HLT-NAACL 2004 Workshop: Biolink 2004,
Linking Biological Literature, Ontologies and Databases, pp. 41-48.
Association for Computational Linguistics.

Clustering MeSH Representations of Biomedical Literature

Craig A. Struble, Chitti Dharmanolla
Department of Mathematics, Statistics, and Computer Science
Marquette University
Milwaukee, WI 53201-1881
{craig.struble,chittithall.dharmanolla}@marquette.edu

Abstract

Biomedical literature contains vital information for the analysis and interpretation of experiments in the biological sciences. Human reasoning is the primary method for extracting, synthesizing, and interpreting the results contained in the literature, yet the rate at which publications are produced is exponential. With the advent of digital, full-text publication and increasing computational power, automated techniques for knowledge discovery and synthesis are being developed to assist humans in making sense of growing literature databases.

We investigate the use of ontological information provided by the Medical Subject Headings (MeSH) project to discover groupings within a collection of medical literature stored in PubMed. Vector representations of documents based on MeSH terms are presented. Results of agglomerative hierachical clustering on two collections of biomedical literature, the Rat Genome Database and Tourette's Syndrome related research, suggest novel and understandable groupings are obtainable.

1 Introduction

In recent years the amount of online documents has grown tremendously that poses challenges for information retrieval from this vast collection. *Text mining* is the application of techniques of machine learning in conjunction with natural language processing, information extraction and algebraic/mathematical approaches to computational information retrieval (Berry and Pottenger, 2003).

Two major subfields of text mining are *document classification* and *document clustering*. Document classification is the automated assignment of textual data to groups or classes. Supervised machine learning techniques, such as neural networks or nearest neighbor classifiers, are often employed in document classification. Document clustering identifies groups of similar documents based on shared features, typically words contained in the documents. This differs from document classification in that topic areas are unknown before clustering.

An important consideration for document classification and document clustering is the representation of the documents for analysis. Traditional approaches represent documents by extracting features from the full-text contents of each document. These features may undergo transformations such as *weighting* or *dimension reduction* with the goal of improving classification accuracy, improving clustering quality, or data reduction.

Our goal is to explore Medical Subject Headings (MeSH), a controlled vocabulary for describing medical literature (National Library of Medicine, 2003), as features for document representation. Exploring this use of MeSH is important for two reasons. First, MeSH terms are assigned to papers by trained indexers, thus many issues involved with natural language processing may be avoided. Second, insights gained with MeSH based representations may be applied to other ontologies under development such as the Gene Ontology (The Gene Ontology Consortium, 2000).

In this paper, we focus on the interplay between MeSH based representations and document clustering. Our application of document clustering is to identify and summarize potential topics within collections of medical literature. The outline of the rest of the paper is as follows. Section 2 discusses methods for obtaining document collections. Representations of documents, including our proposed MeSH representations are described in Section 3. Section 4 outlines the document clustering approach used in our study. Results from a comparative study and an exploratory study are presented in Section 5. Section 6 contains a survey of related work. Conclusions and future opportunities are discussed in Section 7.

2 Document Collections

Collections of documents can be obtained by several means. In the simplest situation, any sample of documents contained in PubMed can be obtained for the purposes of document clustering. Such a sampling may provide insight into the whole of PubMed, but is most likely not useful for specific text mining tasks.

A more useful approach for targeted text mining is to build a query or collection of queries centered around a *concept*. For example, in studying prostate cancer, the query string *prostate cancer* is given to PubMed. The documents matching the query for prostate cancer are retrieved and processed for document clustering. The identified clusters represent potential topics contained in prostate cancer research. This approach has been used to build *concept profiles* for several text mining tasks (Srinivasan and Wedemeyer, 2003; Srinivasan, to appear).

Other possible methods for obtaining document collections exist as well. In obtaining documents for a genome database, such as the Rat Genome Database (RGD) (Twigger et al., 2002), human curators combine queries of PubMed with an exhaustive reading of a limited number of journals. This may be viewed as another form of a concept-based collection. In this case, however, the collection captures several ill defined concepts; ones that cannot be specified with a small number of PubMed queries.

This investigation considers both methods of obtaining document collections.

3 Document Representations

Representing documents for clustering and other text mining tasks is a fundamental step in the knowledge discovery process. The ability to derive useful information from a document collection may be entirely determined by the attributes used to describe the documents. A commonly used representation in text mining and information retrieval is the *vector* representation. A summary of vector representations is presented below and refer the reader to a text on information retrieval (Korfhage, 1997) for a more detailed description.

Suppose D is a collection of documents and $T = \{t_1, t_2, \ldots, t_n\}$ is the collection of unique terms appearing in at least one document in D. Obtaining T is typically accomplished by extracting individual words (e.g., characters between spaces) from the text (e.g. titles, abstracts, and body) of each paper, although more sophisticated parsing may occur. Individual words may be further processed by *stop word removal*, the removal of words without inherent meaning such as articles or pronouns, and *stemming*, the removal of suffixes to extract only root words. This term processing often generates better classification and information retrieval results.

Given T, a document $d \in D$ is represented as a vector

$$v_d \quad = \quad \langle w_1, w_2, \ldots, w_m \rangle, \qquad (1)$$

where w_i is called the *weight* of term t_i within document d. Weights are defined based on specific application needs.

Two examples of commonly used weighting schemes are term frequency (TF) and term frequency inverse document frequency (TFIDF). Let $|t_i|$ be the number of times t_i appears in a document d, $|D|$ be the number of documents in the document collection, and n_i be the number of documents in D containing t_i. The TF scheme is defined by $w_i = |t_i|$. The TFIDF scheme is defined by $w_i = |t_i| / \log_2(|D|/n_i)$.

Consider a document collection D with term collection $T = \{$*cancer*, *diagnosis*, *medical*, *viral*$\}$. If a document d contains three occurences of the term *cancer*, one occurence of the term *diagnosis*, four occurences of the term *medical*, and no occurences of the term *viral*. The representation of d using TF weighting is

$$v_d \quad = \quad \langle 3, 1, 4, 0 \rangle.$$

3.1 MeSH Representations

This investigation builds on the vector space representation of documents described above. Instead of obtaining a term collection T from the full text of titles, abstracts, or content of a paper, T is built from the MeSH assignments for each document. A summary of MeSH is given below.

Medical literature is indexed by MeSH terms by the National Library of Medicine (NLM) for the purpose of subject indexing and searching of journal articles in PubMed (an online literature database that contains citations from more than 4,600 biomedical journals). MeSH terms are assigned to medical literature by human indexers.

MeSH consists of two ontologies: *descriptors* or *headings* are a collection of terms for primary themes or topics contained in the literature; and *qualifiers* or *subheadings* are terms combined with descriptors to indicate the specific aspect of a descriptor. Formally, a *MeSH term* is a tuple (d, q) where d is a descriptor and q is a qualifier (q may be empty if d is unqualified). There exist 21975 descriptors and 83 qualifiers in the 2003 MeSH ontology, which was used in this study.

Both descriptors and qualifiers are organized in directed acyclic graphs (DAGs), where the parent of a descriptor or qualifier is considered more general than the term itself. A descriptor (or qualifier) may have multiple parents, representing that the descriptor (or qualifier) includes multiple concepts in the MeSH ontology simultaneously. For example, in the 2003 MeSH ontology, descriptors have an average of approximately 1.8 parents.

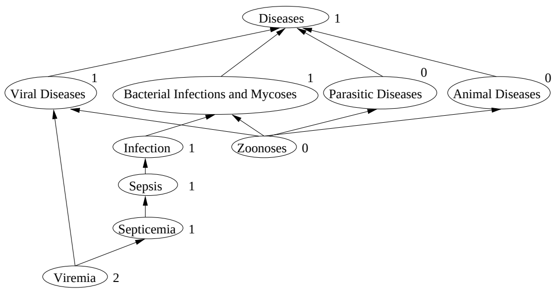

Figure 1: A portion of the MeSH descriptor ontology. The numbers indicate the term weighting if *Virema* is assigned to a document d.

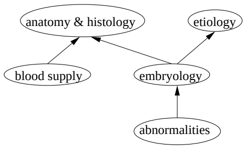

Figure 2: A portion of the MeSH qualifier ontology.

Portions of the descriptor and qualifier ontologies are displayed in Figures 1 and 2.

In MeSH representations, weights are derived from the structure of MeSH. Documents are represented as vectors where the term collection T consists of *descriptors* only, *qualifiers* only, or *combined descriptors and qualifiers* (this will be further referred to as the *combined* representation). Weights are defined by

$$w_i \quad = \quad \begin{cases} 0 & \text{if term } t_i \text{ is not assigned} \\ 1 & \text{if term } t_i \text{ is inferred} \\ 2 & \text{if term } t_i \text{ is assigned} \end{cases} \quad .$$

A term is *inferred* if one of its descendants in the MeSH hierarchy is assigned, but the term itself is not assigned.

Consider d with the term *Viremia* assigned. The descriptors only representation is

$$v_d \quad = \quad \langle 0,1,1,1,0,1,1,1,2,0 \rangle,$$

where the columns correspond to *Animal Diseases, Bacterial Infections and Mycoses, Diseases, Infection, Parasitic Diseases, Sepsis, Septicemia, Viral Diseases, Viremia*, and *Zoonoses* respectively. The relationship between the MeSH hierarchy and the values assigned is demonstrated in Figure 1. In essence, the DAG structure is *flattened*, but allowable vectors for document representation are restricted to the structure imposed by MeSH.

4 Document Clustering

Many clustering algorithms have been proposed for document clustering. In this study, AGNES (Kaufman and Rousseeuw, 1990), an agglomerative hierarchical clustering algorithm, with average linking was employed. Using this algorithm has two advantages for this study. First, dendrograms, a visualization of the substructures contained in a document collection, are produced. Second, AGNES computes an *agglomerative coefficient* a. Let m_d be the height at which d is first merged, and M is the height of the final merge, then

$$a = \operatorname*{mean}_{d \in D} \left(1 - \frac{m_d}{M} \right).$$

Intuitively, the agglomerative coefficient measures the average similarity of d to the members of the first cluster containing d, normalized to a $[0,1]$ range. For document collections of approximately equal size, a larger a indicates better clustering quality (Kaufman and Rousseeuw, 1990).

4.1 Dimension Reduction

The number of unique terms in document collection is typically large (> 1000), resulting in very high dimensional data. Dimension reduction is commonly employed in text mining before further analysis.

Principal components analysis (PCA) and related approaches are methods for dimension reduction (Jolliffe, 1986). A full discussion of PCA is beyond the scope of this paper. Several guidelines exist for PCA to determine the number of dimensions to use. In this study, principal components are selected in descending order until 25% of the variation in the data is captured.

4.2 Document Similarity

Many clustering algorithms require a measure of *similarity* between two documents be defined. Euclidean distance is one measure used in clustering applications. Another measure, used in information retrieval, is the cosine measure (Korfhage, 1997), which measures similarity by calculating the cosine of the angle between the vector representation of two documents. Cosine distance is used in this paper.

4.3 Cluster Identification and Summarization

For MeSH representations, clusters are identified and summarized to find interesting groups in the document collection. Individual clusters are identified by cutting the dendrogram at different heights. The clusters are then summarized by computing the *cluster center*, a vector consisting of the mean term weights across constituent documents, using the full dimensional representation. Terms are ranked in descending order according to the resulting mean weight.

5 Experiments and Results

Two document collections were analyzed using document clustering: documents in RGD (Twigger et al., 2002), and documents retrieved by the PubMed query "Tourette's Syndrome." Each data set is described in more detail below.

The following procedure was employed for each collection.

1. Documents are encoded in a vector representation. The term collection T is derived from terms in abstracts and titles, MeSH descriptors, MeSH qualifiers, or a combination of MeSH descriptors and qualifiers.

   For full-text, terms from abstracts and titles were obtained using `rainbow` with stop word removal and stemming options (McCallum, 1996). TF weighting was used.

   For the MeSH descriptors and qualifiers, the assignments were obtained from PubMed XML entries, and inferring was determined by the 2003 MeSH.

2. PCA was performed on the represented documents, and principal components capturing 25% of the data variance were selected. The documents were projected onto the selected components.

3. The reduced dimension representation was clustered by AGNES using average linking. The cosine distance measure was used for document similarity.

4. Clusters were identified and summarized.

Computations were performed using R version 1.7.1 (R Development Core Team, 2003). Clustering was accomplished using the `agnes` function in the `cluster` package. PCA calculations used the `prcomp` function in the `mva` package.

5.1 Rat Genome Database

The Rat Genome Database (RGD) is a NIH (National Institutes of Health) project developed at Medical College of Wisconsin (MCW) whose main objective is to collect, consolidate and integrate data generated from rat research (Twigger et al., 2002). Rat is the dominant preclinical model organism used to study human diseases involving heart, lung, kidney, blood and vasculature, such as hypertension and renal failure. Researchers at MCW curate approximately 200 articles from 30 journals every month. This is a small portion of the 1200 articles published on rat research every month. The concepts embodied by this document collection are ill defined. Several conversations with the RGD curators resulted in no clear specification of interests or search terms.

	Document Representaion			
#	**Descriptors**	**Qualifiers**	**Combined**	**Full-text**
1	0.9575034	0.9999974	0.9920269	0.9183985
2	0.9568182	0.9999954	0.9919998	0.9196760
3	0.9575754	0.9999963	0.9926404	0.9216638
4	0.9597954	0.9999977	0.9926714	0.9219455
5	0.9594162	0.9999967	0.9923353	0.9212369
6	0.9574500	0.9999971	0.9920885	0.9192635
7	0.9570196	0.9999883	0.9921626	0.9169283
8	0.9561051	0.9999972	0.9920506	0.9168467
9	0.9567221	0.9999963	0.9923461	0.9176114
10	0.9591231	0.9999945	0.9921960	0.9197660
11	0.9557686	0.9999958	0.9922536	0.9196875
12	0.9552862	0.9999971	0.9922055	0.9165249
13	0.9567133	0.9999963	0.9918659	0.9190341
14	0.9557888	0.9999955	0.9917869	0.9157648
15	0.9583430	0.9999974	0.9926036	0.9177084
16	0.9590242	0.9999968	0.9929164	0.9200056
17	0.9568303	0.9999974	0.9920061	0.9187594
18	0.9554807	0.9999956	0.9922523	0.9160838
19	0.9566919	0.9999966	0.9923793	0.9187453
20	0.9592093	0.9999971	0.9925647	0.9240020

Table 1: Agglomerative coefficients from 20 bootstrap samples.

A comparative study of full text (abstracts and titles), MeSH descriptors, MeSH qualifiers, and a combined MeSH descriptors and qualifiers representation was performed. The document collection consists of 2713 papers. The term collection T for the full-text representation contained 17177 unique terms after stemming and stop word removal; and for the MeSH representations, T contained 5013 descriptors and 64 qualifiers. After PCA, the number of principal components used for the descriptors, qualifiers, combined, and full-text representations are 16, 2, 62, and 37 respectively.

The clustering quality of each representation was evaluated using 20 bootstrap samples (i.e., sampling with replacement) of size 2713 from the 2713 documents. Each sample was represented and clustered. The resulting agglomerative coefficients were tabulated (Table 1). To show a significant difference in the agglomerative coefficients obtained between MeSH representations and the full-text representation, the *Wilcoxon rank sum* test, a non-parametric version of the paired t-test, was applied. The p-values in Table 2 indicate that each of the MeSH representations are significantly different than the full-text representation. By observing that larger agglomerative coefficients indicate higher quality clustering, we conclude that MeSH representations offer higher quality clustering than the full-text representation.

The full text and combined MeSH representations are further explored. Dendrograms for the full text representation (Figure 4) and combined representation (Figure 3) show the structure of the document collection. The combined representation results in two clearly distinct clus-

Comparison	p-value
Descriptors and Full-text	1.451e-11
Combined and Full-text	1.451e-11
Qualifiers and Full-text	1.907e-06

Table 2: Results of Wilcoxon rank sum tests.

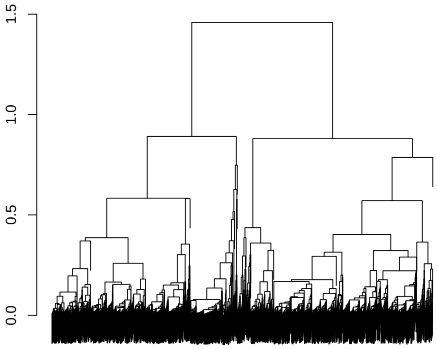

Figure 3: Dendrogram using combined MeSH representation, average linking, and cosine distance. The vertical axis represents the intercluster distance, or height, at which the clusters are merged.

ters identified at height 1.0. Furthermore, the tree contains several small and tight clusters at a low height, indicating the existence of possible subconcepts. In contrast, the resulting tree for the full text representation does not reveal the same structure, suggesting subconcepts are not clearly identified.

Depicted in Figures 5 and 6 are two dimensional scatterplots of the documents projected on the first two principal components of the combined representation and full-text representation respectively. These plots also show a structure with the descriptors and qualifiers representation, there are two distinguished clusters with few outliers. The two clusters in the dendrogram of the combined representation correspond to the left and right groups seen in the scatterplot.

Table 3 presents summary description of the clusters found for the combined representation. Terms with a weight > 0.5 are included. The summary describes the two major groups of papers: one related to sequence and molecular techniques; the other related to metabolism, biochemical phenomena and physiology.

5.2 Tourette's Syndrome

A second, exploratory study was performed on a document collection about the disease *Tourette's Syndrome*. Only the results of using the combined representation are presented here.

Tourette's syndrome is neurological disorder characterized by motor and vocal tics and associated behavioral abnormalities. Chromosomes 2, 7, 11, and 18 have been implicated in causal effects of the disease (OMIM, 2003).

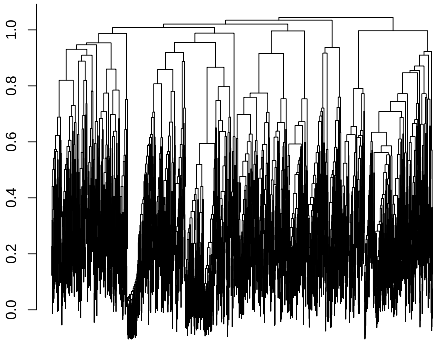

Figure 4: Dendrogram using full-text representation, average linking, and cosine distance. The vertical axis represents the intercluster distance, or height, at which the clusters are merged.

The collection was obtained using the query "Tourette's Syndrome" on PubMed, resulting in 2241 papers. The term collection for the combined representation consists of 6524 MeSH descriptors and 76 MeSH qualifiers. Only 8 principal components were required to capture 25% of the variance in the data set.

Figure 7 contains the resulting dendrogram. Three distinct clusters of documents exist at a height of 1.0. The leftmost cluster in the tree could be split again at a height of approximately 0.9. The clusters at lower heights are not as tightly defined as those in the RGD study, indicating more diversity in the document contents.

Summaries of the three clusters are given in Table 4. In all three clusters, terms associated with Tourettes Syndrome appear with a weight > 0.5 in the cluster center. Documents in the left cluster appear to focus on the psychology and diagnosis associated with the disease, discussing all age groups and genders. The middle cluster consists of papers associated with the genetics and physiopathological diagnosis of Tourette's Syndrome. Of particular interest is the lack of age and gender terms, meaning the papers do not represent consistent themes in ages or genders. Papers associated with drug therapy and pharmacological studies comprise the right cluster, again spanning all age groups and genders. It should be noted that Tourette's Syndrome patients show a therapeutic response to Haloperidol (OMIM, 2003).

The three identified clusters are represented by 1, 2 (in the bottom center of the plot), and 3 in Figure 8, a scatterplot projected onto the first two principal components. The scatterplot along the first two principal components show a correspondence to the dendrogram: 1's

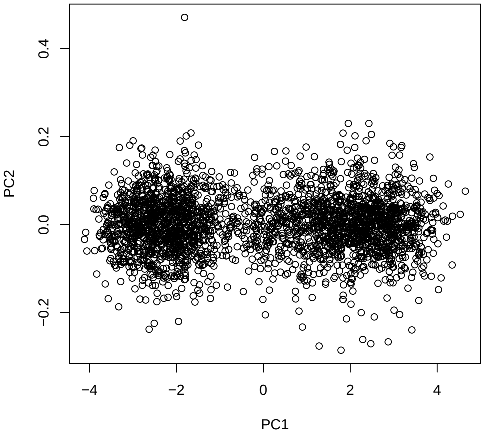

Figure 5: Two-dimensional scatterplot of documents using the combined MeSH representation. The x and y axes are the first two principal components.

correspond to the left cluster in the tree; 2's to the middle cluster; and 3's to the right cluster. The scatterplot suggests the existence of smaller clusters, which agrees with the hierarchical clustering results.

6 Related Work

Srinivasan has extensively investigated the use of MeSH for classification and text mining (Srinivasan, 2001; Ruiz and Srinivasan, 2002; Srinivasan and Rindflesch, 2002; Ruiz and Srinivasan, 2003; Srinivasan and Wedemeyer, 2003; Srinivasan, to appear). Of particular interest is the work on concept profiles to provide targeted summaries of document collections. In comparison to our work, concept profiles provide a global insight into a document collection, whereas document clustering can provide insight into important groups within a document collection.

Document clustering of medical literature in full-text representations has been used for functional annotation of gene products (Renner and Aszódi, 2000) and concept discovery (Iliopoulos et al., 2001). In the latter paper, the authors ignore MeSH, arguing that it is not updated or may not capture the document contents. In our study, we found MeSH indexed documents without abstracts, suggesting that clustering with MeSH terms is complementary work. MeSH descriptors have been considered as additional features in document clustering (Wilbur, 2002), but the hierarchical relationships of MeSH are not used.

Ontology-based clustering has been considered (Hotho et al., 2001). In this work, terms are selected from the ontology based on frequency, employing the parent-child relationships. Adapting this work to MeSH may be inter-

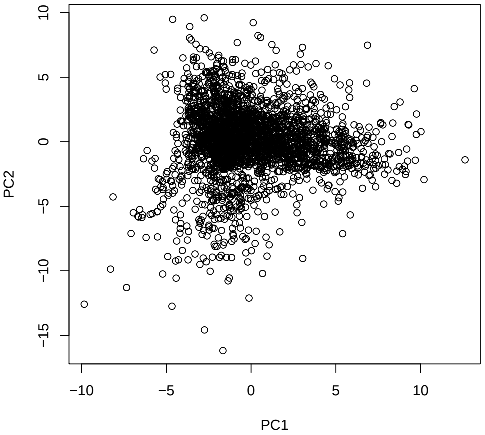

Figure 6: Two-dimensional scatterplot of documents using the full-text representation. The x and y axes are the first two principal components.

Cluster	Terms
Left	Animal; Rats; Support, Non-U.S. Gov't; Muridae; Male; Support, U.S. Gov't, P.H.S.; Rats, Sprague-Dawley; RNA, Messenger; Cells, Cultured; chemistry; cytology; drug effects; etiology; genetics; metabolism; pharmacology; physiology
Right	Animal; Molecular Sequence Data; Rats; Amino Acid Sequence; Base Sequence; Support, Non-U.S. Gov't; Cloning, Molecular; Muridae; Molecular Structure; Documentation; Human; Sequence Homology, Amino Acid; RNA, Messenger; Support, U.S. Gov't, P.H.S.; DNA, Complementary; Genetic Structures; Genetic Techniques; Proteins; Sequence Homology; Mice; analysis; chemistry; etiology; genetics; metabolism; physiology

Table 3: A summary of the two clusters defined at height 1.0 of the agglomerative clustering results for the RGD document collection in terms of descriptors (capitalized) and qualifiers (lower-case).

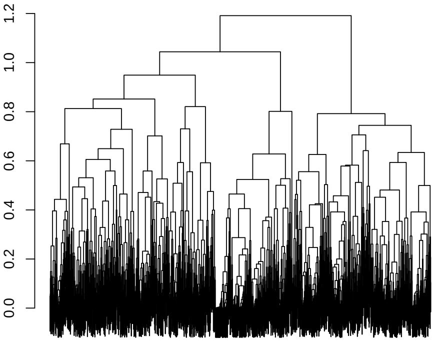

Figure 7: Dendrogram of Tourette's Syndrome document collection using the combined MeSH representation, average linking, and cosine distance. The vertical axis represents the intercluster distance, or height, at which the clusters are merged.

esting.

A distance measure using MeSH has been introduced as part of an algorithm to assign MeSH terms (Ontrup et al., 2003). The distance measure considers a tree based representation; the tree induced by assigned MeSH terms. To overcome combinatorial issues, representative subtrees are constructed. Distance is measured via complements and overlaps of representative subtrees.

Transitive relationships between genes and disease with MeSH terms have also been explored (Hristovski and Peterlin, 2002; Perez-Iratxeta et al., 2002), using association rules and fuzzy logic models, respectively.

7 Conclusions and Future Work

The presented results indicate that MeSH representations are useful for document clustering. MeSH representations provide better quality AGNES clustering than full text representations using TF weighting. Furthermore, clusters are easily summarized and the summaries can be readily interpreted in the context of the document collection.

It is quite surprising that using principal components covering only 25% of the variance provides such obvious structure. Even the first two principal components elicits structure in the two document collections tested here.

Many directions exist for improving MeSH representation. The representations may lose information embedded in the DAGs, since relationship between descriptors (or qualifiers) and their parents are not strictly maintained. Explicit associations between qualifiers and descriptors were removed to simplify representation; these should be reintroduced. MeSH "major themes", an annotation indicating emphasis on a term, remains to be incor-

Cluster	**Terms**
Left	Human; Tourette Syndrome; Male; Female; Adolescent; Adult; Tic Disorders; Child; Basal Ganglia Diseases; Heredodegenerative Disorders; Age Groups; Case Report; Support, Non-U.S. Gov't; diagnosis; etiology; physiology; psychology
Middle	Human; Tourette Syndrome; Basal Ganglia Disorders; Heredodegenerative Disorders, Nervous System; diagnosis; etiology; genetics; physiology; physiopathology
Right	Tourette Syndrome; Human; Male; Tic Disorders; Basal Ganglia Disorders; Heredodegenerative Disorders, Nervous System; Child; Female; Adolescent; Adult; Age Groups; Haloperidol; Case Report; drug therapy; etiology; pharmacology; therapeutic use; therapy; adverse effects

Table 4: A summary of the three clusters defined at height 1.0 of the agglomerative clustering results for the Tourette's Syndrome document collection in terms of descriptors (capitalized) and qualifiers (lower-case).

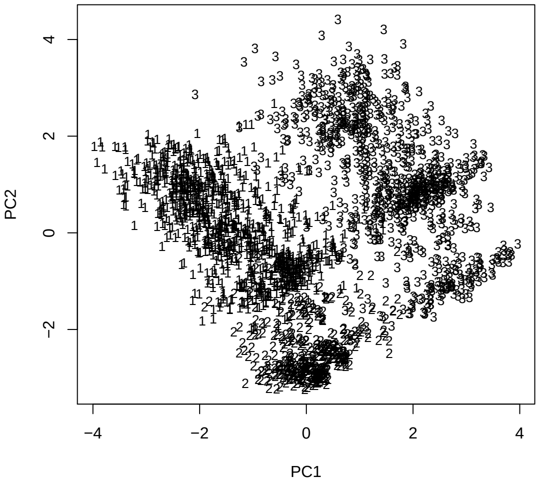

Figure 8: Two-dimensional scatterplot of documents using the combined MeSH representation. The x and y axes are the first two principal components.

porated.

The summarization approach is straightforward, but presents terms that are not insightful. For example, *Rats* is frequently a term with high ranking, bus is not informative in the RGD context. Similar observations have been previously made (Kankar et al., 2002). Term weighting and more flexibility in summarization should help.

MeSH representations have disadvantages compared to full text. The manual curation process requires several weeks for indexing. Yearly revision of MeSH implies systems must adapt to changes. Full text in abstracts and papers contain more precise information. We feel, however, that combining ontology and full text representations should be beneficial.

8 Acknowledgements

We would like to thank Susan Bromberg and the Rat Genome Database team for document information, discussions of curation tasks, and visitation space.

References

Michael W. Berry and William M. Pottenger. 2003. Theme statement. In *Proceedings of the Workshop on Text Mining, SIAM Third International Conference on Data Mining*, San Francisco, CA, May 3. SIAM.

A. Hotho, A. Maedche, and S. Staab. 2001. Ontology-based text clustering.

Dimitar Hristovski and Borut Peterlin. 2002. Improving literature based discovery support by background knowledge integration. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*.

I. Iliopoulos, A. J. Enright, and C. A. Ouzounis. 2001. Textquest: Document clustering of medline abstracts for concept discovery in molecular biology. In *Pacific Symposium on Biocomputing*, pages 384–395.

I. T. Jolliffe. 1986. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag.

P. Kankar, S. Adak, A. Sarkar, K. Murari, and G. Sharma. 2002. MedMeSH summarizer: Text mining for gene clusters. In *Proceedings of the Second SIAM International Conference on Data Mining*, pages 548–565.

L. Kaufman and P. Rousseeuw. 1990. *Finding Groups in Data*. Wiley-Interscience.

R. Korfhage. 1997. *Information Storage and Retrieval*. Wiley Computer Publishers, New York.

Andrew Kachites McCallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering.

National Library of Medicine. 2003. Medical subject headings, MeSH. URL: `http://www.nlm.nih.gov/mesh/`.

OMIM. 2003. Online mendelian inheritance in man, OMIM^{TM}. MIM Number: 137580:11/10/2003: URL: `http://www.ncbi.nlm.nih.gov/omim/`.

Jörg Ontrup, Tim Nattkemper, Olaf Gerstung, and Helge Ritter. 2003. A MeSH term based distance measure for document retrieval and labeling assistance. In *Proceedings of the 25th Annual Int. Conf. of the IEEE Eng. in Medicine and Biology Society (EMBS)*, Cancun, Mexico.

C. Perez-Iratxeta, P. Bork, and M. A. Andrade. 2002. Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, 31(3):316–319.

R Development Core Team, 2003. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.

A. Renner and A. Aszódi. 2000. High-throughput functional annocation of novel gene products using document clustering. In *Proceedings of the Fifth Pacific Symposium on Biocomputing*, pages 54–65.

Miguel E. Ruiz and Padmini Srinivasan. 2002. Hierarchical text categorization using neural networks. In *Information Retrieval*, volume 1, pages 87–118.

Miguel E. Ruiz and Padmini Srinivasan. 2003. Hybrid hierarchical classifiers for categorization of medical documents. In *Proceedings of the 2003 Conference of ASIST*, Long Beach, CA, October.

Padmini Srinivasan and Thomas Rindflesch. 2002. Exploring text mining from medline. In *Proceedings of the Annual Conference of the American Medical Informatics Association*, pages 722–726.

Padmini Srinivasan and Micah Wedemeyer. 2003. Mining concept profiles with the vector model or where on earth are diseases being studied. In *Proceedings of the Text Mining Workshop. Third SIAM International Conference on Data Mining*, San Francisco, CA, May.

Padmini Srinivasan. 2001. MeSHmap: A text mining tool for medline. In *Proceedings of the Annual Conference of the American Medical Informatics Association*, pages 642–646, March.

Padmini Srinivasan. to appear. Text mining: Generating hypotheses from medline. *JASIST*. Accepted September 2003.

The Gene Ontology Consortium. 2000. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25:25–29.

Simon Twigger, Jian Lu, Mary Shimoyama, Dan Chen, Dean Pasko, Hanping Long, Jessica Ginster, Chin-Fu Chen, Rajni Nigam, Anne Kwitek, Janan Eppig, Lois Maltais, Donna Maglott, Greg Schuler, Howard Jacob, and Peter J. Tonellato. 2002. Rat Genome Database (RGD): mapping disease onto the genome. *Nucleic Acids Res*, 30(1):125–128, Jan.

W. John Wilbur. 2002. A thematic analysis of the AIDS literature. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 386–397.