# Exploring Text Mining from MEDLINE

**Padmini Srinivasan[@!], Ph.D.** **Thomas Rindflesch[!], Ph.D.**
**[@]The University of Iowa, Iowa City, IA 52242**
**[!]National Library of Medicine, Bethesda, MD 20894**

**{padmini|tcr}@nlm.nih.gov**

*We present a text mining application that exploits the MeSH heading subheading combinations present in MEDLINE records. The process begins with a user specified pair of subheadings. Co-occurring concepts qualified by these subheadings are regarded as being conceptually related and thus extracted. A parallel process using SemRep, a linguistic tool, also extracts conceptually related concept pairs from the titles of MEDLINE records. The pairs extracted via MeSH and the pairs extracted via SemRep are compared to yield a high confidence subset. These pairs are then combined to project a summary view associated with the selected subheading pair. For each concept the "diversity" in the set of related concepts is assessed. We suggest that this summary and the diversity indicators will be useful a health care practitioner or researcher. We illustrate this application with the subheading pair "drug therapy" and "therapeutic use" which approximates the treatment relationship between Drugs and Diseases.*

## INTRODUCTION

The goal for information extraction systems is to extract nuggets of information from collections of texts [1-6]. The semi-structured nature and natural language format of texts offer particular challenges. The extracted information may be referential as for example the names of cellular locations or the names of drugs. The information may be attributive, for example: a list of terms representing each gene of a set of genes [5]. Or, it may be relational such as highly specific predicates depicting particular interactions between proteins.

Text mining extends the domain of information extraction systems although the distinction between them is often blurred. As in data mining the key goal is the discovery of new knowledge [7]. The emphasis is on the extraction of knowledge that is at least not explicitly present in the source being mined. A pioneering example of text mining is ARROWSMITH designed to identify connections between unrelated literatures [8]. More current examples include the effort to build gene networks from MEDLINE through co-occurrence data [3,4,9,10]. In this paper we present initial research exploring some of the text mining opportunities offered by the MEDLINE database. In particular we present a methodology that allows us to generate a summary view of a group of concepts. The unit of representation in this view is a concept pair. The inclusion of a concept pair in such a summary view indicates that there is at least one document in MEDLINE whose key focus is on the nature of the relationship between the two concepts. That is, the corresponding documents are about some aspect of the interaction between the two concepts.

We illustrate our text mining methodology by focusing on concept pairs where one concept is, broadly speaking, a substance that is being studied for its therapeutic value, generally referred to as a Drug. The second concept is a problem for which a drug therapy is being explored, generally referred to here as a Disease. But the same method may be applied to other combinations of concepts as well such as Diseases and Organisms. We postulate that such summaries will provide informative overviews to the health care practitioner and researcher. It will provide confirmation for known facts while also supporting the generation of new ideas and hypotheses. Our intent is to generate these summaries from the entire MEDLINE database. However if needed they may also be generated from a subset limited to some specialty. The advantage in a comprehensive summary is that it can depict interdisciplinary, conceptual connections. For example, the Drug - Disease summary we generate from the full database identifies the set of diseases against which a drug has been studied. This is irrespective of the type of disease since it is not limited to specializations such as neurological or circulatory diseases.

There are three significant steps in our methodology. (1) Specify the type of conceptual pair that is of interest by specifying a pair of MeSH subheadings. (2) Extract concept pairs from each MEDLINE record and (3) Combine the extracted pairs to form the summary view. We detail these next.

## METHODS

**Specification of Subheading Pair:** The MeSH web site* lists fewer than 100 subheadings. Together the MeSH headings and subheadings offer a powerful indexing tool for MEDLINE. For instance the MeSH concept Hypertension may appear in two documents with the following different subheadings: Hypertension/treatment and Hypertension/adverse effects. Although both documents are about hypertension they cover identifiably different aspects. Thus the MeSH subheadings offer powerful retrieval points. MEDLINE indexers use the asterisk symbol to identify MeSH headings and subheadings that represent a major emphasis in the document.

Our first step is to select a pair of MeSH subheadings. Concepts that co-occur in the MEDLINE records where each member of the pair has one of the specified subheadings are then potential candidates for extraction. For example, if a document is indexed by Colchicine/therapeutic use and Back Pain/drug therapy then the pair Colchicine - Back Pain is extracted.

In contrast to the body of work utilizing MeSH concepts [eg., 4,9,10] there is much less attention given to the MeSH subheadings. One important example is the research of Cimino et al., [11,12]. In their study, subsets of documents on cardiovascular diseases corresponding to searches on therapy, diagnosis, etiology, and prognosis are examined. In each subset they examine the co-occurring pairs of MeSH subheadings to determine if these are statistically significant. The intent is to derive meaningful units of information from them. In previous work we presented MeSHmap, a prototype system that displays for the user the distribution of MeSH concepts in the retrieved document set. The concepts are distributed according to their subheading categories [13]. This paper extends our work with the prototype by utilizing SemRep, a linguistic tool, and by generating summary views. We anticipate that the user will specify the subheading combination to be explored but with guidance from the system. In parallel research we are exploring criteria for designing guidelines for this step.

**Extractions via MeSH:** For each document we extract all MeSH concepts that have been qualified by one of the two selected subheadings and marked as major. We then derive all pairs of these concepts such that they have different qualifiers.

*http://www.nlm.nih.gov/mesh/topcat.html

For the Drug - Disease example, we have selected the subheadings "drug therapy" and "therapeutic use". This pair is selected as an approximation to the treatment or therapy relation that potentially connects drug concepts with disease concepts. However, given that this is only an approximation, we do not claim that there definitely exists a treatment relationship between an extracted drug - disease concept pair. We only claim that the source document may be about some aspect regarding the interaction between the drug and the disease. Also that this aspect is in some way represented by the combination of the particular subheadings. Therefore a different pair of subheadings such as "drug therapy" and "adverse effects" would yield a different set of concept pairs and also differ in the underlying semantics.

**Extractions via SemRep:** SemRep is a natural language processing application designed to identify semantic relationships asserted in biomedical text [6]. For example, from the text in (1), SemRep identifies the relationship in (2).

(1) Methotrexate therapy in systemic lupus erythematosus

(2) methotrexate TREATS lupus erythematosus, systemic

The program relies on an underspecified syntactic analysis [14] to identify simple noun phrases in the text being processed: "Methotrexate therapy" and "systemic lupus erythematosus" in the example. Such phrases are then mapped [18] to concepts in the UMLS Metathesaurus, thereby determining that the concept "methotrexate" has been assigned the semantic type (or category) 'Pharmacologic Substance', while "systemic lupus erythematosus" has semantic type 'Disease or Syndrome' in the Metathesaurus.

On the basis of the syntactic analysis and the semantic type information, a set of argument identification rules refer to the UMLS Semantic Network and determine that the syntactic structure in (3) matches the Semantic Network relationship in (4).

(3) [Pharmacologic Substance] [ "in" [Disease or Syndrome]]

(4) Pharmacologic Substance TREATS Disease or Syndrome

When the corresponding Metathesaurus concepts are substituted for the semantic types in the Semantic Network Relationships, the result is the semantic interpretation given in (2). We apply SemRep to the titles of selected documents and

the arguments of the resulting relationships are the extracted concept pairs.

**High Precision/Confidence Filters:** We would like to be cautious about the extracted pairs of concepts. Thus given that our focus is on an approximation of the treatment relationship between drug and disease concepts, we limit our analysis to the subset of MEDLINE that corresponds to a query for documents on "therapy". We use the Haynes et al., filter criteria designed for high specificity results. This and other search filters designed by them [15] are available through the PubMed site[†]. Generally this filter extracts clinical studies conducting controlled experiments. It is not domain specific. Instead it targets the quality of the underlying experiments. Filters are optional in our methodology. Filters such as for species-specific publications may also be of interest.

**Combining MeSH based and SemRep based evidence:** Thus far we have described two independent processes for extracting pairs of objects. We now jointly assess the outputs of the two methods in order to further raise the integrity of the extracted pairs.

The MeSH based and the SemRep based approaches differ significantly in their operations. The former is based on co-occurrences while the latter utilizes linguistic criteria. The MeSH terms are assigned by human indexers while SemRep is an automated tool that we apply to the titles. Given that they are very different algorithms, we expect the errors made by one to be independent of the errors made by the other. Thus if a concept pair is extracted from a document using both methods then we are more confident about it than if it were extracted by either method alone. Thus for each document, we compare the MeSH based and the SemRep based pairs looking for matches and retain only matched pairs.

More specifically, each concept of a SemRep pair is compared against each concept of a MeSH pair. If both SemRep concepts match (match criteria described later) MeSH concepts then this results in the SemRep pair and the MeSH pair being added to the pool of extracted pairs. (If they are identical then only one instance is added.) A further constraint is that the subheadings of the matched MeSH concepts must be different. Thus for example they cannot both be qualified by "drug therapy". The criteria for matching is that either the two concepts are lexically identical or they are conceptually related to each other as determined using the UMLS Metathesaurus [‡] which has a total of $9,599,838$ conceptual relations in the 2001 version. Relations such as parent, child, sibling are included in this set, but we do not distinguish between them. If the matched concepts are different from each other, we select both the SemRep and the MeSH pairs. For instance two concepts will match if the UMLS identifies one as the parent of the other. In this case both pairs will appear in the output.

**Examples:** Consider the documents in Table 1. For the first title, "hay fever" is found to be related to Allergic rhinitis. Unfortunately, no relationship is observed between "Betamethasone valerate" and "betamethasone 17-valerate". Therefore no pair is extracted from this document. In the second document, we have the appropriate matches and thus two pairs are extracted that are shown in the table.

**Results:** Of the more than 11 million records in the MEDLINE database $15,254$ records satisfied the Hayne's et al., filter and also had at least one co-occurring pair of MeSH concepts with one concept qualified by therapeutic use* and the other by drug therapy*. Out of these, SemRep produced non-null output for $12,288$ record titles. SemRep extracted a total of $25,570$ pairs for these $12,288$ records which is on average two pairs per document. After the matching process was applied (described previously), a total of $7,332$ unique concept pairs were extracted. These contributed to a total of $12,845$ instances of pairs.

## SUMMARIZATION OF PAIRS FOR TEXT MINING

We now combine the individual record extractions to form a summary view. Table 2 shows the frequency distribution of the extracted pairs while table 3 shows some of the most frequent pairs. Both tables also show the number of documents from which the pairs were extracted (N). The last row of Table 2 shows for example that three concept pairs (0.05%) occurred in 16 documents in the dataset. As expected most (97%) concept pairs occur in only 1 MEDLINE record.

Table 4 illustrates summary information that may be useful to the health care practitioner as well as the researcher. It displays the set of diseases studied in relation to three different drugs. For each disease, the row shows the number of documents from which that pair was extracted.

[†]http://www.ncbi.nlm.nih.gov:80/entrez

[‡]http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

| |
|---|
| Title: Intranasal betamethasone valerate in seasonal rhinitis<br>MeSH: hay fever/drug therapy*<br>betamethasone 17-valerate/therapeutic use*<br>SemRep<br>Betamethasone valerate-Allergic rhinitis<br>Match<br>Allergic rhinitis: hay fever<br>Betamethasone valerate: NONE<br>Extracted Pairs: None |
| Title<br>Effects of prazosin in patients with hypertension<br>MeSH: hypertension/drug therapy*<br>antihypertensive agents/therapeutic use*<br>quinazolines/therapeutic use*<br>SemRep: Prazosin-Hypertension<br>Match<br>Prazosin: antihypertensive agents<br>Hypertension: hypertension<br>Extracted Pairs<br>antihypertensive agents AND hypertension<br>Prazosin AND Hypertension |

Table 1: Examples to Illustrate Procedure

Thus 30% of the documents about pyridazines are about congestive heart failure whereas only 10% are related to memory disorders. Additionally, we may assess the "diversity index" of a drug. At the intuitive level the greater the variety in the disease context within which a drug has been studied, the greater the diversity index of the drug. We formalize this notion as follows. Let $X = \{C1, C2, ..., Cn\}$ be the set of concepts (eg. the set of diseases) associated with the concept $Cx$ (eg. drug A) for which we are computing the diversity index. Compare each element of $X$ with all the other elements in $X$ to determine if there is a UMLS based relationship between them. If there is no relationship then this event contributes to the diversity index. More formally, for a concept $Cx$, $Diversity - Index(Cx) =$

$$[1 - \frac{\sum_{i=1}^{n} \sum_{k=1,k\neq i}^{n} related(Ci,Ck)}{n*(n-1)}]$$

where $related(Ci, Ck)$ returns a 1 if concepts $Ci$ and $Ck$ are related and otherwise a 0. Table 4 shows the calculated diversity index (score) for three drugs. We can conclude that pyrithioxin (score = 0.17) has been studied in the context of a more homogeneous set of health problems than pyridazines (score = 1.0) which spans very different health problems from the common cold to memory disorders. We suggest that drugs exhibiting greater diversity offer more points of connection between seemingly disparate problems. Informal observations from a physician at Lister Hill indicate that the Drug - Disease summary represents meaningful information. Our next step is to conduct formal evaluations.

| N | Num Pairs | % | N | Num Pairs | % |
|---|---|---|---|---|---|
| 1 | 5,464 | 97 | 17 | 5 | 0.08 |
| 2 | 940 | 17 | 18 | 8 | 0.1 |
| 3 | 362 | 7 | 19 | 1 | * |
| 4 | 164 | 3 | 20 | 2 | * |
| 5 | 111 | 2.3 | 22 | 2 | * |
| 6 | 64 | 1 | 23 | 3 | * |
| 7 | 54 | 1 | 24 | 1 | * |
| 8 | 35 | 0.6 | 25 | 1 | * |
| 9 | 23 | 0.4 | 26 | 2 | * |
| 10 | 17 | 0.3 | 28 | 1 | * |
| 11 | 10 | 0.2 | 29 | 1 | * |
| 12 | 16 | 0.3 | 31 | 2 | * |
| 13 | 12 | 0.2 | 34 | 2 | * |
| 14 | 6 | 0.1 | 41 | 2 | * |
| 15 | 6 | 0.1 | 43 | 1 | * |
| 16 | 3 | 0.05 | 66 | 1 | * |

Table 2: Frequency Distribution for Extracted Pairs.*: occurs in $< 0.05\%$ of documents

| Disease Concept | Drug Concept | N |
|---|---|---|
| hypertension | antihypertensive agents | 66 |
| angina pectoris | nifedipine | 43 |
| angina pectoris | calcium channel blockers | 41 |
| hypertension | atenolol | 39 |
| hypertension | propanolamines | 34 |
| depressive disorder | antidepressive agents | 34 |
| hypertension | hydrochlorothiazide | 31 |
| angina pectoris | diltiazem | 31 |
| asthma | bronchodilator agents | 29 |
| hypertension | calcium channel blockers | 28 |
| angina pectoris | atenolol | 27 |

Table 3: Sample of Most Frequent Pairs.

## CONCLUSIONS

This paper presents a system that has been developed for the extraction of pairs of concepts from the MEDLINE dataset. A pair is extracted if there is at least 1 document in the database that is on the nature of the relationship between the member concepts. As an example we consider pairs where one concept represents a drug and the other represents a problem or disease. We use several techniques to raise the integrity of the extracted pairs. We demonstrate text mining options that build upon the extracted pairs. In particular the overview generated allows one to assess each drug (or disease) in terms of the various disease (drug) contexts in which it has been studied.

| Drug with List of Diseases |
|---|
| pyrithioxin, Score = 0.17<br>alzheimer disease, 40%<br>cerebrovascular disorders, 20%<br>dementia, multi-infarct, 20%<br>dementia, 20% |
| triprolidine, Score = 0.5<br>hay fever, 41%<br>allergic rhinitis, nos, 16%<br>urticaria, 16%<br>otitis media, 16%<br>rhinitis, 11% |
| pyridazines, Score = 1.0<br>heart failure, congestive, 30%<br>depressive disorder, 30%<br>hypertension, 20%<br>common cold, 10%<br>memory disorders, 10% |

Table 4: Diversity Index Score for 3 Drugs.

Such assessments very naturally enable comparisons between drugs (or diseases) in terms of their diversity.

One limitation in this work is that we need to refine the output of the summarization step since the UMLS tends to be incomplete. For example, allergic rhinitis and rhinitis (see Table 4) could be combined. Also, entries that are at different levels of generality confound the output, as for example the entry for antibiotics versus the entries for the individual members of this group. We notice a tendency for more general concepts to have higher diversity scores which tells us that comparisons are best made at a given level of generality/specificity. We plan to explore this aspect by considering the depth of the classification tree in which the concept is located. These and testing of the summaries are planned for future research.

**Acknowledgments:** This research was accomplished while the first author was on a research visit to the Lister Hill Center from the University of Iowa. The first author acknowledges the support offered by the University of Iowa Faculty Scholar Award.

## References

1. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 2001;1(1);1:9.

2. Rindflesch TC, Rajan JV, Hunter L. Extracting molecular binding relationships from biomedical text. Appl. Nat. Lang. Process., 2000:188-95.

3. Shatkey H., Edwards S, Wilbur WJ, Boguski M. Genes, themes and microarrays. Using information retrieval for large-scale gene analysis. Proc. ISMB, 2000;8;317-328.

4. Stapley and Benoit. Bibiometrics: Information retrieval and visualization from co-occurrences of gene names in MEDLINE abstracts. Proc. PSB, 2000;5;526-537.

5. Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics, in press.

6. Rindflesch TC, Bean CA, Sneiderman CA. Argument identification for arterial branching predications asserted in cardiac catheterization reports. Proc. AMIA Annual Symposium, 2000, 704-8.

7. Hearst MA. Untangling text data Mining. Proc. ACL Conf., 1999.

8. Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimuls to scientific discovery. Artificial Intelligence, 1997;91;183-203.

9. Jenssen TK, Laegreid A, Komorowski J, Hovig E. Literature network of human genes for high-throughput analysis of gene expression. Nature Genetics 2001;28(1):21-8.

10. Masys DR, Welsh JB, Fink JL, Gribskov M. Klacansky I, Corbeil J. Use of keyword hierarchies to intepret gene expression patterns. Bioinformatics 2001;17(4);319:326.

11. Cimino JJ, and Barnett GO. Automatic knowledge acquisition from MEDLINE. Methods of Information in Medicine, 1993;32(2);120-130.

12. Mendonaa EA, Cimino JJ. Automated knowledge extraction from MEDLINE citations. Proc. AMIA Symp., 20 Suppl, 2000;575:579.

13. Srinivasan P. MeSHmap: A text mining tool for MEDLINE. Proc. AMIA Symp., 2001;642:646.

14. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc. AMIA Symposium, 2001;17:21.

15. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. JAMIA 1994 Nov-Dec;1(6):447-58.