Semantic Indexing for a Complete Subject Discipline

Yi-Ming Chung, Qin He, Kevin Powell and Bruce Schatz
CANIS - Community Architectures for Network Information Systems
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign, Champaign, IL 61820
{chung, hqin, powell, schatz}@canis.uiuc.edu
http://www.canis.uiuc.edu

ABSTRACT

As part of the Illinois Digital Library Initiative (DLI) project we developed *"scalable semantics"* technologies. These statistical techniques enabled us to index large collections for deeper search than word matching. Through the auspices of the DARPA Information Management program, we are developing an integrated analysis environment, the Interspace Prototype, that uses *"semantic indexing"* as the foundation for supporting concept navigation. These semantic indexes record the contextual correlation of noun phrases, and are computed generically, independent of subject domain.

Using this technology, we were able to compute semantic indexes for a subject discipline. In particular, in the summer of 1998, we computed concept spaces for 9.3M MEDLINE bibliographic records from the National Library of Medicine (NLM) which extensively covered the biomedical literature for the period from 1966 to 1997. In this experiment, we first partitioned the collection into smaller collections (repositories) by subject, extracted noun phrases from titles and abstracts, then performed semantic indexing on these subcollections by creating a concept space for each repository. The computation required 2 days on a 128-node SGI/CRAY Origin 2000 at the National Center for Supercomputer Applications (NCSA). This experiment demonstrated the feasibility of scalable semantics techniques for large collections. With the rapid increase in computing power, we believe this indexing technology will shortly be feasible on personal computers.

KEYWORDS: Semantic Indexing, Semantic Retrieval, Concept Space, Scalable Semantics, Interspace, MEDSPACE, MEDLINE, Medical Informatics

1 INTRODUCTION

Our major research goal has been pursuing *"scalable semantics"*, a technology of indexing that would scale to large collections yet support deeper search than word matching [21]. We have been developing statistical algorithms for processing the term co-occurrences for disciplinary collections comprising millions of documents. These algorithms have been evolved to operate across different domains, with no special tuning required for each subject area.

Throughout the course of the Digital Libraries (DLI) project, we conducted several large-scale semantic indexing experiments for engineering disciplines [21][20][9]. The Compendex database was used to supply documents which provided broad coverage across all the engineering domains, with 800K abstracts chosen from 600 categories from the hierarchical subject classification. Documents from INSPEC were used to supply deep coverage for the core domains of physics, electrical engineering, and computer science, with 400K abstracts chosen from 300 categories. In the Engineering experiment, each abstract was classified into roughly 3 categories so there was overlap across repositories. This generated approximately 4M bibliographic abstracts across 1,000 repositories.

In the summer of 1998, we employed the semantic indexing technique to index the National Library of Medicine's (NLM) MEDLINE collection of 9.3 million bibliographic records covering from 1966 to 1997. This collection covers the entire medical discipline both in breadth and depth. In this experiment, we first partitioned the collection into subdisciplinary repositories and then performed semantic indexing on these sub-collections by creating a concept space for each repository. On average, each document was placed into 4.5 partitions. Thus, the final computation involved 40M abstracts across 10,000 partitions and required 50 hours computation time using a 128-node SGI/CRAY Origin 2000, NCSA's largest supercomputer array at that time.

The purpose of this experiment is to test the scalability of the semantic indexing techniques on a large scale collection with the ultimate goal to create an unique large-scale testbed for *"semantic interoperability"* [17]. To test semantic interoperability, we need semantic indexes for large collections across multiple subjects. The current semantic indexing technology has shown effective primarily for text documents. Obtaining a suitable collection of text documents with breadth across subjects and depth within subjects mandates bibliographic abstracts. So partitioning the MEDLINE collection into subject repositories and indexing these sub-collections with concept spaces can provide a large testbed for future semantic interoperability experiments.

This paper is organized as follows. Section 2 describes the semantic indexing techniques including noun phrase parsing techniques and concept space algorithm. Section 3 describes the characteristics of the MEDLINE collection and the MeSH Tree classification hierarchy. Section 4 discusses our partitioning strategy for the MEDLINE collection using the MeSH subject classification hierarchy and the semantic indexing computation experiment using NCSA's SGI/CRAY Origin 2000. Section 5 describes the potential usage of the MEDLINE indexing experiment results, called MEDSPACE, as a testbed in the Interspace Prototype. Finally, Section 6 describes the future directions.

2 SEMANTIC INDEXING: CONCEPTS FROM CONTEXT USING CO-OCCURRENCE ANALYSIS

The concept space algorithm has been used in numerous experiments to generate and integrate multiple semantic indexes [6][9]. Previous experiments have shown that creating domain independent semantic indexes automatically is feasible and that these indexes can pave the way for cross-domain information retrieval [8].

The concept space algorithm is based upon statistical correlations of the context within documents. To create a concept space, first find the context of terms or phrases within documents using a noun phrase parser and then compute term (noun phrase) relationships using co-occurrence analysis. The algorithms used by this experiment are briefly summarized below.

2.1 Noun Phrase Extraction

AZ Noun Phraser [22], the noun phrase extractor developed in collaboration with the AI (Artificial Intelligence) group at the Department of Management Information Systems of University of Arizona, was used for this experiment. The phraser is based upon the Brill tagger [1][2] and the noun phrase identification rules of NPtool [23]. The program was chosen since it is generic–the trained lexicon was derived from several different sources including the Wall Street Journal and Brown corpora, hence the lexicon has a fairly general coverage of the English language. It can be applied across domains without further domain customization while maintaining a comparable parsing quality. The phraser operates in three distinct phases: tokenization, part-of-speech tagging and noun phrase identification.

2.1.1 Tokenization Text must be tokenized in order to be parsed correctly by the noun phrase parser. The goal of tokenization is to determine sentence boundaries and separate the text into individual tokens by removing irrelevant punctuation. Tokens are strings of characters separated by whitespace. Punctuation characters such as ", ; : ." are treated specially in the tokenization process. Since these characters typically signal noun phrase boundaries, they are considered separate tokens. The output of the tokenization phase is a list of tokens which will be analyzed by the tagger.

2.1.2 Part-of-Speech Tagging The Part-of-Speech tagger is internally divided into two phases of operation: lexical analysis and contextual analysis. The first phase involves looking up each token in a trained lexicon of commonly used English words. Each word (entry) in the lexicon is marked with all of its possible parts of speech. If a token (word) is found in the lexicon, the first part of speech of this token is then assigned to it. If a token does not appear in the lexicon, the tagger marks it as an unknown noun. Lexical rules are then used by the tagger to choose the closest matching part of speech for these unknown tokens.

At this point in the process, the exact part of speech of each token is still in question. To resolve this, the Brill tagger uses a number of contextual rules to disambiguate the term's part of speech. For each token, these rules examine the tokens immediately preceding and following the current token. With this information the tagger is finally able to determine the best part of speech for each token and eventually create a list of tagged tokens.

2.1.3 Noun Phrase Identification Noun phrases are recognized by a set of patterns, or rules, that are composed of different orderings of parts of speech. For our experiment, the limit for the longest noun phrase pattern recognizable is seven words in length. The shortest pattern is a noun phrase of length one (e.g. just the noun itself).

The identification of noun phrases is realized by moving a sliding window through the tagged token list. The window starts from the head of the list with a size of seven tokens, i.e., the size of the longest noun phrase recognizable. If the content of the window, the parts of speech of seven tokens, matches one of the noun phrase rules, a noun phrase is identified. If the content does not match any of the rules, the last token in the window will be truncated and the remaining content is compared to the noun phrase rules again. This process is repeated until the window content matches a rule, which could be a single word at the end. The window will then move on and start from the token following the noun phrase just identified, with a size reset to seven. Tokens for characters such as ", ; : ." and tokens tagged as verb are treated as noun phrase delimiters. They will truncate the window when they are encountered.

There are a number of limitations of the noun phraser in the current implementation. In scientific literature noun phrases of greater than seven words in length appear to be more common than in general text. This results in the tagger misidentifying long noun phrases as two separate shorter phrases. Additionally, the tagger cannot effectively differentiate between context sensitive noun phrases, which only have meaning in the context of use, and more general forms that have a useful degree of context free meaning. Finally, this type of noun phrase identification is computationally expensive when compared with more ad hoc techniques that rely solely on tokenization and concatenation of adjacent tokens to identify phrases.

The AZ Noun Phraser can also be customized to a particular domain by training the Brill tagger, editing stop words list or incorporating a domain specialized lexicon, e.g. SPECIALIST Lexicon from the Unified Medical Language System (UMLS) by NLM. According to a study conducted by by our collaborators, a SPECIALIST Lexicon enhanced Noun Phraser performed slightly better than the generic version on a collection of 630K CancerLit abstracts but the difference is not statistically significant [22]. The generic nature and ease of customization enable the parser to fulfill the range of noun phrase parsing needs.

2.2 Co-occurrence Analysis

During the noun phrase extraction process, we also keep track of noun phrase frequency information, which is used to compute weights for each noun phrase in the documents. The weight of noun phrase j in document i, d_{ij}, is computed based on the product of "noun phrase frequency" and "inverse document frequency" (a common technique adopted in vector space models of IR). Noun phrase frequency, tf_{ij}, represents the number of occurrences of noun phrase j in document i. Document frequency, df_j, represents the number of documents in a collection in which noun phrase j occurs. This is represented by equation (1).

$$d_{ij} = tf_{ij} \times log(\frac{N}{df_j} \times w_j) \tag{1}$$

where *N* represents the total number of documents in a collection and w_j represents the number of tokens (words) in noun phrase j. Multiple-word terms are assigned heavier weights than single-word terms because multiple-word terms usually convey more precise semantic meaning than single-word terms.

The co-occurrence analysis is computed based on an asymmetric similarity function as follows:

$$Weight(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \times WeightingFactor(Tk) \tag{2}$$

The above equation indicates the similarity weight from term T_j to term T_k. d_{ij} is calculated by equation (1). d_{ijk} represents the combined weight of both term descriptors T_j and T_k in document i and is defined as follows:

$$d_{ijk} = tf_{ijk} \times \log{(\frac{N}{df_{jk}} \times w_j)} \tag{3}$$

where tf_{ijk} represents the smaller number of occurrences of term j and term k in document i. df_{jk} represents the number of documents in which terms j and k occur together. w_j represents the number of words in descriptor T_j.

A weighting factor is also used to further penalize high collection frequency terms in the co-occurrence analysis:

$$WeightingFactor(T_k) = \frac{\log \frac{N}{df_k}}{\log N} \tag{4}$$

Terms with a higher document frequency value, df_k, (possibly more general terms) have a smaller weighting factor value, which causes the co-occurrence probability to become smaller.

The resulting co-occurrence matrix represents a network of noun phrases and their probabilistic relationships. The relationships between noun phrases reflect the strengths of their context associations within a collection.

3 MEDLINE: A COMPLETE DISCIPLINE COLLECTION

MEDLINE is a medical bibliographic database maintained by the National Library of Medicine (NLM). It covers the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and pre-clinical science. It contains bibliographic citations from over 3900 biomedical journals published in the United States and 70 foreign countries. In early 1998, we acquired all the bibliographic records in MEDLINE and its backup databases, BACKFILES, covering from 1966 to 1997. The size, 9.3 million abstracts, and the coverage of the corpus is comprehensive for the subject discipline of biomedicine.

3.1 Indexing MEDLINE

Our first task toward indexing this large medical corpus was to partition it into segments by subject domains. The goal was to partition the corpus into a set of domains which would allow users to navigate across these domains.

We investigated several partitioning strategies, based on standard clustering algorithms (a completely automated approach) and on partitions derived from human assigned subject headings (a semi-automatic approach requiring subject categorization). For example, one simple approach is to partition the collection based upon the top (one thousand) most frequently occurring terms. This approach, while straightforward, produces a semantically weak set of clusters, i.e. the clusters produced are highly variable in their meaningfulness. In contrast, human generated classification and thesaurus systems have better properties semantically and can be used to produce more meaningful clusters.

We also ruled out the possibility of using automatic clustering algorithms in this particular experiment. Most of the clustering algorithms are computation extensive. With the collection size of 9.3M, the clustering task itself would not be computationally feasible within a reasonable time frame. These led us to adopt the semi-automatic approach based on an existing human classification scheme. The NLM's Medical Subject Headings (MeSH) was chosen in this experiment.

3.2 Medical Subject Headings (MeSH)

MeSH consists of a set of terms or subject headings that are arranged in both an alphabetic and a hierarchical structure known as MeSH Tree structure. It has the properties of a thesaurus and a classification system. The hierarchical structure helps users browsing the thesaurus to navigate from a broader concept, a broader term (BT), to narrow concepts, narrow terms (NT) and vice versa. This hierarchical structure is also useful as an aid for retrieval. For example, a user can specify the retrieval of all citations indexed to a particular heading, as well as those indexed to all the narrow terms of the heading.

The 1998 MeSH includes 32284 tree nodes in the hierarchical structure and 18934 main headings, among which 18849 headings were used as labels to MeSH Tree nodes. In the hierarchical tree structure, tree nodes are identified by a unique tree number named MeSH Tree Number (MN), an alphanumerical hierarchical representation of the terms in MeSH. The top level has 15 broad categories and each category can be up to nine levels in depth. Each tree node corresponds to a main heading. As a result, some main headings have duplicate occurrences in the tree structure.

To better understand MeSH subject headings and MeSH Tree structure, we use the MeSH heading "*Suppuration*" as an example to demonstrate the relationship between MeSH headings and MeSH Tree. Figure 1 shows the MeSH entry of *Suppuration*. The record shows *Suppuration* is a main heading (MH) and the heading is assigned with two MeSH numbers (MN): C01.539.830 and C23.739.487.85. The annotation (AN) defines suppuration as "a type of abscess; NIM; TN 178: for suppurative dis & coord."

The MeSH Tree numbers (MN) allow us to browse the broad terms and narrow terms in the tree hierarchy. Since *Suppuration* is assigned with two tree numbers, we examine the both hierarchical branches in parallel. Figure 2 shows *Suppuration*: C01.539.830 is branched from C01 (Bacterial Infections and Mycoses) and has 6 sub-nodes. In contrast, C23.739.487.856, is branched from C23 (Symptoms and General Pathology) and has 3 sub-nodes.

Strictly speaking, since both tree nodes use the same MeSH heading, *Suppuration*, they should represent the same "concept." However, the tree structure allows users to conceive

```
MH = Suppuration
MN = C01.539.830
MN = C23.739.487.856
AQ = BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM
     ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH TM UR US
     VE VI
AN = a type of abscess; NIM; TN 178: for suppurative dis
     & coord
MS = The formation of pus. (Dorland, 27th ed)
MR = 940527
DC = 1
UI = D013492
```

Figure 1: MeSH entry of *Suppuration*: MH: MeSH Main Heading; MN: MeSH Tree Number; AQ: Allowable Topical Qualifiers; AN: Annotation; MS: MeSH Scope Note; MR: Major Revision Date; DC: Descriptor Class; UI: Unique Identifier. For a detailed description of the fields, please refer to *Medical Subject Headings, Annotated Alphabetic List* and other printed publications via ftp from the host nlmpubs.nlm.hin.gov in the directory online/medlars/manuals.

```
Bacterial Infections and Mycoses;C01
  Infection;C01.539
    Suppuration;C01.539.830
      Abscess;C01.539.830.025
        Abdominal Abscess;C01.539.830.025.020
          Liver Abscess;C01.539.830.025.020.455
            Liver Abscess, Amebic;
            C01.539.830.025.020.455.460
          Subphrenic Abscess;C01.539.830.025.020.810
        Brain Abscess;C01.539.830.025.160
        Lung Abscess;C01.539.830.025.490
        Periapical Abscess;C01.539.830.025.650
        Periodontal Abscess;C01.539.830.025.665
        Peritonsillar Abscess;C01.539.830.025.675
        Psoas Abscess;C01.539.830.025.700
        Retropharyngeal Abscess;C01.539.830.025.780
      Cellulitis;C01.539.830.200
      Empyema;C01.539.830.305
        Empyema, Pleural;C01.539.830.305.310
          Empyema, Tuberculous;C01.539.830.305.310.320
        Empyema, Subdural;C01.539.830.305.330
      Otitis Media, Suppurative;C01.539.830.694
      Thyroiditis, Suppurative;C01.539.830.840
      Uveitis, Suppurative;C01.539.830.900
```

```
Symptoms and General Pathology;C23
  Pathologic Processes;C23.739
    Inflammation;C23.739.487
      Suppuration;C23.739.487.856
        Abscess;C23.739.487.856.025
        Cellulitis;C23.739.487.856.200
        Empyema;C23.739.487.856.305
```

Figure 2: The C01.539.830 and C23.739.487.856 subtrees of *Suppuration*

the particular concept in different contexts (branches) of the concept hierarchy.

4 MEDSPACE: SEMANTIC INDEXING EXPERIMENT FOR A MEDICAL DISCIPLINE

4.1 Domain Partitions

As mentioned earlier, the current partitioning of MEDLINE is based on MeSH headings, by using the subject hierarchy to specify the collections for the sub-disciplinary repositories. Each MEDLINE bibliographic record has an average of 12.36 MeSH headings assigned by the professional indexers at NLM. Of these headings, an average of 4.6 MeSH terms are indicated by the indexer as main concepts of the article.

Main concepts are those concepts (MeSH headings) which best describe the article. These main concepts were used to determine into which collections to place the particular abstract. Since some abstracts do not have main concepts assigned to them, about 1.4% of abstracts are dropped from the final partitions. Also, since each abstract has an average of 4.6 placements in the partitioned repositories, the multiple placements of each abstract caused an expansion of about 4.6 times which resulted in a total of 46,733,751 repository abstracts computed from raw 9,315,615 abstracts.

We also used the above results to create an "inclusive" partition set. Contrary to the above partition scheme, all the abstracts in the narrow-term nodes were propagated to their parent node. We refer the former partition scheme as the "exclusive" partition set and this scheme as the "inclusive" partition set. The inclusive partitioning caused an expansion of the data by a factor of about 19 and produced 178+ million of repository abstracts.

Table 1 shows both of the top 10 largest domains in the inclusive and exclusive partition sets. In the exclusive set, the largest domain is *Liver* with 123,082 abstracts while the largest domain in the inclusive set is *Amino, Acid, Peptides, and Proteins* with 1,373,592 abstracts.

Table 1: Top 10 inclusive and exclusive domains

Inclusive partition		Exclusive Partition	
Domain	# of abs	Domain	# of abs
Amino Acids, Peptides, and Proteins	1373592	Liver	123082
Proteins	1169296	Brain	105614
Neoplasms	910074	Kidney	77040
Organic Chemicals	886533	Neoplasms	69721
Immunologic and Biological Factors	780897	Hypertension	66907
Cells	777939	Muscles	65341
Symptoms and General Pathology	678656	Escherichia coli	64652
Cardiovascular Diseases	655243	Calcium	63467
Diagnosis	606924	DNA	58535
Enzymes, Coenzymes, and Enzyme Inhibitors	602995	Coronary Disease	57662

In our experiment, the exclusive partition set was used to build the medical disciplinary repositories. Table 2 shows the statistics of the exclusive domains' collection sizes. Since the partitioning was based on the MeSH Tree hierarchical structure, it created 32,284 domains in total, including the domains with a collection size of zero. We computed concept spaces for each exclusive domain which had a collection size greater than 1000 abstracts. This resulted in 9894 disciplinary repositories.

Table 2: MEDLINE partition results

collection size	Num of Domains	Accum. Total
10,000+	748	748
5,000 - 9,999	1,407	2,155
1,000 - 4,999	7,739	**9,894**
500 - 999	4,592	14,486
100 - 499	9,724	24,210
1 - 99	7,053	31,263
= 0	1,021	32,284

The final computation involved 45,444,790 unique concepts (noun phrases) and about 19 billion concept co-occurrences within 40,628,964 abstracts of the 9894 repository collections. The rest of the small domains were not computed in this experiment based on the consideration that these domains were too small to be useful, even thought it would only take few seconds for each to be computed. In the inclusive partitions, these small domains were merged to their parent node.

4.2 Parallel Computing

The SGI/CRAY Origin 2000 by Silicon Graphics is a scalable shared memory multiprocessor (S2MP) designed to provide the benefits of both a shared memory multiprocessor approach and a distributed memory message-passing multiprocessor approach. The architecture uses physically distributed memories but treats them as a unified, global address space. This design preserves the benefits of ease of programming on a shared memory architecture without sacrificing the performance scalability of a distributed memory architecture.

The largest NCSA Origin 2000 machine available in March 1998 was used in this experiment. It had 128 processing nodes, 64GB of memory and 420GB scratch disk. Each processor was a 195 MHz MIPS R10000 processor with a peak performance of 390 MFLOPS.

Table 3 shows concept space execution times on the Origin 2000. In these experiments, we performed noun phrase extraction and co-occurrence analysis for 5 different collections varying in size from 1K to 100K abstracts. The maximum number of processors used in these benchmarks was 32.

In the final computation experiment, all 128 processors were

Table 3: Benchmark performance on Origin 2000

	NP (h:m:s)	Co-occurrence Analysis (h:m:s)					
Collection size	1 node	1 node	2 nodes	4 nodes	8 nodes	16 nodes	32 nodes
1K	0:57	0:03	0:02	0:02	0:01	0:01	0:01
5K	5:23	0:52	0:26	0:16	0:09	0:07	0:08
10K	11:39	1:59	0:57	0:33	0:19	0:14	0:15
50K	1:06:38	24:11	10:40	6:00	3:41	2:17	1:18
100K	2:28:16	1:02:58	27:02	15:10	9:04	5:45	3:07

used and the whole computation finished in about 2 days, i.e. 50 hours of dedicated execution time. The first 16 hours were used to extract noun phrases for the 9.3M abstracts, partition noun phrases to all of the 32,284 exclusive partition domains, and perform some housekeeping tasks such as preparing input data, deleting intermediate files and setting up execution scripts.

The rest of the execution involved: tracking term frequency information; applying noun phrase selection using frequency information; computing co-occurrence matrices; creating outputs files. Since all the programs were bundled together, we were unable to separate the execution time by stages or domains. Several extra outputs were created during the same run in order for these concept spaces to be imported into the full-scale Interspace Prototype (see Section 5).

This MEDLINE computation was an order of magnitude bigger than the Compendex computation (40M versus 4M abstracts). The actual computation time required, however, was about the same scale - 10 days for test debugging and 2 days for final production. This was possible because the intervening 2-year period since the last experiment had made the high-end NCSA supercomputer an order of magnitude better. The 128-node, 64GB SGI/Cray Origin 2000 had 4 times more processors, and the faster processors and bigger memories combined with optimized parallel algorithms further improved the performance.

5 INTERSPACE PROTOTYPE

5.1 Concept Spaces and Concept Switching

The Interspace Prototype [4], a research effort in the DARPA Information Management program being developed at CANIS, is an analysis environment for semantic indexing of multimedia information in a testbed of real collections. It is an integrated information analysis environment which allows interactive concept navigation based on semantic indexing and semantic clustering.

Part of the MEDLINE concept spaces have been incorporated into the analysis environment of the Interspace Prototype. Figure 3 gives an example of our MEDLINE concept spaces with the early demonstration client of the analysis environment. The subject domains “Colorectal Neoplasms, Hereditary Nonpolyposis” and “Genes, Regulator” were chosen and their concept spaces were displayed in the middle and the right respectively. “Hereditary cancer” was entered as a search term in the first concept space and all concepts which are lexical permutations are returned. Indented levels in the display indicate the hierarchy of the co-occurrence list. Navigating in the concept space moves from “hereditary nonpolyposis colorectal cancer” to the related “mismatch repair genes”.

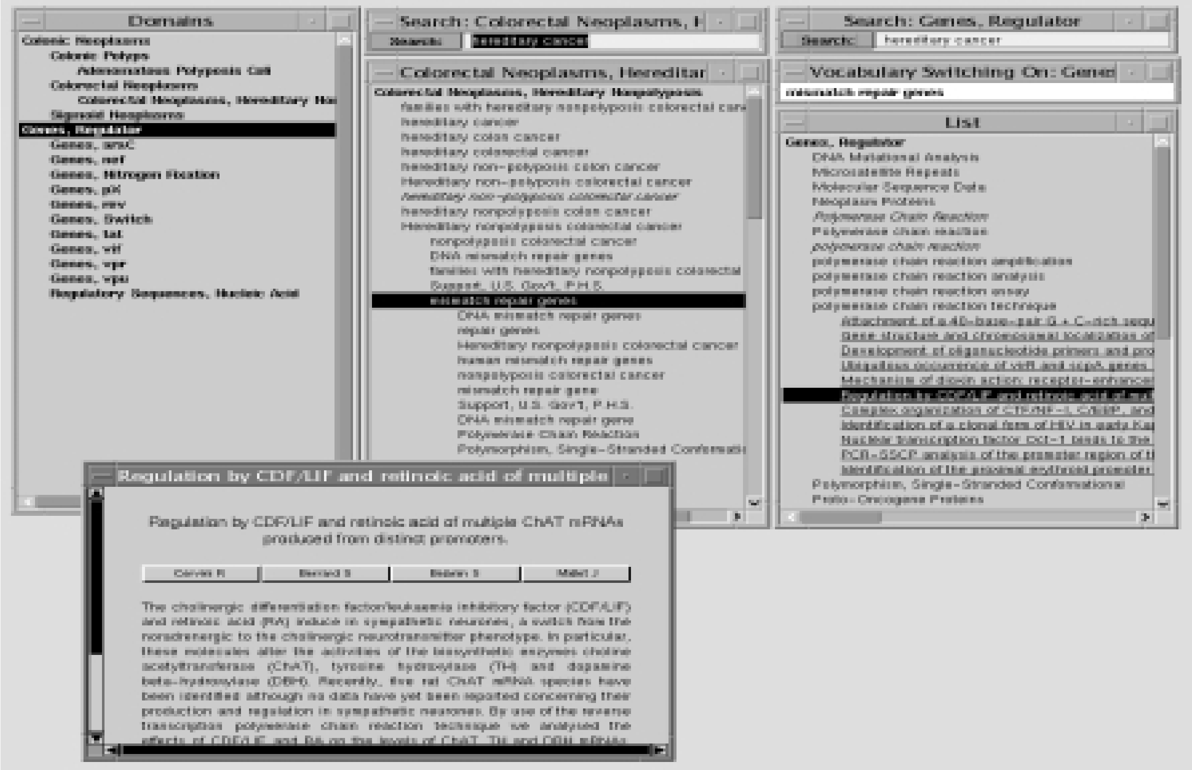
The user then tries to search for this desired term in another domain repository, “Genes, Regulator”. A straight text search at top right returns no hits. So *“Concept Switching”* is invoked to switch concepts from one domain to another across their respective concept spaces. The concept switch takes the term “mismatch repair genes” and all related terms from its indented co-occurrence list in the source concept space for “Colorectal Neoplasms” and intersects this set into the target concept space for “Genes, Regulator”.

After syntactic transformations, the concept switch produces the list (in the right most window panel) of concepts computed to be semantically equivalent to “mismatch repair genes” within “Genes, Regulator”. Navigating the concept space down to the object (document) level locates the article displayed at the bottom. This article discusses a leukaemia inhibitory factor which is related to colon cancer. Note that this article was located without doing a search, by concept switching across repositories starting with the broad term “hereditary cancer”.

5.2 Interspace Analysis Environment

Figure 4 shows the current stage of the Interspace Prototype analysis environment. The interface is based upon a custom user interface toolkit called Morphic implemented in Smalltalk. The Interspace environment consists of multiple levels of abstraction: the category, the concept space, and document space. It enables navigation across multiple levels of indexes in a seamless navigation session, e.g., the spaces can be navigated from concept to concept without the need for text search.

The sample session begins within the subject domain of Rheumatoid Arthritis, one of the community repositories from the



Domains

Colonic Neoplasms
Colonic Polyps
Adenomatous Polyposis Coli
Colorectal Neoplasms
Colorectal Neoplasms, Hereditary Nonpolyposis
Sigmoid Neoplasms
Genes, Regulator
Genes, arsC
Genes, nef
Genes, Nitrogen Fixation
Genes, pX
Genes, rev
Genes, Switch
Genes, tat
Genes, vif
Genes, vpr
Genes, vpu
Regulatory Sequences, Nucleic Acid

Search: Colorectal Neoplasms, H

Search: hereditary cancer

Colorectal Neoplasms, Hereditar

Colorectal Neoplasms, Hereditary Nonpolyposis
families with hereditary nonpolyposis colorectal can
hereditary cancer
hereditary colon cancer
hereditary colorectal cancer
hereditary non-polyposis colon cancer
Hereditary non-polyposis colorectal cancer
hereditary non-polyposis colorectal cancer
hereditary nonpolyposis colon cancer
Hereditary nonpolyposis colorectal cancer
nonpolyposis colorectal cancer
DNA mismatch repair genes
families with hereditary nonpolyposis colorectal
Support, U.S. Gov't, P.H.S.
mismatch repair genes
DNA mismatch repair genes
repair genes
Hereditary nonpolyposis colorectal cancer
human mismatch repair genes
nonpolyposis colorectal cancer
mismatch repair gene
Support, U.S. Gov't, P.H.S.
DNA mismatch repair gene
Polymerase Chain Reaction
Polymorphism, Single-Stranded Conformati

Search: Genes, Regulator

Search: hereditary cancer

Vocabulary Switching On: Genes

mismatch repair genes

List

Genes, Regulator
DNA Mutational Analysis
Microsatellite Repeats
Molecular Sequence Data
Neoplasm Proteins
Polymerase Chain Reaction
Polymerase chain reaction
polymerase chain reaction
polymerase chain reaction amplification
polymerase chain reaction analysis
polymerase chain reaction assay
polymerase chain reaction technique
Attachment of a 40-base-pair G + C-rich sequ
Gene structure and chromosomal localization of
Development of oligonucleotide primers and pro
Ubiquitous occurrence of virB and vapA genes
Mechanism of dioxin action: receptor-enhancer
Regulation by CDF/LIF and retinoic acid of mult
Complex organization of CTF/NF-I, C/EBP, and
Identification of a clonal form of HIV in early Ka
Nuclear transcription factor Oct-1 binds to the
PCR-SSCP analysis of the promoter region of t
Identification of the proximal erythroid promoter
Polymorphism, Single-Stranded Conformational
Proto-Oncogene Proteins

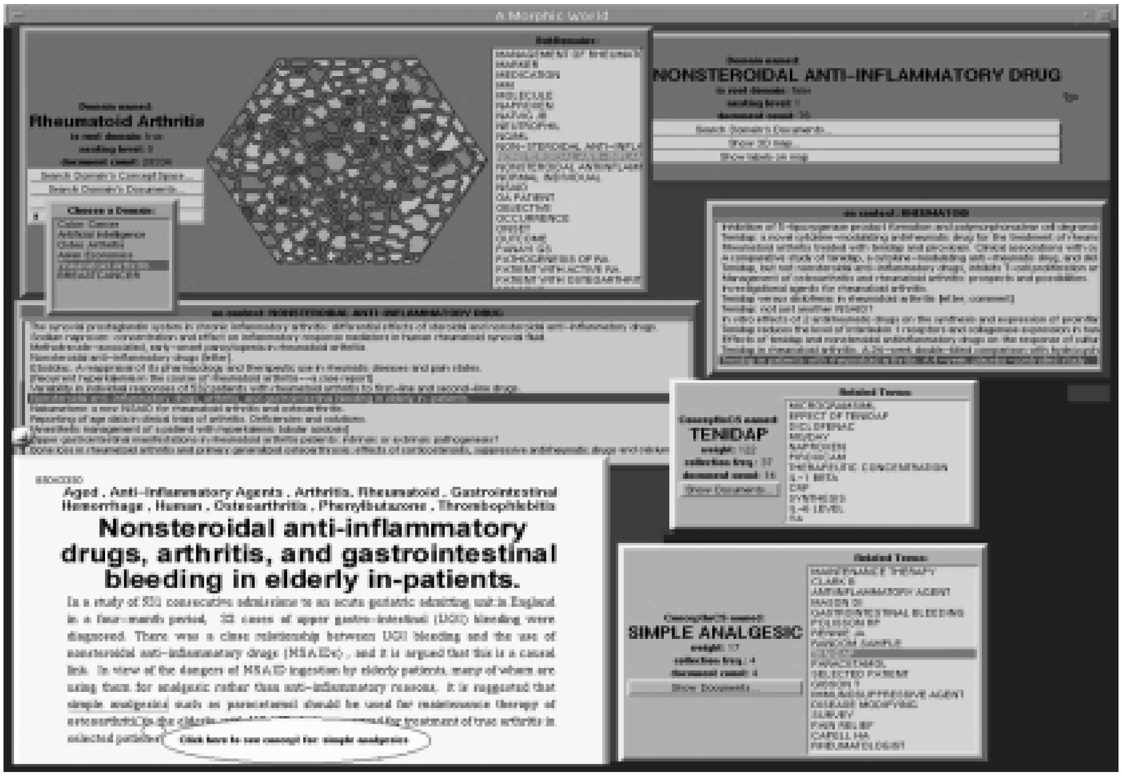
Regulation by CDF/LIF and retinoic acid of multiple

Regulation by CDF/LIF and retinoic acid of multiple ChAT mRNAs produced from distinct promoters.

Cervini R | Berrard S | Bejanin S | Mallet J

The cholinergic differentiation factor/leukaemia inhibitory factor (CDF/LIF) and retinoic acid (RA) induce in sympathetic neurones, a switch from the noradrenergic to the cholinergic neurotransmitter phenotype. In particular, these molecules alter the activities of the biosynthetic enzymes choline acetyltransferase (ChAT), tyrosine hydroxylase (TH) and dopamine beta-hydroxylase (DBH). Recently, five rat ChAT mRNA species have been identified although no data have yet been reported concerning their production and regulation in sympathetic neurones. By use of the reverse transcription polymerase chain reaction technique we analysed the effects of CDF/LIF and RA on the levels of ChAT, TH and DBH mRNAs

Figure 3: Concept Switching in MEDLINE



A Morphic World

Domain named: Rheumatoid Arthritis
is root domain: true
nesting level: 0
document count: [illegible]

Search Domain's Concept Space...
Search Domain's Documents...

Choose a Domain:
Colon Cancer
Artificial Intelligence
Osteo Arthritis
Asian Economics
Rheumatoid Arthritis
BREAST CANCER

Subdomains:
MANAGEMENT OF RHEUMATOI
MARKER
MEDICATION
MM
MOLECULE
NAPROXEN
NATAL [illegible]
NEUTROPHIL
NORMAL
NON-STEROIDAL ANTI-INFLA
[illegible]
NONSTEROIDAL ANTIINFLAM
NORMAL INDIVIDUAL
NSAID
OA PATIENT
OBJECTIVE
OCCURRENCE
ONSET
OUTCOME
[illegible]
PATHOGENESIS OF RA
PATIENT WITH ACTIVE RA
PATIENT WITH OSTEOARTHRI

Domain named: NONSTEROIDAL ANTI-INFLAMMATORY DRUG
is root domain: false
nesting level: 1
document count: 35

Search Domain's Documents...
Show 2D map...
Show labels on map

as context: RHEUMATOID

[illegible]
Tenidap versus diclofenac in rheumatoid arthritis [letter; comment]
Tenidap: not just another NSAID?
[illegible]

as context: NONSTEROIDAL ANTI-INFLAMMATORY DRUG

[illegible]
Nonsteroidal anti-inflammatory drugs [letter].
Etodolac. A reappraisal of its pharmacology and therapeutic use in rheumatic diseases and pain states.
Recurrent hyperkalaemia in the course of rheumatoid arthritis—a case report
Variability in individual responses of 532 patients with rheumatoid arthritis to first-line and second-line drugs.
[illegible]
Nabumetone: a new NSAID for rheumatoid arthritis and osteoarthritis.
Reporting of age data in clinical trials of arthritis. Deficiencies and solutions.
[illegible]

Concept(s) named: TENIDAP
weight: 122
collection freq: 37
document count: 16
Show Documents...

Related Terms:
MICROGRAM/ML
EFFECT OF TENIDAP
DICLOFENAC
MG/DAY
NAPROXEN
PIROXICAM
THERAPEUTIC CONCENTRATION
IL-1 BETA
CRP
SYNTHESIS
IL-6 LEVEL
RA

Concept(s) named: SIMPLE ANALGESIC
weight: 17
collection freq: 4
document count: 4
Show Documents...

Related Terms:
MAINTENANCE THERAPY
CLARK B
ANTIINFLAMMATORY AGENT
MASON [illegible]
GASTROINTESTINAL BLEEDING
POLISSON RP
PENNIE [illegible]
RANDOM SAMPLE
[illegible]
PARACETAMOL
SELECTED PATIENT
GRIFFIN [illegible]
IMMUNOSUPPRESSIVE AGENT
DISEASE MODIFYING
SURVEY
PAIN RELIEF
CAPELL HA
RHEUMATOLOGIST

[illegible]

Aged . Anti-Inflammatory Agents . Arthritis, Rheumatoid . Gastrointestinal Hemorrhage . Human . Osteoarthritis . Phenylbutazone . Thrombophlebitis

Nonsteroidal anti-inflammatory drugs, arthritis, and gastrointestinal bleeding in elderly in-patients.

In a study of 531 consecutive admissions to an acute geriatric admitting unit in England in a four-month period, 33 cases of upper gastro-intestinal (UGI) bleeding were diagnosed. There was a close relationship between UGI bleeding and the use of nonsteroidal anti-inflammatory drugs (NSAIDs), and it is argued that this is a causal link. In view of the dangers of NSAID ingestion by elderly patients, many of whom are using them for analgesic rather than anti-inflammatory reasons, it is suggested that simple analgesics such as paracetamol should be used for maintenance therapy of osteoarthritis in the elderly [illegible] for treatment of true arthritis in selected patients.

Click here to see concept for simple analgesics

Figure 4: Smalltalk-Morphic based user interface of the Interspace Prototype system

partitioning of MEDLINE. The window in the upper left shows a category map of the documents in the repository, where each area contains a particular subdomain (sub-collection of the community collection). As shown in the upper right window, the subdomain “Nonsteroidal Anti-Inflammatory Drug” (NSAID) is selected. The documents within this subdomain are displayed in summary form in the middle left window. One such document is selected and displayed in full in the bottom left window.

Although the document appears to be a text abstract, it is actually represented as a sequence of objects, each of which is a concept in the space. As the user moves the mouse cursor over the “text”, the corresponding concepts are highlighted within a pop-up bubble. Here the concept “simple analgesics” is selected, with the bubble appearing at the bottom of the screen. Within the integrated analysis environment, a selected concept can simply be navigated into a different semantic index. Here the related concepts (in the concept space) are navigated from “simple analgesic” to “tenidap” in the windows at the bottom right and middle right respectively. The list of documents containing the concept “tenidap” is displayed in the middle right. The user has located a document describing a new type of NSAID named tenidap, without initially knowing its existence or ever needing to issue a search.

6 FUTURE DIRECTIONS

The resulting concept spaces of the disciplinary repositories from the MEDLINE experiment will be gradually incorporated into the Interspace prototype to serve as a large testbed for real users. To identify illustrative uses of the Interspace Prototype, it is being tested by a few early adopters, who are practicing physicians. We are currently evaluating the system utility versus their information needs. Our plan over the next several years is to evaluate an evolved production version of the MEDSPACE (Medical Interspace) for a large user group. The target user subjects include medical students and practitioners across Illinois.

To assist users in concept navigation across multiple spaces, concept switching capabilities are necessary. We are investigating several concept association techniques, mostly machine learning techniques, to support *concept-to-concept* mapping across spaces. The hypothesis is that such techniques should be more effective than the syntactic transformations of earlier prototypes, e.g., Figure 3. *Concept*, therefore, will be represented by a symbolic label which groups entire equivalence classes of semantically related terms. The key tasks of mapping will be to identify the “semantic” region of the given query term(s) in one space and to “map” the region into a “equivalent” region in another domain.

One of the promising techniques is the Hopfield Network algorithm[13]. We have used this neural net algorithm in several semantic indexing experiments [7][12]. Our hypothesis is that the parallel relaxation property of a Hopfield net can be used to activate related terms across concept spaces, as a form of spreading activation of concepts. A previous experiment showed the algorithm is promising for domain switching tasks[11]. The MEDSPACE will provide a large scale testbed for future experiments in concept switching.

Several new algorithms and system improvements are currently under development for the MEDSPACE testbed. These future researches can be categorized as: noun phrasing parsing, domain partition and distributed computation.

6.1 Noun Phrase Parsing

Several techniques are under development to improve the noun phrase parsing quality, including term normalization techniques and a noun phrase weighting scheme. The goal of the noun phrase weighting module is to assign lower weights to noun phrases composed of common English words such that the system can discriminate specific noun phrases from general noun phrases.

Another continuing effort in the Interspace project has been to extract the conceptual structures of the vocabulary used in a particular corpus of documents. These conceptual structures ideally should be independent of the underlying terminology used in the documents. One approach to approximate the conceptual structure of the vocabulary is to extract noun phrases from the documents in the collection (as discussed above). Another, is to use a name finding system to enhance noun phrase parsing since names are the major practical use of concept spaces, to find specific terms for effective search. Finally, by using term conflation, the system can map multiple specific terms into a single broader one. The details of the second and third methods follow.

A potential enhancement to the process of concept space creation would be the identification of the names of people and things in the underlying source documents. A number of different approaches exist that attempt to extract arbitrary names from text. Some use elaborate heuristic rules to determine if the term in question is a name, while other researchers are working on systems which use dynamic techniques to discover name-like patterns in the text. Using such a name identification system deepens the conceptual knowledge used in concept space generation as well as enhancing other elements of the Interspace system.

Noun phrase extraction with simple normalizations are one means of getting at the underlying concepts that compose a sentence’s meaning. This technique suffers from being too precise, i.e. the majority of noun phrases only occur once in a given set of documents. One way of alleviating this problem involves a process known as term conflation, a technique whereby noun phrases with different constituent terms are mapped into one broader term. Based on the work in [15], we are incorporating part of the metarules into our noun phrases extracting system to map highly specific terms into broader ones. This serves to make the concept space calculation less intensive and provides for a richer set of relationships among the vocabulary extracted from the documents.

6.2 Domain Partition

The current MEDLINE partitioning was based on MeSH headings. Our next partition approach will be based on SNOMED, the Systematized Nomenclature of Human Medicine, which is a subject classification hierarchy specifically for clinical medicine. In the studies of the Large-Scale Vocabulary Test on the Unified Medical Language System (UMLS) at NLM [14], and the CPRI Work Group on Codes and Structures [3], SNOMED performed best of the source vocabularies and was found to have the richest clinical taxonomy. We will use the UMLS Metathesaurus to map MeSH subject headings to SNOMED categories in the Metathesaurus and use SNOMED categories to partition the MEDLINE collection.

In the situations where the target corpus lacks human generated subject classification systems, the subject domains partition will have to rely on completely automatic approaches. We are investigating potential solutions using term discrimination models, topic spotting methods, text categorization and clustering algorithms. The intuition is that by selecting the most discriminated terms or the major topics within a corpus, we will be able to assign documents to these terms or topics to form partitions. Text classifying and clustering algorithms can also achieve the same purpose. We chose to examine the term discrimination model before other approaches due to its relatively low computational cost. Other than medium frequency terms as suggested by Salton [18][19], we are looking into different algorithms of selecting the most discriminating terms to improve clustering performance.

Another approach we are interested in is Kohonen's self organizing feature map (SOM) [16]. We have generated SOMs at different resolutions for 500K MEDLINE abstracts. The major advantages of this particular method are its ability to construct a hierarchical topical structure by creating a multilayer category map and its visualization effects [10][6]. These enable document space navigation and enhance fine grain searching.

6.3 Distributed Computation

All of our large scale semantic indexing experiments have been performed on high-end supercomputers. Since supercomputers are not available to most researchers, we are investigating the feasibility of conducting large scale experiments on a group of PCs or workstations.

Recently, we have implemented the concept space algorithm and Interspace Prototype on a network of Sun workstations based on a distributed computing environment design and conducted a series of experiments on different collection sizes [5]. The Sun cluster consists of one Ultra 2 Model 1200 with 896MB memory and four Ultra 1 Model 170 with 128MB memory. The workstations were connected by a 100Mbs Ethernet network. Using this network of five workstations, we were able to compute a collection of about 10K documents within three hours. Note that unlike the supercomputer run which simply computed concept spaces' co-occurrence matrices and stored data in a flat file format, the complete Interspace Prototype involves creating persistent objects in a database environment in order to support the Interspace analysis client. This preliminary study shows that community-sized collections of 10K to 100K documents are computationally feasible using a group of workstations.

In order to be able to distribute the Interspace Prototype to general communities, our next step will be porting our system to a PC cluster. We recently built a Beowulf cluster of 8 nodes, each of which has a 450 MHz Pentium II processor, and 512MB of memory. Each node is estimated to be about 2 to 3 times more powerful than the current Sun workstations used in the experiment. If we can obtain the similar speedup as we saw in the previous Sun workstation cluster experiment, it is reasonable to expect PC clusters will shortly be able to compute semantic indexes for large real collections.

7 Acknowledgments

Financial support for this research was provided by the DARPA Information Management program under contract N66001-97-C-8535 and the NSF/DARPA/NASA Digital Libraries Initiative under cooperative agreement NSF IRI 94-11318COOP.

We would like to thank Nuala Bennett, Conrad Chang, Daniel Pape, Leslie Tyrrell, Kristin Tolle, William Pottenger and other members of the Interspace team both at UIUC and at the AI Lab of University of Arizona MIS Department for their time and effort without which this project would not have been possible. Richard Berlin, MD, at the HealthAlliance HMO in Urbana, and Jonathan Silverstein, MD, at the University of Illinois at Chicago, kindly served as early adopters for testing the prototypes.

REFERENCES

1. E. Brill. *A Corpus-Based Approach to Language Learning*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, 1993.

2. E. Brill. Transformation-Based Error-Driven Learning and Natural Language Processing. *Computational Linguistics*, 21(4):543–565, 1995.

3. J. R. Campbell, P. Carpenter, C. Sneiderman, S. Cohn, C. G. Chute, and J. Warren. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. CPRI Work Group on Codes and Structures. *Journal of American Medical Informatics Association*, 4(3):238–251, May 1997.

4. CANIS. *The Interspace Prototype*. CANIS, Community Architectures for Network Information Systems laboratory, University of Illinois at Urbana-Champaign,

http://www.canis.uiuc.edu/interspace. visited May 7, 1999.

5. C. Chang and B. R. Schatz. Performance and Implications of Semantic Indexing in a Distributed Environment. *ACM CIKM'99 Eighth International Conference on Information and Knowledge Management*, 1999. submitted.

6. H. Chen, A. Houston, R. Sewell, and B. Schatz. Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques. *Journal of the American Society for Information Science*, 49(7):582–603, 1998.

7. H. Chen and D. T. Ng. An algorithmic Approach to Concept Exploration in a Large Knowledge Network (Automatic Thesaurus Consultation): Symbolic Branch-and-bound Search vs. Connectionist Hopfield Net Activation. *Journal of the American Society for Information Science*, 46(5):723–728, June 1995.

8. H. Chen, D. T. Ng, J. Martinez, and B. R. Schatz. A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System. *Journal of the American Society for Information Science*, 48(1):17–31, January 1997.

9. H. Chen, B. Schatz, D. Ng, J. Martinez, A. Kirchhoff, and C. Lin. A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project. *IEEE Trans Pattern Analysis & Machine Intelligence, special issue on Digital Libraries: Representation and Retrieval*, 18(8):771–782, Aug 1996.

10. H. Chen, C. Schuffels, and R. Orwig. Internet Categorization and Search: A Machine Learning Approach. *Journal of Visual Communication and Image Representation, Special Issue on Digital Libraries*, 7(1):88–102, 1996.

11. Y. Chung. Machine learning approaches to information retrieval: Using genetic algorithms and neural networks for internet search and vocabulary switching. Master's thesis, Department of Management Information Systems, University of Arizona, 1997.

12. Y. Chung, W. Pottenger, and B. Schatz. Automatic Subject Indexing Using an Associative Neural Network. In *Proceedings of The Third ACM Conference on Digital Libraries*, pages 59–68, Pittsburgh, PA, June 1998.

13. J. J. Hopfield. Neural network and physical systems with collective computational abilities. *Proceedings of the National Academy of Sciences, USA*, 79(4):2554–2558, 1982.

14. A. T. Humphreys, B. L. McCray and M. L. Cheh. Evaluating the coverage of controlled health data terminologies: report on the results of the nlm/ahcpr large scale vocabulary test. *Journal of American Medical Informatics Association*, 4(6):484–500, Nov-Dev 1997.

15. C. Jacquemin and J. Royaute. Retrieving Terms and their Variants in a Lexicalized Unification-based Framework. In *Proceedings of the 17th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 132–141, 1994.

16. T. Kohonen. *Self-Organization and Associative Memory. Third Edition*. Springer-Verlag, Berlin Heidelberg, 1989.

17. C. Lynch and H. (eds) Garcia-Molina. Interoperability, Scaling, and the Digital Libraries Research Agenda: A Report on the IITA Digital Libraries Workshop, August 1995.

18. G. Salton. *A Theory of Indexing*. Philadelphia: Society for Industrial and Applied Mathematics, 1975.

19. G. Salton. *Dynamic Information and Library Processing*. Englewood Cliffs: Prentice-Hall, 1975.

20. B. R. Schatz. Information retrieval in digital libraries: Bringing search to the net. *Science*, 275(5298):327–334, Jan. 1997.

21. B. R. Schatz, W. Mischo, T. Cole, A. Bishop, S. Harum, E. Johnson, L. Neumann, H. Chen, and D. Ng. Federated Search of Scientific Literature: A Retrospective on the Illinois Digital Library Project. *IEEE Computer*, 32:51–59, Feb 1999.

22. K. M. Tolle and H. Chen. Comparing Noun Phrasing Techniques for Use With Medical Digital Library Tools. *Journal of the American Society for Information Science, Special Issue on Digital Libraries*, 1999. forthcoming.

23. A. Voutilainen. *A Short Introduction to NPtool*. Lingsoft, Inc. http://www.lingsoft.fi/doc/nptool/intro. visited May 7, 1999.