers have found discipline-based parameters quite useful for certain types of MEDLINE (and before that, batch MEDLARS®) queries since the beginning of computerized searching in the mid-1960's. Examples of such queries would be: 1) Neurotransmitters in Cardiology (intersecting words with disciplines), and 2) Pediatric Cardiology (intersecting disciplines), where expressing the Cardiology parameter in the search would be particularly problematic. One way this has been done traditionally is forming the union of cardiology journal titles or codes of which there are 75 currently indexed in MEDLINE, i.e., Acta Cardiol OR Adv Cardiol OR Am Heart J OR Am J Cardiol OR Br Heart J OR etc. Specifying this parameter would be quite difficult, if not impossible, in current retrieval systems

A more fundamental problem is that JD's are restricted to specific journals, e.g., the JD 'Cardiology' is available only for journals in that discipline. For example, using this parameter for the Neurotransmitters in Cardiology search would miss retrieving the citation titled "Plasma neurotransmitters, blood pressure and heart rate ..." from the journal *Psychotherapy and Psychosomatics*. However, entering this title as a document to be indexed by the JD indexing methodology results in 'Cardiology' being returned immediately as the top-ranked indexing term, and therefore makes this citation retrievable for this type of query. Thus, to enhance searching by discipline, the JD indexing system has the potential to quickly assign discipline-based indexing terms to citations in the entire MEDLINE database for retrieval regardless of the discipline associated with the particular journal.

JD indexing can potentially be extended to areas other than biomedicine. Many Library of Congress (LC) subject classes are disciplinary in nature. For example, the journal *Architecture* has

ISSN: 2324-9773

Humphrey, S., Rindflesch, T., & Aronson, A. (2000). Automatic indexing by discipline and high-level categories: Methodology and potential applications. . 11th ASIS SIG/CR Classification Research Workshop, 103-116. doi:10.7152/acro.v11i1.12773

a call number beginning NA1, in the range of NA1 - NA9428 corresponding to the 'Architecture' subclass; *Journal of Climate* is QC851, in the range corresponding to the subclass 'Meteorol. Climatol'. Thus, associations between LC subclasses, serving as JD's for journals in various disciplines, and words in titles and abstracts in these journals can form the basis for indexing documents by discipline, in turn, enabling these documents to be retrieved by discipline.

The Web provides numerous examples of indexing by discipline. For example, the American Academy of Pediatrics (AAP) Web page http://www.pediatrics.org/collections provides selectable subspecialty collections of studies published in their journal *Pediatrics*, using categories such as 'Endocrinology', 'Nutrition & Metabolism', 'Surgery', 'Allergy & Dermatology', and so forth. The editors of this journal have provided this indexing "to extend users' exploration of the topic area." Based on checking a few citations, it was noted that the editors placed each article in only a single category. For example, the article titled, "Distal Sensory Polyneuropathy in a Cohort of HIV-Infected Children Over Five Years of Age" appears in the 'Infectious Disease & Immunity' collection, but is missing from the 'Neurology & Psychiatry' collection. Authors of this article publish in neurology journals. This obviously would be problematic for the searcher expecting to find all neurology articles in the 'Neurology & Psychiatry' collection. An example where discipline-based indexing is altogether unavailable but would be useful is the New England Journal of Medicine (File 444) searchable using Dialog (http:www.dialogweb.com).

While JD indexing is designed to provide the discipline or field with which the content of a document is associated, ST indexing often gives a general sense of the content using high-level categories, namely, the 134 nodes of the UMLS semantic network to which all UMLS concepts have an IS-A relationship. In the ST indexing in Figure 12, 'Laboratory Procedure' succinctly conveys a cross-disciplinary slant to the content of the document (Figure 5) that would be difficult to express using conventional search terms. This can be compared to the actual human MeSH indexing of this document that corresponds to this sense, but is much more specific (Figure 13).

MH - Colony Count, Microbial
MH - Sputum/MICROBIOLOGY
MH - Tuberculosis, Pulmonary/MICROBIOLOGY

Figure 13. Manually assigned MeSH indexing terms corresponding to Laboratory Procedure, a highly ranked indexing term assigned automatically by ST indexing.

Further examples illustrate some of the characteristics of the automatic indexing methodologies being pursued. In comparison to traditional indexing terms, the coarser granularity of the UMLS semantic types allows techniques associated with labeled document categorization to be used in cooperation with JD indexing to support innovative methods of information management and retrieval. We provide several examples based on informal investigations that suggest some of the benefits. For this preliminary study, we considered only the first three JD's and semantic types from the ranked lists assigned to each document. We intend to call on statistical techniques to formalize the most promising results of this exploratory research.

Several UMLS semantic types refer to classes of broadly-defined ontological concepts, such as 'Spatial Concept', 'Temporal Concept', and 'Quantitative Concept'. These appear to have

ISSN: 2324-9773

Humphrey, S., Rindflesch, T., & Aronson, A. (2000). Automatic indexing by discipline and high-level categories: Methodology and potential applications. . 11th ASIS SIG/CR Classification Research Workshop, 103-116. doi:10.7152/acro.v11i1.12773

unexpectedly beneficial consequences for categorization and retrieval. MEDLINE citations assigned the 'Spatial Concept' as one of the top three semantic types, for example, typically discuss surgery, anatomy, or diagnostic imaging. Some titles illustrating this are given in Figure 14.

Mitral valve repair in a patient with severe porcelain aorta.
Repair of pseudoaneurysms via ultrasound-guided compression.
MRI and CT of metastatic hepatocellular carcinoma causing spinal cord compression.
Comparison of B-mode, M-mode and echo-tracking methods for measurement of the arterial distension waveform.
Angioarchitectural classification of the fungiform papillae on the dorsal surface of the bullfrog tongue.

Figure 14. Titles of MEDLINE citations assigned semantic type 'Spatial Concept'.

Other highly-ranked semantic types and the top-ranked JD's that co-occur with 'Spatial Concept' further isolate the specific content of these documents. Several examples illustrate the correlation of document content with the indexing terms assigned automatically. The article associated with the title in Figure 15 discusses the technique employed during surgical intervention addressing the condition noted. The semantic type 'Therapeutic or Preventive Procedure' suggests this focus, while 'Medical Device' is due to the crucial role of "combined axillary and femoral arterial cannulations" during the procedure. The top three JD's assigned to this citation specify the therapeutic modality and associated body systems.

Title: Mitral valve repair in a patient with severe porcelain aorta.
Last sentence: Likewise, the combination of a superior mitral approach and profound hypothermic fibrillatory arrest in conjunction with low-flow cardiopulmonary bypass allowed us to repair the mitral valve successfully.
Top three ST's: Spatial Concept; Therapeutic or Preventive Procedure; Medical Device
Top three JD's: Pulmonary Disease (Specialty); Cardiology; Surgery

Figure 15. ST and JD indexing of text discussing surgical intervention technique in cardiac disease.

A second example, outlined in Figure 16, illustrates an article that concentrates on general methodology for diagnostic imaging in cardiology, rather than a specific clinical case. The semantic types 'Diagnostic Procedure', 'Medical Device', and 'Spatial Concept' indicate the broadly defined semantic content of this article, while more detail is provided by the JD's 'Diagnostic Imaging', 'Cardiology', and 'Vascular Diseases'.

The journal article referred to in Figure 17 is primarily concerned with the elucidation of anatomical structures through the use of a scanning electron microscope. Diagnostic methodology and therapeutic intervention are not discussed in this paper. Phrases such as *dorsal **surface*** and *three-**dimensional*** highly correlate with the assigned semantic type 'Spatial Concept'. The other highly-ranked indexing terms accurately reflect the thrust of the research reported. Although

ISSN: 2324-9773

Humphrey, S., Rindflesch, T., & Aronson, A. (2000). Automatic indexing by discipline and high-level categories: Methodology and potential applications. . 11th ASIS SIG/CR Classification Research Workshop, 103-116. doi:10.7152/acro.v11i1.12773

Title: Comparison of B-mode, M-mode and echo-tracking methods for measurement of the arterial distension waveform.
Last sentence: On the basis of this relative comparison of methods, we conclude that, although echo tracking offers high resolution for continuous measurements, the reproducibility of discontinuous measurements of carotid artery distension is no better with echo tracking than can be obtained from 30-Hz B-mode images.
Top three ST's: Diagnostic Procedure; Medical Device; Spatial Concept
Top three JD's: Diagnostic Imaging; Cardiology; Vascular Diseases

Figure 16. ST and JD indexing of text discussing diagnostic methodology in cardiology.

dentistry is not actually discussed, the closely-associated terminology used for the anatomy of the tongue accounts for the JD 'Dentistry'.

Title: Angioarchitectural classification of the fungiform papillae on the *dorsal* ***surface*** of the bullfrog tongue.
First sentence: In this study, the *three-**dimensional*** anatomy of the microvascular structure of fungiform papillae (FuP) on the dorsal surface of the bullfrog tongue has been investigated using scanning electron microscopy...
Top three ST's: Medical Device; Organ or Tissue Function; Spatial Concept
Top three JD's: Dentistry; Anatomy; Physiology

Figure 17. ST and JD indexing of text discussing anatomy of oral structure.

A final set of examples illustrates the cooperative contribution of the more general UMLS semantic types and the more specific journal descriptors, but considers a specific JD as being primary. The following three citations discuss gastroenterological issues from various points of view: diagnostic technique (Figure 18), surgical intervention (Figure 19), and health care research (Figure 20). The associated automatically assigned indexing terms accurately reflect the focus of the content in each case.

Title: Angiographic evaluation and management of nonvariceal upper gastrointestinal bleeding.
Top three ST's: Diagnostic Procedure; Spatial Concept; Pathologic Function
Top three JD's: Gastroenterology; Cardiology; Vascular Diseases

Figure 18. ST and JD indexing of text discussing diagnostic technique in gastroenterology.

A possible application of automatic indexing would be the use of ST's in organizing retrieved output. One means to achieve this would be to rank or to cluster the retrieval in terms of the ST indexing. For example, a retrieval in the field of neoplasia (cancer) may be organized to first display results reflecting cancer management, in contrast to the literature reflecting cancer pathology or experimental research. Retrieved documents with high rankings involving the ST

ISSN: 2324-9773

Humphrey, S., Rindflesch, T., & Aronson, A. (2000). Automatic indexing by discipline and high-level categories: Methodology and potential applications. . 11th ASIS SIG/CR Classification Research Workshop, 103-116. doi:10.7152/acro.v11i1.12773

Title: Early and late complications after surgical gastric resection for peptic ulcer.
Top three ST's: Therapeutic or Preventive Procedure; Pathologic Function; Spatial Concept
Top three JD's: Gastroenterology; Surgery; Chemistry

Figure 19. ST and JD indexing of text discussing surgical intervention of gastrointestinal disease.

Title: Physician specialty and variations in the cost of treating patients with acute upper gasrointestinal bleeding.
Top three ST's: Health Care Activity; Health Care Related Organization; Research Activity
Top three JD's: Gastroenterology; Health Services Research; Public Health

Figure 20. ST and JD indexing of text discussing health care research in gastroenterology.

"Therapeutic or Preventive Procedure' combined with ST's such as 'Pharmacologic Substance' (for chemotherapy), 'Spatial Concept' (for surgery and radiotherapy), 'Quantitative Concept' (for results of treatment), and 'Medical Device' (for delivery of treating agents) may serve to bring out the cancer management aspect.

5. CONCLUSION

Many performance issues remain for the automated JD and ST indexing methodologies. These issues aside, we are particularly interested in discussing the potential usefulness and application of this indexing in retrieval. For the example of indexing a MEDLINE title and abstract (Figure 5 and Figure 12), to what extent are the highest ranked automatically assigned JD's ('Drug Therapy', 'Antibiotics', 'Pulmonary Disease (Specialty)', 'Communicable Diseases', 'Microbiology') and ST's ('Antibiotic', 'Clinical Drug', 'Laboratory Procedure', 'Pharmacologic Substance', 'Disease or Syndrome') useful for retrieval, and how can this general indexing be applied during the retrieval process?

6. REFERENCES

Aronson, A.R., Bodenreider, O., Chang, H.F., Humphrey, S.M., Mork, J.G., Nelson, S.J., Rindflesch, T.C., & Wilbur, W.J. (2000). The NLM Indexing Initiative. Submitted to *AMIA 2000 Annual Symposium*, Bethesda, MD: American Medical Informatics Association.

Humphrey, S.M. (1998). A new approach to automatic indexing using journal descriptors. In C.M. Preston (Ed.), ASIS '98, *Proceedings of the 61st ASIS Annual Meeting* (pp. 496-500), Medford, NJ: Information Today.

Humphrey, S.M. (1999). Automatic indexing of documents from journal descriptors: A preliminary investigation. *Journal of the American Society for Information Science*, 50, 661-674.

ISSN: 2324-9773

Humphrey, S., Rindflesch, T., & Aronson, A. (2000). Automatic indexing by discipline and high-level categories: Methodology and potential applications. . 11th ASIS SIG/CR Classification Research Workshop, 103-116. doi:10.7152/acro.v11i1.12773

National Library of Medicine (2000). *UMLS knowledge sources*. (11th ed.). Bethesda, MD.

Salton, G., & McGill, M.J. (1983). *Introduction to modern information retrieval*. (p. 124). New York: McGraw-Hill.

ISSN: 2324-9773