Biomedical Ontologies and Text Mining for Biomedicine and Healthcare: A Survey

Illhoi Yoo
Department of Health Management and Informatics, School of Medicine,
University of Missouri-Columbia, USA
yooil@health.missouri.edu

Min Song
Information Systems department
College of Computing Sciences
New Jersey Institute of Technology
min.song@njit.edu

In this survey paper, we discuss biomedical ontologies and major text mining techniques applied to biomedicine and healthcare. Biomedical ontologies such as UMLS are currently being adopted in text mining approaches because they provide domain knowledge for text mining approaches. In addition, biomedical ontologies enable us to resolve many linguistic problems when text mining approaches handle biomedical literature. As the first example of text mining, document clustering is surveyed. Because a document set is normally multiple-topic, text mining approaches use document clustering as a preprocessing step to group similar documents. Additionally, document clustering is able to inform the biomedical literature searches required for the practice of evidence-based medicine. We introduce Swanson's UnDiscovered Public Knowledge (UDPK) model to generate biomedical hypotheses from biomedical literature such as MEDLINE by discovering novel connections among logically-related biomedical concepts. Another important area of text mining is document classification. Document classification is a valuable tool for biomedical tasks that involve large amounts of text. We survey well-known classification techniques in biomedicine. As the last example of text mining in biomedicine and healthcare, we survey information extraction. Information extraction is the process of scanning text for information relevant to some interest, including extracting entities, relations, and events. We also address techniques and issues of evaluating text mining applications in biomedicine and healthcare.

Journal of Computing Science and Engineering, Vol. 2, No. 2, June 2008, Pages 109-136.

1. INTRODUCTION

An overwhelming amount of biomedical text information is available because vast quantities of biomedical discoveries and studies have been reported, documented, and stored to digital libraries or biomedical literature databases such as MEDLINE [NLM 2008]. For example, MEDLINE, the largest biomedical bibliographic text database, has nearly 18 million articles since 1966.

In order to overcome this text information overload and transform the text information into machine-understandable knowledge for knowledge management, text mining techniques have been used along with machine learning, data mining, information retrieval, etc. Text mining has been defined as the non-trivial discovery process for uncovering novel patterns in unstructured text [Fan et al. 2005; Mooney and Nahm 2003; Karanikas and Theodoulidis 2002]. Text mining techniques have been advanced by using techniques and methods from information retrieval, natural language processing, data mining, machine learning, and statistics. However, some old methods are rarely used nowadays. For example, while Bayesian models and hierarchical clustering were widely used in the early days, more advanced machine learning methods, such as artificial neural networks, support vector machines, and semantic-based clustering algorithms, have been applied in recent years.

Text mining has been applied to many different areas of biomedicine and healthcare such as gene clustering, protein structure prediction, spike signal detection, clinical diagnosis, biomedical hypothesis generation, measurement of patient care quality, and evidence-based medicine. In this paper, we briefly survey some of the relevant research in the field, covering biomedical ontologies and the applications of learning techniques in text mining in biomedicine and healthcare. More exhaustive and detailed reviews and discussions of selected text mining techniques and applications in biomedicine can be found in the subsequent sections in this paper.

The rest of the paper is organized as follows. Section 2 surveys biomedical ontologies. We briefly discuss ontology and major biomedical ontologies. In Sections 3, 4, 5, 6, and 7, document clustering, Swanson's Undiscovered Public Knowledge (UDPK), document classification, information extraction, and evaluation are discussed respectively. Section 7 concludes this paper.

2. BIOMEDICAL ONTOLOGIES

In this section, we discuss what ontology is and what major biomedical ontologies are.

2.1 Ontology

First of all, we briefly discuss what ontology is. Although ontology in philosophy is a field that studies what exists in the world or being, ontology in computer science (especially artificial intelligence), information science, bioinformatics and biomedical informatics is a sharable, reusable, machine-readable data structure, emphasizing practical usage. A well-known definition of ontology [Gruninger and Lee 2002] is "a formal, explicit specification of a shared conceptualization for a domain of interest" [Gruber 1995]. The following shows how some important terms in the definition should be interpreted:

• *conceptualization* refers to a model formed with concepts and their relationships in a domain for conceptual interpretation.

• *shared* implies the community of the domain should reach a consensus on the conceptualization.

• *formal specification* means that the specification must be machine-readable and machine-understandable.

• *explicit specification* indicates that concepts (meanings) and relationships between the concepts are explicitly defined.

2.2 Major Biomedical Ontologies

There are many biomedical ontologies; interested users should refer to Open Biomedical Ontologies (OBO[*]) and Unified Medical Language System (UMLS[†]) source vocabularies[‡] for a comprehensive list of biomedical ontologies. Each ontology except UMLS has its intended purpose and biomedical aspect. In this section, we briefly discuss Medical Subject Headings (MeSH[§]) and UMLS because they are most widely-used for biomedical information retrieval and biomedical literature mining and because National Library of Medicine (NLM) has developed them. For well-structured ontologies such as the MeSH or UMLS, the corresponding domain communities can reach a consensus on the knowledge in the ontologies. For this reason, ontologies can be used as domain knowledge for knowledge-based systems or intelligent agents. They enable us to go beyond traditional mathematics/statistics-based machine learning approaches.

2.2.1 MeSH

MeSH is formally classified as biomedical terminology rather than biomedical ontology. In fact, the distinction between ontology and terminology is not always clear at least in the biomedical informatics field [Bodenreider 2006]. This is because ontologies normally concentrate on the relationships among concepts and terminologies just record terms for concepts or entities of domain significance. However, many terminologies, including MeSH and Gene Ontology (GO), also provide a sort of semantic networks (e.g., MeSH Tree) as well as lexicons. However, MeSH Tree and GO supply simple relationships among concepts such as *isa* and *part of* (GO only), while UMLS defines the relationships among concepts in detail using, for example, *result_of*, *affects*, *derives_from*, etc. Accordingly, the distinction between terminologies and ontologies could be the detailness of defined relationships among concepts. Here, we assume that MeSH is a biomedical (pre)ontology because of MeSH Tree (discussed soon).

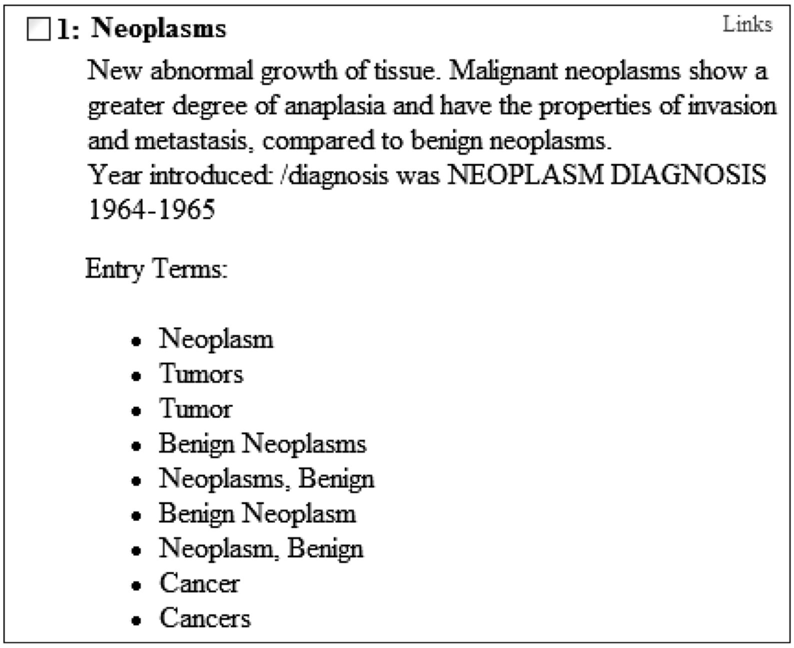
The MeSH, developed by the NLM in 1954, mainly consists of the controlled vocabulary and MeSH Tree. The controlled vocabulary contains several different types of terms such as Descriptor, Qualifiers, Publication Types, Geographics, and Entry terms. Among them, Descriptors and Entry terms are normally used for text

[*] http://www.obofoundry.org

[†] http://www.nlm.nih.gov/research/umls/

[‡] http://www.nlm.nih.gov/research/umls/metaa1.html

[§] http://www.ncbi.nlm.nih.gov/sites/entrez?db=mesh



□1: **Neoplasms**

Links

New abnormal growth of tissue. Malignant neoplasms show a greater degree of anaplasia and have the properties of invasion and metastasis, compared to benign neoplasms.
Year introduced: /diagnosis was NEOPLASM DIAGNOSIS 1964-1965

Entry Terms:

- Neoplasm
- Tumors
- Tumor
- Benign Neoplasms
- Neoplasms, Benign
- Benign Neoplasm
- Neoplasm, Benign
- Cancer
- Cancers

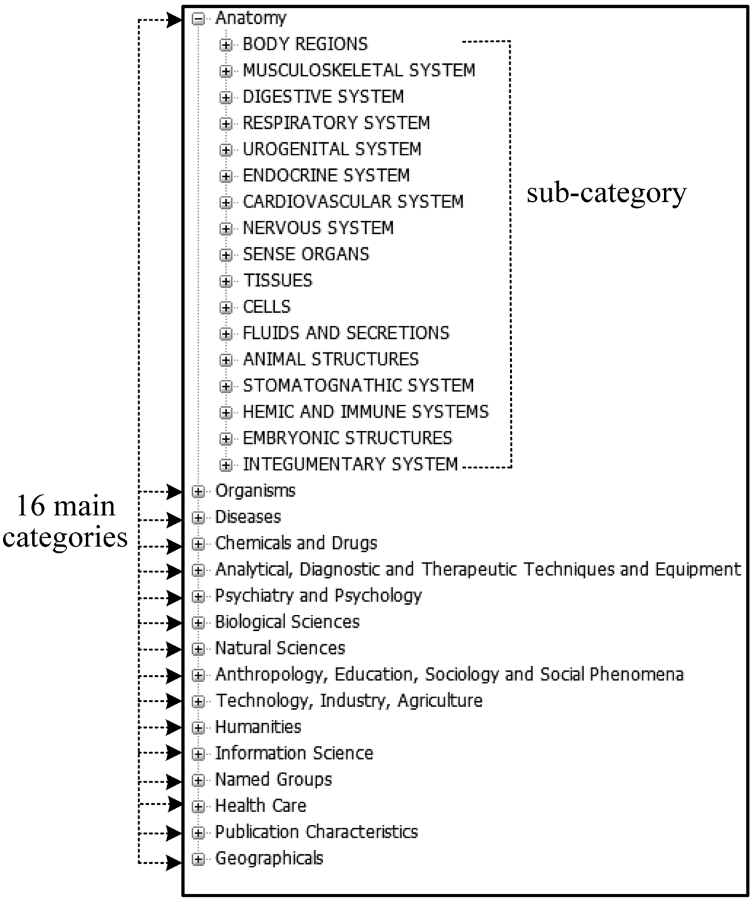
Figure 1. For the MeSH Descriptor of "Neoplasms", its definition and Entry Terms are provided. This figure was captured at www.ncbi.nlm.nih.gov and edited for this purpose.

mining approaches. The rest of them are reserved for information retrieval (i.e. MEDLINE search) so that Descriptor and Entry terms are discussed.

Descriptors are main concepts or main headings. Entry terms are basically the synonyms of Descriptors. Entry terms include linguistic variations (in word order and plurality) of the synonyms as well as Descriptors. For example, as shown in Figure 1, "Neoplasms" as a descriptor has the following entry terms: {"Neoplasm", "Tumors", "Tumor", "Benign Neoplasms", "Neoplasms, Benign", "Benign Neoplasm", "Neoplasm, Benign", "Cancer", "Cancers"}. The NLM annually revises and updates the MeSH controlled vocabulary including Descriptors and Entry terms. There are 24,767 Descriptors and more than 97,000 Entry terms in 2008 MeSH while there are 22,997 Descriptors in 2007 MeSH [NLM 2008].

MeSH Descriptors are organized in the MeSH Tree, which can be seen as the MeSH Concept Hierarchy, as shown in Figure 2. In other words, MeSH Descriptors are arranged by parent/child. In the MeSH Tree there are 16 categories [NLM-MeSH 2008], and each category is further divided into subcategories; for example, *Anatomy* category has *Body Regions*, *Musculoskeletal System*, *Digestive System*, etc. For each subcategory, its descriptors are hierarchically arranged from most general to most specific. The nodes in the tree are MeSH Descriptors. Because the 2008 MeSH Tree has 48,442 nodes, on average, a descriptor is placed in two places in the tree (48,442/ 24,767≈2). The nodes do not include the main branches of the MeSH Tree such as *Anatomy*, *Organism*, etc. that are not MeSH terms.

In addition to the ontology role of the MeSH, MeSH Descriptors have been used to



- Anatomy
  - BODY REGIONS
  - MUSCULOSKELETAL SYSTEM
  - DIGESTIVE SYSTEM
  - RESPIRATORY SYSTEM
  - UROGENITAL SYSTEM
  - ENDOCRINE SYSTEM
  - CARDIOVASCULAR SYSTEM
  - NERVOUS SYSTEM
  - SENSE ORGANS
  - TISSUES
  - CELLS
  - FLUIDS AND SECRETIONS
  - ANIMAL STRUCTURES
  - STOMATOGNATHIC SYSTEM
  - HEMIC AND IMMUNE SYSTEMS
  - EMBRYONIC STRUCTURES
  - INTEGUMENTARY SYSTEM
- Organisms
- Diseases
- Chemicals and Drugs
- Analytical, Diagnostic and Therapeutic Techniques and Equipment
- Psychiatry and Psychology
- Biological Sciences
- Natural Sciences
- Anthropology, Education, Sociology and Social Phenomena
- Technology, Industry, Agriculture
- Humanities
- Information Science
- Named Groups
- Health Care
- Publication Characteristics
- Geographicals

Figure 2. MeSH Tree with 16 hierarchies (categories).

index MEDLINE articles as well as NLM media. Some of the MeSH Tree categories (such as *Publication Characteristics* in Figure 2) are used only for indexing and retrieval purposes. For these purposes, about 10 to 20 MeSH terms are manually assigned to each article by highly-trained curators (after reading full papers). Upon the assignment of MeSH terms to an article, roughly 3 to 4 MeSH terms are set as “MajorTopics” that primarily represent that article depending on its contents.

2.2.2 UMLS

UMLS, started by NLM as a long-term R&D project in 1986, provides a mechanism for integrating all the major biomedical vocabularies such as the MeSH, the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), etc. Ultimately, UMLS has been designed for paving the way for the development of intelligent biomedical systems that can read Biomedical and Healthcare data, comprehend the meaning of them and further make inferences from them [NLM-UMLS, 2008]. UMLS is normally updated three times per year; each updated version has a name like 2008 AA, AB, or AC. UMLS consists of three knowledge sources; Metathesaurus, Semantic Network, and SPECIALIST lexicon.

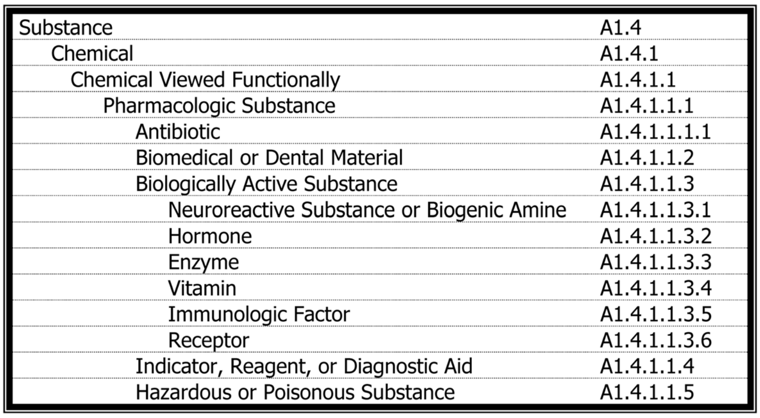
The Metathesaurus is a very large vocabulary database (nearly 5GB in text) whose data are collected from various biomedical thesauri. Currently, Metathesaurus (2007 AC version) contains more than 1 million biomedical concepts (meanings), 5 million unique concept names from nearly 150 different source vocabularies, and more than 17 million relationships between concepts. The Metathesaurus is basically organized by concepts (meanings). Each concept in the Metathesaurus has a unique concept identifier called CUI. Because a concept is an idea of how something is conceived, a concept can be represented in several different names (see Table I). Those names come from source vocabularies such as the MeSH. Each name from a source vocabulary is an atom; atoms are the “basic building blocks” of Metathesaurus [NLM-UMLS 2008]. Each atom has a unique identifier called AUI. Strings (concept names) from different source vocabularies may be the same so the Metathesaurus has a unique and permanent string identifier called SUI to combine the same strings. For example, *Benign Neoplasm* (A7569072) and *Benign Neoplasm* (A0029820) in Table I are the same and they have a common SUI S0018300. In addition, the Metathesaurus connects strings that can be seen as lexical variants or minor linguistic variations with LUI. For example, in Table I, the terms “*Benign Neoplasm*”, “*BENIGN NEOPLASM NOS* ”, “*Neoplasms, Benign*”, “*Benign Neoplasms*”, “*Benign neoplasms*”, “*BENIGN NEOPLASMS*”, and “*Neoplasm, Benign*” are linked to a single

Table I. UMLS Metathesaurus Concept Construction (The source vocabularies of the first five strings are present.).

Concept (CUI)	**Term (LUI)**	**String (SUI)**	**Atom (AUI)**
meaning	lexical variations	the same strings	string (concept name) from each vocabulary source
C0027651	**L0086795**	S0018300	A7569072: Benign Neoplasm *from National Cancer Institute Thesaurus*
			A0029820: Benign Neoplasm *from UMLS Metathesaurus*
		S0006783	A12968446: BENIGN NEOPLASM NOS *from ICD-9-CM, 2008*
		S0065757	A0090952: Neoplasms, Benign *from Medical Subject Headings (MeSH), 2008*
		S0219311	A7777134: Benign Neoplasms *from Medical Subject Headings (MeSH), 2008*
		S0358710	A0394473: Benign neoplasms
		S0595352	A8358142: BENIGN NEOPLASMS
		S0390371	A0438783: Neoplasm, Benign
	L0085417	S4182061	A7631348: Benign Tumor
		S0288291	A0317997: BENIGN TUMOR
		S0519216	A10765549: benign tumor
		S5926552	A6996890: Tumors, Benign
		S5926565	A6996903 Benign Tumors
	L0180889	S0245416	A8343815: Benign neoplasm of unspecified site
	L1030850	S1243886	A1207657: NONMALIGNANT TUMORS
	L6527860	S7590779	A12079848: BENIGN NEOPL
		S7600019	A12090680: NEOPL BENIGN

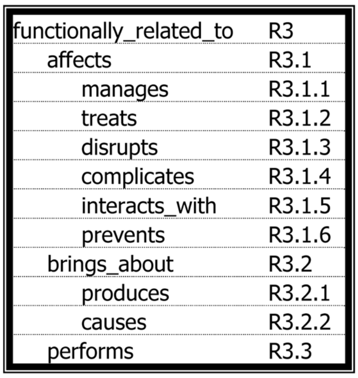
LUI L0086795 because they are merely lexical variants or minor linguistic variations.

The Semantic Network of UMLS consists of semantic types and semantic relations. semantic types are simply categories for concepts so that every Metathesaurus concept is assigned to at least one semantic type as a category. semantic relations are relationships (e.g., *diagnoses*) between semantic types instead of concepts. Thus, in the Semantic Network, semantic types are nodes and semantic relations are links between nodes. The Semantic Network currently (2007 AC version) contains 135 semantic types and 54 distinct semantic relations. Both the semantic types and the semantic relations are hierarchically arranged from most general to most specific as are MeSH Descriptors. Each semantic type or each semantic relation has its own tree number (for example, *A1.1*, *A1.1.1*, *A1.1.2*). Figures 3 and 4 show sample semantic types and sample semantic relations, respectively, as well as their tree numbers.



Substance	A1.4
Chemical	A1.4.1
Chemical Viewed Functionally	A1.4.1.1
Pharmacologic Substance	A1.4.1.1.1
Antibiotic	A1.4.1.1.1.1
Biomedical or Dental Material	A1.4.1.1.2
Biologically Active Substance	A1.4.1.1.3
Neuroreactive Substance or Biogenic Amine	A1.4.1.1.3.1
Hormone	A1.4.1.1.3.2
Enzyme	A1.4.1.1.3.3
Vitamin	A1.4.1.1.3.4
Immunologic Factor	A1.4.1.1.3.5
Receptor	A1.4.1.1.3.6
Indicator, Reagent, or Diagnostic Aid	A1.4.1.1.4
Hazardous or Poisonous Substance	A1.4.1.1.5

Figure 3. Sample semantic types arranged in a hierarchical structure.



functionally_related_to	R3
affects	R3.1
manages	R3.1.1
treats	R3.1.2
disrupts	R3.1.3
complicates	R3.1.4
interacts_with	R3.1.5
prevents	R3.1.6
brings_about	R3.2
produces	R3.2.1
causes	R3.2.2
performs	R3.3

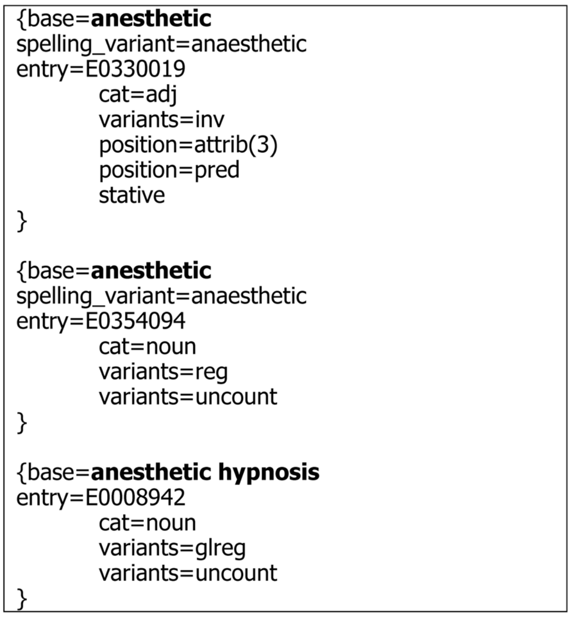
Figure 4. Sample semantic relations arranged in a hierarchical structure.

Basically, 558 “stem” relationships between semantic types have been made. Because the semantic types are hierarchically arranged, those relationships are inherited through the hierarchy. Through this inheritance process 6,752 relations are stated between semantic types from the stem relationships in the Semantic Network. For example, the stem relationship <*Biologic Function*|*process_of*| *Organism*> generates 208 derived relationships because *Biologic Function* semantic type has 13 child semantic types and *Organism* semantic type has 16 child semantic types; 13*16=208.

The SPECIALIST lexicon has been designed for the SPECIALIST Natural Language Processing (NLP) System. In order to supply the NLP system with lexical information, the lexicon contains general English words and many biomedical terms. Those words and biomedical terms come from the American Heritage Word Frequency Book, English dictionaries such as Longman's Dictionary of Contemporary English, the UMLS Test Collection of MEDLINE abstracts, and Dorland's Illustrated Medical Dictionary [NLM-UMLS 2008].

As shown in Figure 7, the lexicon records for each word/term are delimited by braces ("{ }"). Figure 7 shows 3 lexical entries. Each record basically has slot (e.g., base) and filler (e.g., anesthetic) format. Each record provides lexical information for each word/term in terms of morphology (base form and inflectional type (e.g., showing the pluralization pattern of a base form is regular or irregular)), orthography (spelling variants of a base form), and syntax (syntactic category (e.g., nouns, verbs, adjectives), position information).

Figure 7 shows that the first lexicon entry has the base form of *anesthetic* with a spelling variant of *anaesthetic* (orthographical information). The *anesthetic* is an adjective (adj) (syntactic information). The adjective (*anaesthetic*) is invariant (inv) so that it compares with *less*, *least*, *more*, or *most* (e.g., most anesthetic agents) (morphological information). *attrib(x)* and *pred* indicate the adjective is attributive (e.g., anesthetic drugs) and predicative (the drug is anesthetic) respectively (syntactic



```
{base=anesthetic
spelling_variant=anaesthetic
entry=E0330019
        cat=adj
        variants=inv
        position=attrib(3)
        position=pred
        stative
}

{base=anesthetic
spelling_variant=anaesthetic
entry=E0354094
        cat=noun
        variants=reg
        variants=uncount
}

{base=anesthetic hypnosis
entry=E0008942
        cat=noun
        variants=glreg
        variants=uncount
}
```

Figure 5. Sample SPECIALIST lexicon.

information). The number (1-3) in the parentheses ("()") indicates the location of the adjective in the typical sequence of adjectives in a noun phrase (syntactic information). There are three kinds of adjectives: qualitative (attrib(1)), color (attrib(2)), and classifying (attrib(3)) [Browne et al. 2000]. Qualitative adjectives normally precede color adjectives and color adjectives generally precede classifying adjectives. The adjective *anaesthetic* is classifying one (attrib(3)). Last, stative indicates the adjective is static. The stative adjective does not describe an activity or event but a state or condition. For example, we do not say that "the drug is being anesthetic." but "the drug is anesthetic".

3. DOCUMENT CLUSTERING

Unlike classification, document clustering is an unsupervised learning process. Basically, document clustering is to group unlabeled documents into meaningful document clusters whose documents are similar to one another within the cluster, without any prior information about the document set. The problem of document clustering is formally defined as follows: Given a set of n documents called DS, DS is clustered into a user-defined number of k document clusters DS_1, DS_2,…DS_k, ($\{DS_1, DS_2…,DS_k\} = DS$) so that the documents in a document cluster are similar to one another while documents from different clusters are dissimilar. In order to measure similarities between documents, documents have been represented based on the vector space model. In this model, each document d is represented as a high dimensional vector of words'/terms' frequencies (as the simplest form), where the dimensionality indicates the vocabulary of DS; the size of the vocabulary is the number of distinct words/terms in the document set. Although there are a number of similarity measurements, similarity between two documents has been traditionally measured by the cosine of the angle between their vector representations. Documents are grouped based on a cluster criterion function as an iterative optimization process that measures key aspects of inter-cluster and intra-cluster similarities.

3.1 Classification of Document Clustering Approaches

A number of document clustering approaches have been developed over several decades. Most of these document clustering approaches are based on the vector space representation and apply various clustering algorithms to the representation. As a result, most of the approaches are categorized as either hierarchical or partitional [Kaufman and Rousseeuw 1999].

3.1.1 Hierarchical clustering

Hierarchical agglomerative clustering algorithms were used for document clustering. The algorithms successively merge the most similar objects based on the pairwise distances between objects until a termination condition holds. Thus, the algorithms can be classified by the way they select a pair of objects for calculating the similarity measure (e.g., single-link, complete-link, and average-link). An advantage of the algorithms is that they generate a document hierarchy so that users can drill up and drill down for specific topics of interest. However, due to their cubic time complexity, they are limited in their scalability so that they cannot handle a very large number of documents.

3.1.2 Partitional clustering

Partitional clustering algorithms (especially K-means) are the most widely-used algorithms in document clustering [Steinbach et al. 2000]. Most of the algorithms first randomly select k centroids and then decompose the objects into k disjoint groups through iteratively relocating objects based on the similarity between the centroids and the objects. The clusters become optimal in terms of certain criterion functions. As the most widely-used partitional algorithm K-means minimizes the sum of squared distances between the objects and their corresponding cluster centroids. K-mean's complexity is $O(k^*T^*n)$, where k is the number of clusters, T is the number of iterations for relocating objects, and n is the number of objects. As a variation of K-means, BiSecting K-means [Steinbach et al. 2000] first pick a cluster (normally the biggest one) to split and then splits the objects into two groups (i.e. $k = 2$) using K-means. One major drawback of partitional clustering algorithms is that clustering results are heavily sensitive to the initial centroids because the centroids are randomly selected.

3.1.3 Hybrid clustering

There are some hybrid document clustering approaches that combine hierarchical and partitional clustering algorithms. For instance, Buckshot [Cutting et al. 1992] follows a K-means approach but uses the average-link to set cluster centroids with the assumption that hierarchical clustering algorithms provide superior clustering quality to K-means. In order to create cluster centroids for K-means, Buckshot first picks $\sqrt{kn}$ objects randomly and then uses an average-link algorithm whose complexity is $(O(n^2 log\ n))$. In order to make the overall complexity linear, Buckshot selects $\sqrt{kn}$ objects. However, as Larsen & Aone pointed out [Pantel and Lin 2002], using a hierarchical algorithm for centroids does not significantly improve the overall clustering quality compared with the random selection of centroids.

3.1.4 Semantic document clustering

Hotho et al. introduced the semantic document clustering approach that uses background knowledge [Hotho et al. 2002]. The authors apply an ontology during the construction of a vector space representation by mapping terms in documents to ontology concepts and then aggregating concepts based on the concept hierarchy, which they called concept selection and aggregation (COSA). As a result of COSA, they resolve a synonym problem and introduce more general concepts in the vector space to easily identify related topics [Hotho et al. 2002]. Their method, however, cannot reduce the dimensionality (i.e. the document features) in the vector space; it still suffers from the "*Curse of Dimensionality*".

3.1.5 Non-vector-space document clustering

While all the approaches mentioned above represent a document as a feature vector, Suffix Tree Clustering (STC) [Zamir and Etzioni 1998] does not rely on the vector space model. STC does not treat a document as "a set of words", where the order is not important, but rather as an ordered sequence of words (i.e. a set of phrases). In fact, phrases instead of words have long been used in IR systems [Buckley et al. 1995]. One of major drawbacks of STC is that semantically similar nodes may be distant within a suffix tree, because STC does not consider the semantic relationships among phrases (nodes or base clusters). In addition, some common expressions may lead to combining unrelated documents. Recently Eissen et al. applied STC to the RCV1 document collection of Reuters Corporation and showed STC did not produce good clustering results; the average F-measure was 0.44 [Eissen et al. 2005].

3.1.6 Graph-based semantic document clustering

Recently, Yoo et al. introduced two graph-based semantic document clustering approaches (a scale-free, graph-based, semantic document clustering method and a bipartite, graph-based, semantic document clustering method) and successfully applied them to MEDLINE documents [Yoo et al. 2007; Yoo et al. 2006]. The key difference of the two approaches is how they represent documents: scale-free graph or bipartite graph.

The key of the first approach is the use of the graphical representation method using a biomedical ontology. The graphical representation method represents a set of documents as an ontology-enriched scale-free graph. This ontology-enriched graphical representation method has several advantages over the traditional vector space based approaches. First, it provides a very natural way to portray the contents of documents. Second, it provides *document representation independence* that means that the graphical representation of a document does not affect other representations. Last, it guarantees better scalability on text mining than the traditional vector space model. The ontology-enriched graph (i.e., the corpus-level graphical representation of documents) is clustered under the consideration of the power law distribution of terms in documents to identify document cluster models as semantic chunks capturing the semantic relationships among the terms in the document clusters. These document cluster models are used for assigning documents to clusters to group semantically similar documents in accordance with the similarity between each document and each document cluster model.

The second approach models a set of documents as a bipartite graph, where the two sets on the graph are documents and corpus-level significant semantic features. The semantic features are the selected co-occurrence concepts based on Mutual Information. The use of co-occurrence concepts has several advantages. First, the co-occurrence concepts have been regarded as more important than the single word/term [Hristovski et al. 2001; Jenssen et al. 2001; Perez-Iratxeta et al. 2002] because they capture potential relationships between two co-occurring concepts in text. Second, the use of co-occurrence concepts prevents noise terms (which are unrelated to the topic of a document but are found in the document) from affecting the similarity measures during document clustering. After selecting co-occurrence concepts as significant semantic features, they are classified according to their relationships with documents and their semantic similarities in a concept hierarchy in an ontology. Then the documents are clustered based on each document's contribution to each significant semantic feature group. To refine the initial document clustering, Yoo and colleagues developed a new spectral co-clustering algorithm that uses the mutual-refinement relationship between the significant semantic feature groups and the document groups so that the two groups are mutually recursively refined.

3.2 Applications of Document Clustering

Document clustering was initially investigated for improving information retrieval (IR) performance in terms of precision and recall or F-measure because similar documents grouped by document clustering tend to be relevant to the same user queries [Wang et al. 2002; Zamir and Etzioni 1998]. Document clustering has been recently used to facilitate nearest-neighbor search [Buckley and Lewit 1985], to support an interactive document browsing paradigm [Cutting et al. 1992; Koller and Sahami 1997; Gruber 1993], and to construct hierarchical topic structures [van Rijsbergen 1979]. Thus, document clustering plays a more important role for IR and text mining communities since the most natural form for storing information is text, and text information has increased exponentially.

In the biomedical domain, document clustering technologies have been used to facilitate the practice of evidence-based medicine. This is because document clustering enhances biomedical literature searching (e.g., MEDLINE searching) in several ways and literature searches are one of the core skills required for the practice of evidence-based medicine [Evidence-based Medicine Working Group 1992]. For example, Pratt and her colleagues [Pratt et al. 1999; Pratt and Fagan 2000], and Lin and Demner-Fushman [Lin and Demner-Fushman 2007] introduced interesting semantic document clustering approaches that automatically cluster biomedical literature (MEDLINE) search results into document groups for better understanding of literature search results.

Unlike traditional document clustering methods, DynaCat [Pratt et al. 1999; Pratt and Fagan 2000] takes advantage of keywords assigned by article authors and MeSH terms assigned by NLM and tries to map them to UMLS semantic types to classify MEDLINE articles. Because a term is assigned at least one UMLS semantic type, DynaCat provides soft categorization. In addition, DynaCat utilizes user's query to categorize search results through Query Model (created by the authors) and Terminology Model (borrowed from UMLS). In other words, document categorization is dependent on kinds of user's query: there are nine query types (for example *symptoms-diagnoses*) A major drawback of this approach is that the method cannot cover all kinds of user queries because it is nearly impossible to create a comprehensive list of user queries.

Recently, Lin and Demner-Fushman developed a MEDLINE document clustering approach focusing on therapy and diagnosis questions [Lin and Demner-Fushman 2007]. In order to group similar documents, terms (in MEDLINE articles) belonging to the *Chemicals & Drugs, Devices, and Procedures* UMLS semantic type are extracted using MetaMap** and ranked based on some heuristics (e.g., term's location in an abstract). The highest ranked term for each article becomes the representative for the article. The approach combines the representative UMLS terms based on UMLS

**http://mmtx.nlm.nih.gov/

Metathesaurus concept relationships (that are hierarchically structured), until no representatives are left to combine. In this way, MEDLINE articles are hierarchically arranged and each article is labeled. A major drawback of this approach is that each MEDLINE article is labeled using only one UMLS term. However, this does not reflect the reality that most MEDLINE articles discuss several concepts. In addition, there is a high possibility that the approach may miss the most important term in MEDLINE articles.

4. SWANSON'S UNDISCOVERED PUBLIC KNOWLEDGE (UDPK)

The huge volume of the biomedical literature provides a promising opportunity to induce novel knowledge by finding novel connections among logically-related medical concepts. For example, Swanson introduced Undiscovered Public Knowledge (UDPK) model to generate biomedical hypotheses from biomedical literature such as MEDLINE [Swanson 1986]. According to Swanson, UDPK is "a knowledge which can be public, yet undiscovered, if independently created fragments are logically related but never retrieved, brought together, and interpreted".

The UDPK model formalizes a procedure to discover novel knowledge from biomedical literature as follows (see Figure 6): Consider two separate sets of biomedical literature, BC and AB, where the BC document set discusses biomedical concepts B and C and the AB document set discusses biomedical concepts B and A. However, none of the documents in the BC or AB sets primarily discusses biomedical concepts C and A together. The goal of the UDPK model is to discover some novel connections between a starting concept C (e.g., a disease) and target concepts A (e.g., possible medicine or treatments to the disease) by identifying biomedical concept B (called a bridge concept). For example, Swanson discovered that fish oils (as concept A) could be a potential medicine for Raynaud disease (as concept C) by identifying the bridge concept (as concept B) blood viscosity. This discovery (UDPK) is accomplished by finding two different biomedical document sets in which one set (the CB document set) discusses that Raynaud disease (as concept C) aggravates blood viscosity (as concept B) and the other set (the BA document set) discusses that fish

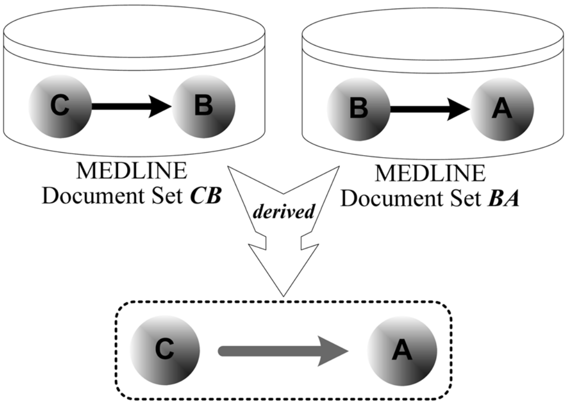


Figure 6. Swanson's Undiscovered Public Knowledge model.

oils (as concept A) improve blood viscosity (as concept B).

Swanson's UDPK model can be described as a process to induce "C implies A", which is derived from both "C implies B" and "B implies A"; the derived knowledge or relationship "C implies A" is not conclusive but, rather, hypothetical. The B concepts are the bridge between concepts C and A. The following steps summarize the procedure [Swanson 1987].

1. Specify the user's goal (a starting concept C such as a disease, or symptom, etc.)
2. Search the relevant documents BC from the biomedical literature (e.g., MEDLINE) for C.
3. Generate a set of selected biomedical terms (called "B" list) from the BC document set using predefined stop-list filter; B concepts are chosen from only the titles of the documents.
4. Search MEDLINE for each term in the B list to retrieve AB documents related to the B concepts.
5. Generate a set of biomedical terms (A candidates) from the AB documents; A concepts also come from only the titles of the documents.
6. Check whether each of the A candidates and C are co-cited together in any MEDLINE articles. If not, keep the A candidate.
7. Rank the selected A terms based on how many linkages are made with B terms.

One of the drawbacks of Swanson's method is that a large amount of manual intervention is required. Although he and his colleague designed an interactive tool called Arrowsmith to automate some of the steps [Swanson and Smalheiser 1999], the procedure still requires much manual intervention, such as the choice of proper lists of stop words and filtering through a large number of C-B and B-A connections to identify the real novel connections/hypotheses. Another problem is that the relationships or associations among a huge number of the biomedical concepts grow exponentially. As a result, the key of the UDPK problem is how to exclude meaningless C-B and B-A concept pairs because the automated process yields too many irrelevant suggestions.

Several algorithms have been developed to overcome the limitations of Swanson's approach. Hristovski et al. use the MeSH descriptors rather than the title words of MEDLINE articles [Hristovski et al. 2001]. They use association rule algorithms to find the co-occurrence of the words. Their methods identify as bridges all B concepts that are related to the starting concept C. Then, all A concepts related to B concepts are found through MEDLINE searching. However, in MEDLINE each concept can be associated with many other concepts, the number of C-B and B-A concept pairs can be extremely large. In order to deal with this problem, the algorithm incorporates filtering and ordering capabilities [Hristovski et al. 2003; Hristovski et al. 2001; Joshi et al. 2004].

Pratt and Yetisgen-Yildiz used UMLS concepts instead of MeSH terms assigned to MEDLINE documents [Pratt and Yetisgen-Yildiz 2003]. Similar to Swanson's method, their search space is limited to only the titles of MEDLINE articles. As a result, they can reduce the number of possible B-A concept pairs. In addition to that, they reduce the number of terms/concepts by pruning out terms that are "too general" (e.g., problem, test, etc), "too closely related to the starting concept", and "meaningless" in the following ways: They defined terms as "too general" if the terms are found in titles of MEDLINE articles more than 10,000 times. For "too closely related to the starting concept", they tracked all the parents and children concepts of the starting concept and then eliminated the related terms. To avoid "meaningless" terms, they manually selected a subset of semantic types to which the collected terms should belong, as did Hristovski et al. [Hristovski et al. 2001]. After eliminating these terms, they group B concepts and A concepts based on their similarity. They then remove "too general" concepts in each group by looking at their UMLS hierarchy level, and finally remove non-UMLS concepts. With the qualified and grouped UMLS concepts (for B and A), they use the well-known Apriori algorithm [Agrawal et al. 1995] to find correlations among the concepts. Although they manage to simulate Swanson's migraine-magnesium case only through concept grouping, their method still requires strong domain knowledge, especially for selecting semantic types for A and B concepts and also some vague parameters for defining "too general" concepts.

Srinivasan viewed Swanson's method as involving two dimensions [Srinivasan 2004]. The first dimension is about identifying relevant concepts for a given concept. The second dimension is about exploring the specific relationships between concepts. However, Srinivasan deals with only the first dimension. The key of this approach is that MeSH terms are grouped into the semantic types of UMLS to which they belong. However, only a small number of semantic types (8 out of 134) are considered because the author believes those semantic types are relevant to B and A concepts. For each semantic type, the MeSH terms that belong to the semantic type are ranked based on a modified TF*IDF. There are some limitations in the method. First, the author used manually-generated semantic types for filtering. Second, the author applied the same semantic types to both A and B terms even though the roles of the A and B terms for the C term are different.

These research work mentioned above represent significant progress on Swanson's method. However, none of the approaches considers the different roles of concepts A and B for concept C. We believe the UDPK association problem should be tackled by means of not only information measures such as TF*IDF but also the semantic relationships among the concepts.

5. DOCUMENT CLASSIFICATION

Medical experts have hard time identifying the relevant medicine related literature. Human classification is a highly time consuming task, and slow process. Furthermore, it has been applied to several aspects of biomedical research, such as database construction [Stapley et al. 2002], gene/protein function annotation [Raychaudhuri et al. 2002], and clinical records data mining [Pakhomov et al. 2003]. Within the field of IR, text classification provides the means to drastically improve the efficiency of medical experts. Text classification is a technique to automatically determine the category that a document or part of a document belongs to based on the particular topics or characteristics of interest that a document contains. Classification methods primarily rely on supervised machine learning techniques.

Features used for classification are not specified explicitly by the user; instead the user only provides a set of documents that contain the characteristics of interest, which is known as the positive training set, and another set that does not contain the characteristics, which is known as the negative training set.

The goal of a text classification system is to assign the class labels to the new unseen documents. Classification can be defined into two different phases, namely model construction or also widely known as the training phase, and model usage also known as the prediction phase.

In the training phase, given a set of positively labeled sample documents, the goal is to automatically extract features relevant to a given class. Hence, it would help distinguish the positive documents from the negative ones. Once such features have been identified then those features should be applied to the candidate documents using some kind of decision-making process. In the model usage phase, check the accuracy of the model and use to it classify new unseen data.

5.1 Classification Methods

Several methods have been proposed for document classification, including naïve Bayesian [Raychaudhuri et al. 2002], maximum entropy [Joachims 1999; Pakhomov et al. 2002], support vector machine [Stapley et al. 2002; Jenssen et al. 2001].

5.1.1 Naïve Bayes

Naïve Bayes assumes that the features in the input feature vector are statistically independent, hence the order of words, and presence of a word does not affect the presence or absence of the other words [Sebastiani 2002; Rish 2001]. This assumption makes the computation of Bayesian approaches more efficient. Although the assumption is not true in any language, it has been shown that the classification accuracy is insignificantly affected by such disruptions [Domingos and Pazzani 1997]. It is the simplest yet very effective, and due to its simplicity also the single most researched classifier. Naïve Bayes is also the canonical classifier [Rish 2001] such that it is a baseline against which other new classification methods are measured [Pant and Srinivasan 2005]. Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured.

Naïve Bayesian approaches have been studied frequently in data mining even before the topic of document classification became popular. Their performance is as well as the newer, more sophisticated methods [Witten and Frank 2000]. Additionally, they show a very good computational performance during classification of new documents [Herrmann et al. 2001].

5.1.2 SVM

The support vector machine (SVM) method was introduced for text classification in [Joachims 1998; Joachims 1999], and has been subsequently used in several other text classification tasks [Vapnik 1995; Witten and Frank 2000; Yang and Liu 1999]. The goal of SVM is to automatically learn a separation hyperplane from a set of positive and negative training examples, which splits classified entities into two subsets according to a certain classification property. As the name SVM suggests, it is required that the text documents be represented as vectors but this problem has been handled in various classification and clustering algorithms. Generally, each word in the vocabulary of the corpus becomes a dimension, and a vector represents the number of occurrences of the respective words in the document. Stemming is utilized so that words such as "run" and "running" are counted together. SVMs work well for text classification as there are many words in the vocabulary, yielding a high-dimensional vector space. At the same time, each document might only use a small subset of the thousands of words in the vocabulary of the corpus. SVMs are thus well suited for such document vectors that are sparse but contain dense concepts (i.e., the words that are present in a document are important). SVMs were normally adapted for binary classification method that combines statistical learning and optimization techniques with kernel mapping [Vapnik 1995], but have been successfully applied to multiclass classification [Crammer and Singer 2001].

The optimization part is used to maximize the distance, also called the margin of each of the two subsets from the hyperplane. The document representatives closest to the decision surface are called support vectors. The result of the algorithm remains unchanged if documents that do not belong to the support vectors are removed from the set of training data. An advantage of SVM is its superior runtime-behavior during the categorization of new documents because only one dot product per new document has to be computed. A disadvantage is the fact that a document could be assigned to several categories because the similarity is typically calculated individually for each category. Nevertheless, SVM is a very powerful method and has outperformed other methods in several studies [Dumais et al. 1998; Hearst et al. 1998; Joachims 1998; Siolas and d'Alché-Buc 2000; Yang and Liu 1998].

According to [Joachims 1998], unlike many other classification methods that have difficulties coping with huge dimensions, one of the major advantages of the SVM approach is that its performance does not depend on the dimensionality reduction. SVMs tend to be fairly robust to overfitting and can scale up to considerable dimensionalities. [Brank 2002] also indicated that feature selection tends to negatively effect the performance of SVMs. Even, it has been shown that the removal of stop-words is not necessary [Leopold and Kindermann 2002].

Few approaches have focused on tuning the original SVM approach by selecting different features, or by using different feature weights and kernels, mostly for the text classification task. For example, [Leopold and Kindermann 2002] have discussed the impact of different feature weights on the performance of SVMs in the case of document classification in English and German. Additionally, they have reported that an entropy-like weight performed better than idf, especially for larger documents. Moreover, they proposed that when using single word as features, lemmatization was not required, as it had no significant effect on the performance.

Additionally, SVMs have been used extensively in the classifying documents in the biomedical domain, as discussed below.

In Donaldson and his colleagues' study [2003], SVM is used to train on the bag-of-words in MEDLINE abstracts to distinguish abstracts containing information on protein-protein interactions, prior to curating this information into their BIND database. A small evaluation with 100 abstracts found a precision of 96% with a recall of 84%. They estimated that the classification system would reduce the number of abstracts that the curators needed to read by 2/3rds.

5.1.3 Maximum entropy

Maximum entropy takes advantage of feature distribution characteristics in the training set, and is suitable for classification tasks having a discriminated feature set.

In Liu and her colleagues' study [2004], the authors have focused on classifying figures captions in full-text scientific papers. They applied text mining and classification techniques on the text in figure legends, in order to identify an interesting figure based on schematic representations of protein interactions and signaling events. Furthermore, each figure was classified as relevant/non-relevant using a maximum entropy classifier, by assigning an estimated likelihood. One of the major advantages of the maximum entropy technique is that it provides a probability score for the decision, instead of a binary assignment. Also, this accelerated the experts manual work, while the results showed positively the linkage between picture and text.

6. INFORMATION EXTRACTION

In terms of what is to be extracted by the systems, most studies can be broken into the following three major areas: 1) named entity extraction, named entities may include *proteins* or *genes*, 2) relation extraction whose main task is to extract relationships among entities, and 3) abbreviation extraction. Most of these studies adopt information extraction techniques, using a curated lexicon or natural language processing for identifying relevant tokens such as words or phrases in text [Shatkay & Feldman, 2003].

Fukuda et al. [1998] extract protein names with hand-crafted rules. Although they reported that experimental results were competitive based on F-value of 0.92, the results were not replicated and their method relied on manually created rules. Proux et al. [2000] use single word names only with selected test set from 1,200 sentences coming from Flybase. Collier et al. [2000] adopt Hidden Markov Models (HMMs) for 10 test classes with small training and test sets. Krauthammer et al. [2000] use BLAST database with letters encoded as 4-tuples of DNA. Demetriou and Gaizuaskas [2002] pipeline the mining processes including hand-crafted components and machine learning components. For the study, they use large lexicon and morphology components. Narayanaswamy et al. [2003] use a Part of Speech (POS) tagger for tagging the parsed MEDLINE abstracts. Although Narayanaswamy and his colleagues implement an automatic protein name detection system, the number of words used is 302 and thus it is difficult to see the quality of their system in that the size of the test data is too small. Yamamoto et al. [2003] use morphological analysis techniques for preprocessing protein name tagging and apply SVM for extracting protein names. They found that increasing training data from 390 abstracts to 1,600 abstracts improved F-value performance from 70% to 75%. Lee et al. [2003] combined an

SVM and dictionary look-up for named entity recognition. Their approach is based on two phases, the first phase is identification of each entity with an SVM classifier and the second phase is post-processing to correct the errors by the SVM with a simple dictionary look-up. Bunescu et al. [2004] studied protein name identification and protein-protein interaction. 5,206 names extracted from MEDLINE database were used in their experiment. Among several approaches used in their study, the main two ways are one using POS tagging, and the other using the generalized dictionary-based tagging. Their dictionary-based tagging presents higher F-value. Table I summarizes the work in the areas of named entity extraction in biomedical literature.

The second extraction task of biomedical literature extraction is relation extraction. Leek [1997] applies HMM techniques to identify gene names and chromosomes through heuristics. Craven and Kumlien [1999] extract location relations for proteins in Yeast database based on HMM techniques. Blaschke et al. [1999] extract protein-protein interactions based on co-occurrence of the form "… p1 I1… p2" within a sentence, where p1, p2 are proteins and I1 is an interaction term. Protein names and interaction terms (e.g., activate, bind, inhibit) are provided as a "dictionary." Rindflesch [1999] extracts binding relations for Unified Medical Language System (UMLS) from MEDLINE. Proux [2000] extracts an "interact" relation for the gene entity from Flybase database. Pustejovsky [2002] extracts an "inhibit" relation for the gene entity from MEDLINE. Jenssen et al. [2001] extract gene-gene relations based on co-occurrence of the form "… g1…g2…" within a MEDLINE abstracts, where g1 and g2 are gene names. Gene names are provided as a "dictionary", harvested from HUGO, LocusLink, and other sources. Although their study uses 13,712 named human genes and millions of MEDLINE abstracts, no extensive quantitative results are reported and analyzed. Friedman et al. [2001] extract a pathway relation for various biological entities from a variety of articles. In their work, the precision of the experiments is high (from 79-96%). However, the recalls are relatively low (from 21-72%). Bunescu

Table II. A summary of works in Biomedical Entity Extraction.

Authors	Named Entities	Database	No. of Words	F value
Fukuda et al.	Protein	MEDLINE	20,000	93
Proux et al	Gene	Flybase summary	12,000	92
Collier et al.	Proteins and DNA	MEDLINE	30,000	73
Krauthammer et al.	Gene and Protein	Review articles	5,000	75
Demetriou and Gaizauskas	Protein, Species, and 10 more	MEDLINE	30,000	83
Narayanaswamy	Protein	MEDLINE	302	75.86
Yamamoto et al.	Protein	GENIA	1,600 abstracts	75
Lee et al.	Protein DNA RNA	GENIA	10,000	77

Table III. A summary of relation extraction for biomedical data.

Source	Relation	Entity	DB	Precision	Recall
Leek	Location	Gene	OMIM	80%	36%
Craven	Location	Protein	Yeast	92%	21%
Rindflesch	Binding	UMLS	MEDLINE	79%	72%
Blaschke	Interact	Protein	MEDLINE	n/a	n/a
Proux	Interact	Gene	Flybase	81%	44%
Pustejovsky	Inhibit	Gene	MEDLINE	90%	57%
Jenssen	Location	Gene	MEDLINE	n/a	n/a
Friedman	Pathway	Many	Articles	96%	63%

et al. [2004] conducted protein/protein interaction identification with several learning methods such as pattern matching rule induction (RAPIER), boosted wrapper induction (BWI), and extraction using longest common subsequences (ELCS). ELCS automatically learns rules for extracting protein interactions using a bottom-up approach. They conducted experiments in two ways; one with manually crafted protein names and the other with extracted protein names by their name identification method. In both experiments, Bunescu et al. compared their results with human-written rules and showed that machine learning methods provides higher precisions than human-written rules. Table II summarizes work on relation extraction in biomedical literature.

The third extraction task is abbreviation extraction. Chang et al. [2003] use an algorithm with a logistic regression technique to extract abbreviations. Their algorithm scores abbreviation expansions based on the similarity to a training set of human-annotated abbreviations from MEDLINE abstracts. The algorithm is reported to have a maximum recall of 83% at 80% precision. Major limitations of their approach are that an abbreviation must be enclosed in parentheses and a set of rules applied to abbreviation extraction was not comprehensive compared to other rule-based extraction techniques.

Yu et al. [2002] present a system (i.e., AbbRE) with a rule-based algorithm. Their system contains pattern-matching rules for mapping abbreviations to their full forms in biomedical text. AbbRE is reported to have an average 70% recall and 95% precision for defined abbreviations. However, their experimental setup was limited to defined abbreviations which constitute only 25 percent of total abbreviations in biomedical articles as their own statistics identify.

Liu and Friedman [2003] propose an algorithm based system to extract a set of related terms from the biomedical literature. Their method is also based on the observation that presentation of abbreviations in text are usually within parentheses. The recall of the algorithm was around 88.5%, and its precision was 96.3%. The limitation of their approach is that the system is not suitable for identifying expansions that occur only once in a text.

Schwartz and Hearst [2003] report a system with a simple algorithm based on the use of parentheses and ad hoc rules for identifying abbreviations' definitions in biomedical texts. The algorithm has an experimental result of 82% of recall and a precision of 96%.

Ao and Takagi [2005] describe an ad hoc algorithm called ALICE. ALICE identifies and extracts pairs of abbreviations and their expansions by using parentheses-searching and heuristic pattern-matching rules. In addition to the strategies used by Yu et al. [2003] and Schwartz and Hearst [2003], this algorithm uses manually expanded patterns, rules, and stop word lists. The authors argue that their system can potentially validate 320 abbreviation-expansion patterns as combinations of the rules. It is reported that the system achieved 95% recall and 97% precision on randomly selected titles and abstracts from the MEDLINE database. ALICE is reported to be limited to disambiguate synonyms and expansions.

Song and Yoo [2007] proposed a hybrid extraction technique to detect the corresponding long forms (i.e., definitions, expansions, and full names) of short forms (i.e., abbreviations, acronyms, and symbols) from biomedical text. It incorporates the proposed method gives us the comparative advantages over others in the following aspects: 1) it incorporates lexical analysis techniques into supervised learning for extracting abbreviations; 2) it makes use of text chunking techniques to identify long forms of abbreviations.

7. EVALUATION

One of the pivotal issues yet to be explored further for biomedical literature mining is how to evaluate the techniques or systems.

The focus of the evaluation conducted in the literature is on extraction accuracy. The accuracy measures used in IE are precision and recall ratios. For a set of N items, where N is either terms, sentences, or documents and the system needs to label each of the terms as "positive" or as "negative" according to some criterion – "positive" if a term belongs to a predefined document category or a term class. Recall is defined as the number of relevant documents retrieved divided by the total number of relevant documents in the collection. For example, suppose there are 80 documents relevant to widgets in the collection. If the system X returns 60 documents and 40 of which are about widgets, then X's recall is 40/80 = 50%. Precision is defined as the number of relevant documents retrieved divided by the total number of documents retrieved. In our example, X's precision is 40/60 = 67%. As discussed earlier, the extraction accuracy is measured by precision and recall ratio. Although these evaluation techniques are straightforward and are well accepted, calculating recall ratios may be criticized when the total number of true "positive" terms is not clearly defined.

Participants in the Message Understanding Conference (MUC) tested the ability of their systems to identify entities in text to resolve co-reference, extract and populate attributes of entities, and perform various other extraction tasks from written text. As identified by Shatkay and Feldman [2003], the important challenge in biomedical literature mining is "the creation of gold-standards and critical evaluation methods for systems developed in this very active field." The framework of evaluating biomedical literature mining systems was recently proposed by Hirschman et al. [2002]. According to this framework the following elements are needed for a successful evaluation: 1) challenging problem, 2) task definition, 3) training data, 4) test data, 5) evaluation methodology and implementation, 6) evaluator, 7) participants, and 8) funding. In addition to these elements for evaluation, the existing biomedical literature mining systems encounter the issues of portability and scalability, and these issues need to be taken as part of evaluation.

The accuracy of a learning system needs to be evaluated before it can become useful. Limited availability of data often makes estimating accuracy a difficult task [Kohavi 1995]. Choosing a good evaluation methodology is very important for machine learning systems development.

There are several popular methods used for such evaluation, including holdout sampling, cross validation, leave-one-out, and bootstrap sampling [Stone 1974; Efron and Tibshirani 1993]. In the holdout method, data are divided into a training set and a testing set. Usually 2/3 of the data are assigned to the training set and 1/3 to the testing set. After the system is trained by the training set data, the system predicts the output value of each instance in the testing set. These values are then compared with the real output values to determine accuracy.

In cross-validation, a data set is randomly divided into a number of subsets of roughly equal size. Ten-fold cross validation, in which the data set is divided into 10 subsets, is most commonly used. The system is trained and tested for 10 iterations. In each iteration, 9 subsets of data are used as training data and the remaining set is used as testing data. In rotation, each subset of data serves as the testing set in exactly one iteration. The accuracy of the system is the average accuracy over the 10 iterations. Leave-one-out is the extreme case of cross-validation, where the original data are split into n subsets, where n is the size of the original data. The system is trained and tested for n iterations, in each of which n–1 instances are used for training and the remaining instance is used for testing.

In the bootstrap method, n independent random samples are taken from the original data set of size n. Because the samples are taken with replacement, the number of unique instances will be less than n. These samples are then used as the training set for the learning system, and the remaining data that have not been sampled are used to test the system [Efron and Tibshirani 1993].

Each of these methods has its strengths and weaknesses. Several studies have compared them in terms of their accuracies. Hold-out sampling is the easiest to implement, but a major problem is that the training set and the testing set are not independent. This method also does not make efficient use of data since as much as 1/3 of the data are not used to train the system [Kohavi 1995]. Leave-one-out provides the most unbiased estimate, but it is computationally expensive and its estimations have very high variances, especially for small data sets [Efron, 1983; Jain et al. 1987]. Breiman and Spector [1992] and Kohavi [1995] conducted independent experiments to compare the performance of several different methods, and the results of both experiments showed ten-fold cross validation to be the best method for model selection.

In light of the significant medical and patient consequences associated with many biomedical data mining applications, it is critical that a systematic validation method be adopted. In addition, a detailed, qualitative validation of the data mining or text mining results needs to be conducted with the help of domain experts (e.g., physicians and biologists), and therefore this is generally a time-consuming and costly process.

8. CONCLUSIONS

This paper showed that major biomedical ontologies such as UMLS and MeSH served as the domain knowledge provider for text mining techniques. This paper also overviewed important text mining approaches (such as document clustering, document classification, and information extraction). In addition, we introduced several text mining applications to biomedicine and healthcare. For example, we showed how semantic-based document clustering helps physicians for the practice of evidence-based medicine when they perform MEDLINE searches.

An overwhelming amount of biomedical text data is available. However, biomedical text is not readily processed for biomedical knowledge extraction and management due to its unstructured nature and the complexity of the biomedicine domain. Because of these issues, the biomedical information and knowledge in biomedical text have not been easily reused. For this problem, text mining techniques enable us to readily put to use research findings and even infer novel biomedical knowledge or reasonable biomedical hypotheses from the research findings. As a result, text mining is the most promising solution to overcome biomedical text information overload. Text mining has attracted considerable attention even though there are challenging issues of applying text mining techniques to biomedicine and healthcare such as full-text mining, cross-language mining techniques, and interactive text mining systems. As the sheer size of the biomedical text databases continues to grow, the importance of text mining for biomedical literature will be further magnified.

REFERENCES

AGRAWAL, R., ET AL. 2000. Fast Discovery of Association Rules, Advances in Knowledge Discovery and Data Mining, U. Fayyad, et al., Editors. AAAI/MIT Press.

ANDRADE, M. A. AND BORK, P. 2000. Automated extraction of information in molecular biology. FEBS Letters, 476:12–7.

ANDRADE, M., BLASCHKE, C. AND VALENCIA, A. 1999. AbXtract: Automatic Abstract eXtraction of keywords associated to protein function. *Bioinformatics*, 14(7):600–7.

AO, H. AND TAKAGI, T. 2005. ALICE: An algorithm to extract abbreviations from MEDLINE, *Journal of the American Medical Informatics Association*, 12: 576–586.

BAEZA-YATES, R. AND RIBEIRO-NETO, B. 1999. Modern Information Retrieval.

BERGER, A., DELLA PIETRA, S., AND DELLA PIERTA, V. 1999. A maximum entropy approach to natural language processing. Computational Linguistics, Vol. 22, p. 39–71

BLASCHKE, C., ANDRADE, M. A., OUZOUNIS, C., AND VALENCIA, A. 1999. Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions, *In Proceedings of the First International Conference on Intelligent Systems for Molecular Biology,* 60–67.

BODENREIDER, O. 2006. *Lexical, Terminological, and Ontological Resources for Biological Text Mining*, *Text Mining for Biology and Biomedicine*, Ananiadou S. and McNaught J. (eds.), Artech House, 43–66.

BRANK, J., GROBELNIK, M., MILIĆ-FRAYLING, N., AND MLADENIĆ, D. 2002. Interaction of feature selection methods and linear classification models. Proceedings of the ICML-02 Workshop on Text Learning, Sydney, AU.

BROWNE, A., MCCRAY, A., AND SRINIVASAN, S., The SPECIALIST Lexicon, Lister Hill National

Center for Biomedical Communications, National Library of Medicine (NLM), http://lexsrv3.nlm.nih.gov/SPECIALIST/Projects/lexicon/current/release/LEX/DOCS/techrpt.pdf.

BUCKLEY, C. AND LEWIT, A. F. 1985. Optimization of inverted vector searches. In *Proceedings of SIGIR-85*, 97–110.

BUCKLEY, C., SALTON, G., ALLEN, J. AND SINGHAL, A. 1995. Automatic query expansion using SMART: TREC-3. In: D. K. Harman (ed.), The Third Text Retrieval Conference (TREC-3). U.S. Department of Commerce, 69–80.

CHANG, J. T., SCHÜTZE, H. AND ALTMAN, R. B. 2002. Creating an Online Dictionary of Abbreviations from MEDLINE, *The Journal of the American Medical Informatics Association*, 9: 612–620.

COLLIER, N., NOBATA, C., AND TSUJII, J. 2000. Extracting the Names of Genes and Gene Products with a Hidden Markov Model. *Proceedings of the 18th International Conference on Computational Linguistics (COLING2000)*, 201–207.

COWIE, J. AND LEHNERT, W. 1996. Information extraction. *Communications of ACM*, 39:80–91.

CRAMMER, K., AND SINGER, Y. 2001. On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research, Vol. 2, p. 265–292.

CRAVEN, M., AND KUMLIEN, J. 1999. Constructing Biological Knowledge Bases by Extracting Information from Text Sources. *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, 77–86.

CUTTING, D., KARGER, D., PEDERSEN, J., AND TUKEY, J. 1992. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, In *Proceedings of SIGIR '92*, 318–329.

DE BRUIJN, B., AND MARTIN, J. 2002. Getting to the (C)ore of knowledge: mining biomedical literature. *International Journal of Medical Informatics,* (67): 7–18.

DEMETRIOU, G., AND GAIZAUSKAS, R. 2002. Utilizing text mining results: The Pasta Web System. *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*, 77–84.

DING, J., BERLEANT, D., NETTLETON, D., AND WURTELE, E. 2002. Mining MEDLINE: abstracts, sentences, or phrases? *Pacific Symposium on Biocomputing*, 326–337.

DOMINGOS, P., AND PAZZANI, M. 1997. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, in: Machine Learning, Vol. 29:2-3, p. 103–130.

DONALDSON, I., MARTIN, J., DE BRUIJN, B., AND WOLTING, C. 2003. "PreBIND and Textomy– mining the biomedical literature for protein-protein interactions using a support vector machine", BMC Bioinformatics, Vol. 4:11, p. 11–23.

DUMAIS, S. T., PLATT, J., HECKERMAN, D., AND SAHAMI, M. 1998. Inductive learning algorithms and representations for text categorization. Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management, eds. G. Gardarin, J.C. French, N. Pissinou, K. Makki & L. Bouganim, ACMPress, New York, US: Bethesda, US, p. 148–155.

Evidence-Based Medicine Working Group 1992. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA, Nov 1992; 268: 2420–2425.

FAN, W., WALLACE, L., RICH, S., AND ZHANG, Z. 2005. Tapping into the power of text mining, Communications of ACM, forthcoming.

FRIEDMAN, C., KRA, P., YU, H., KRAUTHAMMER, M. AND RZHETSKY, A. 2001. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17 Suppl 1, S74–82.

FUKUDA, K., TAMURA, A., TSUNODA, T., AND TAKAGI, T. 1998. Toward information extraction: identifying protein names from biological papers. *Pacific Symposium on Biocomputing*, 707–18.

GRUBER, T. R. A. 1993. Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, Vol. 5, pp. 199–220.

GRUBER, T. R. 1995. *Towards Principles for the Design of Ontologies used for Knowledge Sharing.* International Journal of Human-Computer Studies, 43, 907–928.

GRUNINGER, M., AND LEE, J. 2002. *Ontology applications and design*, Communications of the

ACM, February, Vol. 45, No. 2, 39–41.

HAHN, U., ROMACKER, M., AND SCHULZ, S. 2002. Creating Knowledge Repositories from Biomedical Reports: The MEDSYNDIKATE Text Mining System. *Pacific Symposium on Biocomputing*, 338–349.

HEARST, M. A., SCHOELKOPF, B., DUMAIS, S., OSUNA, E., AND PLATT, J. 1998. Trends and Controversies-Support Vector Machines, in: IEEE Intelligent Systems, Vol. 13:4, p. 18–28.

HERRMANN, K. 2001. Rakesh Agrawal: Athena: Mining-based Interactive Management of Text Databases, URL: http://www3.informatik.tumuenchen.de/lehre/WS2001/HSEM-bayer/textmining.pdf [as of 2002-03-02].

HIRSCHMAN, L., PARK, J. C., TSUJII, J., WONG, L., AND WU, C. H. 2002. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12): 1553–1561.

HOTHO, A., MAEDCHE, A., AND STAAB, S. 2002. Text clustering based on good aggregations. Künstliche Intelligenz (KI), 16, 4, 48–54.

HRISTOVSKI, D., STARE, J., PETERLIN, B., AND DZEROSKI, S. 2001. Supporting discovery in medicine by association rule mining in Medline and UMLS, Medinfo, 10, 1344–1348.

HRISTOVSKI, D., PETERLIN, B., MITCHELL, J. A., AND HUMPHREY, S. M. 2003. Improving literature based discovery support by genetic knowledge integration, Stud. Health Technol. Inform. 95:68–73.

HUMPHREYS, K., DEMETRIOU, G., AND GAIZAUSKAS, R. 2000. Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. *Pacific Symposium on Biocomputing*, 505–16.

JENSSEN, T. K., et al. 2001. A literature network of human genes for high-throughput analysis of gene expression. Nat. Genet., 28, 21–28.

JENSSEN, T. K., LAEGREID, A., KOMOROWSKI, J., AND HOVIG, E. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–8.

JENSSEN, T. K., LAEGREID, A., KOMOROWSKI, J., AND HOVIG, E. 2001. A literature network of human genes for high-throughput analysis of gene expression, Nature Genetics, Vol. 28, p. 21–28.

JOACHIMS, T. 1998. Text categorization with support vector machines: learning with many relevant features. Proceedings of ECML-98, 10th European Conference on Machine Learning, eds. C. Nédellec & C. Rouveirol, Springer Verlag, Heidelberg, DE: Chemnitz, DE, p. 137–142.

JOACHIMS, T. 1999. Transductive inference for text classification using support vector machines. Proceedings of ICML-99, 16th International Conference on Machine Learning, eds. I. Bratko & S. Dzeroski, Morgan Kaufmann Publishers, San Francisco, US: Bled, SL, p. 200–209.

JOSHI, R., LI, X. L., RAMACHANDARAN, S., AND LEONG, T. Y. 2004. Automatic Model Structuring from Text using BioMedical Ontology, In American Association for Artificial Intelligence (AAAI) Workshop, pp. 74–79, San Jose, California, July.

KARANIKAS, H., AND THEODOULIDIS, B. 2002. Knowledge discovery in text and text mining software, Technical report, UMIST−CRIM, Manchester.

KAUFMAN, L., AND ROUSSEEUW, P. J. 1999. Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons.

KOLLER, D., AND SAHAMI, M. 1997. Hierarchically classifying documents using very few words. In *Proceedings of ICML-97*, 170–176.

KRAUTHAMMER, M., RZHETSKY, A., MOROZOV, P., AND FRIEDMAN, C. 2000. Using BLAST for identifying gene and protein names in journal articles. *Gene*, 259(1-2): 245–252.

LEE, K., HWANG, Y., AND RIM, H. 2003. Two-Phase Biomedical NE Recognition based on SVMs. *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, 33–40.

LEEK, T. R. 1997. Information Extraction Using Hidden Markov Models. *MSc Thesis,* Department of Computer Science, University of California, San Diego.

LEOPOLD, E., AND KINDERMANN, J. 2002. Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? Machine Learning, Vol. 46:1-3, p. 423–444.

LIN, J., AND DEMNER-FUSHMAN, D. 2007. Semantic Clustering of Answers to Clinical Questions, Proceedings of the 2007 Annual Symposium of the American Medical Informatics Association (AMIA 2007), Chicago, Illinois, pp. 458–462.

LIU, F., JENSSEN, T. K., NYGAARD, V., SACK, J., AND HOVIG, E. 2004. FigSearch: Using Maximum Entropy Classifier to Categorize Biological Figures. Proceedings of IEEE Computational Systems Bioinformatics Conference, p. 476–477

LIU, H. AND FRIEDMAN, C. 2003. Mining Terminological Knowledge in Large Biomedical Corpora, *Proceedings of the Pacific Symposium on Biocomputing*, 8: 415–426.

MOONEY, R. J., AND NAHM, U. Y. 2003. Text Mining with Information Extraction, Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, 22-23 September, Bloemfontein, South Africa, pp. 141–160.

NARAYANASWAMY, M., RAVIKUMAR, K. E. AND VIJAY-SHANKER, K. 2003. A biological named entity recognizer. *Pacific Symposium on Biocomputing*, 427–438.

National Library of Medicine (NLM), MEDLINE, http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=File&db=PubMed, 2008.

National Library of Medicine (NLM), Medical Subject Headings (MeSH) Fact Sheet, http://www.nlm.nih.gov/pubs/factsheets/mesh.html, 2008.

National Library of Medicine (NLM), Unified Medical Language System (UMLS) Fact Sheet, http://www.nlm.nih.gov/pubs/factsheets/umls.html, 2008.

NG, S. K., AND WONG, M. 1999. Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Informatics Series: Workshop on Genome Informatics*, 10: 104–112.

OHTA, Y., YAMAMOTO, Y., OKAZAKI, T., UCHIYAMA, I., AND TAKAGI, T. 1997. Automatic construction of knowledge base from biological papers. *Proceedings of International Conference on Intelligent System for Molecular Biology*, 5:218–25.

PAKHOMOV, S. V., RUGGIERI, A., AND CHUTE, C. G. 2002. Maximum entropy modeling for mining patient medication status from free text, Proc AMIA Symp, p. 587–91

PANT, G., AND SRINIVASAN, P. 2005. Learning to crawl: Comparing classification schemes. ACM Transactions on Information Systems, Vol. 23, p. 430–462.

PANTEL, P., AND LIN, D. 2002. Document clustering with committees. In Proceedings of the 2002 ACM SIGMOD International Conference on Management of data, 199–206.

PARK, J. C., KIM, H. S., AND KIM, J. J. 2001. Bidirectional Incremental Parsing for Automatic Pathway Identification with Combinatory Categorical Grammar, *Pacific Symposium on Biocomputing,* 396–407.

PEREZ-IRATXETA, C., BORK, P., AND ANDRADE, M. A. 2002. Association of genes to genetically inherited diseases using data mining. Nat. Genet., 31, 316–319.

PRATT, WANDA AND YETISGEN-YILDIZ, Meliha, 2003. LitLinker: capturing connections across the biomedical literature, K-CAP'03, pp. 105-112, Sanibel Island, FL, Oct. 23–25.

PRATT, W., AND FAGAN, L. 2000. The Usefulness of Dynamically Categorizing Search Results, Journal of the American Medical Informatics Association, 7(6), pp. 605–617.

PRATT, W., HEARST, M., AND FAGAN, L. 1999. A knowledge-based approach to organizing retrieved documents, AAAI '99: Proceedings of the 16th National Conference on Artificial Intelligence, Orlando, Florida, pp. 80–85.

PROUX, D., RECHENMANN, F., AND JULLIARD, L. 2000. A pragmatic information extraction strategy for gathering data on genetic interactions. *Proceedings of International Conference on Intelligent System for Molecular Biology*, 8:279–85.

PUSTEJOVSKY, J., CASTANO, J., ZHANG, J., KOTECKI, M., AND COCHRAN, B. 2002. Robust relational parsing over biomedical literature: extracting inhibit relations. *Pacific Symposium on Biocomputing*, 362–73.

RAY, S. AND CRAVEN, M. 2001. Representing Sentence Structure in Hidden Markov Models for Information Extraction. *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, Seattle, WA. Morgan Kaufmann.

RAYCHAUDHURI, S., CHANG, J. T., SUTPHIN, P. D., AND ALTMAN, R. B. 2002. Associating genes with Gene Ontology codes using a maximum entropy analysis of biomedical literature, Genome Research, Vol. 12, p. 203–14.

RINDFLESCH, T. C., TANABE, L., WEINSTEIN, J. N., AND HUNTER, L. 2000. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pacific Symposium on Biocomputing*, 517–28.

RISH, I., An empirical study of naïve Bayes classifier. In IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, p. 41–46.

SCHWARTZ, A. S., AND HEARST, M. A. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text, *Proceedings of the Pacific Symposium on Biocomputing*, 8: 451–462.

SEBASTIANI, F. 2002. Machine Learning in automated text categorization. ACM Computing Surveys, Vol. 34, p. 1–47.

SHATKAY, H., AND FELDMAN, R. 2003. Mining the biomedical literature in the genomic era: An overview. *Journal of Computational Biology*, 10(6): 821–855.

SIOLAS, G., AND D'ALCHÉ-BUC, F. 2000. Support Vector Machines based on a semantic kernel for text categorization, in: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00), p. 205–209.

SONG, M., AND YOO, I. 2007. A Hybrid Abbreviation Extraction Technique for Biomedical Literature, accepted in 2007 IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM 2007), San Jose, CA, USA, Nov. 2–4.

SRINIVASAN, P. 2004. Text mining: Generating hypotheses from MEDLINE, Journal of the American Society for Information Science, Vol. 55, No. 4, pp. 396–413.

STAPLEY, B. J., AND BENOIT, G. 2000. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in MEDLINE abstracts. *Pacific Symposium on Biocomputing*, 529–40.

STAPLEY, B. J., KELLEY, L. A., AND STERNBERG, M. J. E. 2002. Predicting the sub-cellular location of proteins from text using support vector machines, Pacific Symposia in Biocomputing, p. 374–85.

STEINBACH, M., KARYPIS, G., AND KUMAR, V. 2000. A comparison of document clustering techniques. Technical Report #00-034. Department of Computer Science and Engineering, University of Minnesota.

SWANSON, D. R. 1986. Undiscovered public knowledge. Libr. Q. 56(2):103-118.

SWANSON, D. R. 1987. Two medical literatures that are logically but not bibliographically connected. JASIS, 38(4):228–233.

SWANSON, D. R., AND SMALHEISER, N. R. 1999. Implicit text linkages between Medline records: Using Arrowsmith as an aid to scientific discovery. Library Trends, 48(1):48–59.

TANABE, L., SCHERF, U., SMITH, L. H., LEE, J. K., HUNTER, L., AND WEINSTEIN, J. N. 1999. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, 27(6), 1210–4, 1216–7.

THOMAS, J., MILWARD, D., OUZOUNIS, C., PULMAN, S., AND CARROLL, M. 2000. Automatic extraction of protein interactions from scientific abstracts. *Pacific Symposium on Biocomputing*, 541–52.

VAN RIJSBERGEN, C. J. 1979. *Information Retrieval, 2nd edition*, London: Buttersworth.

VAPNIK, V. N. 1995. The nature of statistical learning theory. Springer Verlag: Heidelberg, DE.

WANG, B. B., MCKAY, R. I., ABBASS, H. A., AND BARLOW, M. 2002. Learning Text Classifier using the Domain Concept Hierarchy. In *Proceedings of International Conference on Communications, Circuits and Systems 2002*, China.

WITTEN, I. H., AND FRANK, E. 2000. Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers: San Francisco.

YAKUSHIJI, A., TATEISI, Y., MIYAO, Y., AND TSUJII, J. 2001. Event extraction from biomedical papers using afull parser. *Pacific Symposium on Biocomputing*, 408–19.

YAMAMOTO, K., KUDO, T., KONAGAYA, A., AND MATSUMOTO, Y. 2003. Protein Name Tagging for Biomedical Annotation in Text. *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, 65–72.

YANG, Y., AND LIU, X. 1999. A Re-Examination of Text Categorization Methods, in: Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, p. 42–49.

YOO, I., HU, X., AND SONG, I.-Y. 2007. A Coherent Graph-based Semantic Clustering and Summarization Approach for Bi/omedical Literature and a New Summarization Evaluation Methods, BMC Bioinformatics, 8(Suppl 9):S4.

YOO, I., HU, X., AND SONG, I.-Y. 2006. Integration of Semantic-based Bipartite Graph Representation and Mutual Refinement Strategy for Biomedical Literature Clustering, in the 12th SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 791-796, Philadelphia, USA, August 20–23.

YU, H., HRIPCSAK, G., AND FRIEDMAN, C. 2002. Mapping abbreviations to full forms in biomedical articles, *Journal of the American Medical Informatics Association*, 9: 162–172.

ZAMIR, O., AND ETZIONI, O. 1998. Web Document Clustering: A Feasibility Demonstration, In *Proceedings of SIGIR 98*, 46–54.

ZU EISSEN, S. M., STEIN, B., AND POTTHAST, M. 2005. The suffix tree document model revisited, In Proceedings of the 5th International Conference on Knowledge Management, 596–603.

Illhoi Yoo is an assistant professor of the department of Health Management and Informatics, University of Missouri-Columbia School of Medicine. He gained his M.S. degree in information science from University of Pittsburgh and his Ph.D. in information science and technology from Drexel University. His research interest is in semantic-oriented MEDLINE search and biomedical literature mining using biomedical ontologies (UMLS and MeSH). He has published more than 25 peer-reviewed papers in various journals, conferences and books for the last three years. He won the best paper award as the co-author of a paper published in the Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (IEEE CIBCB 2004) on Oct. 7, 2004.

Min Song is an assistant professor of Department of Information Systems at NJIT. He received his M.S. in School of Information Science from Indiana University in 1996 and received Ph.D. degree in Information Systems from Drexel University in 2005. Min has a background in Text Mining, Bioinfomatics, Information Retrieval and Information Visualization. Min received the Drexel Dissertation Award in 2005. In 2006, Min's work received an honorable mention award in the 2006 Greater Philadelphia Bioinformatics Symposium. In addition, The paper entitled "Extracting and Mining Protein-protein interaction Network from Biomedical Literature" has received the best paper award from 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, which was held in San Diego, USA, Oct. 7-8, 2004. In addition, another paper entitled "Ontology-based Scalable and Portable Information Extraction System to Extract Biological Knowledge from Huge Collection of Biomedical Web Documents" was nominated as the best paper at 2004 IEEE/ACM Web Intelligence Conference, which was held in Beijing, China, Sept. 20-24, 2004.