LitLinker: A System for Searching Potential Discoveries in Biomedical Literature

Meliha Yetisgen-Yildiz
Information School
University of Washington
Seattle, WA, USA
melihay@u.washington.edu

ABSTRACT

The explosive growth in biomedical literature has made it difficult for researchers to keep up with advancements, even in their own narrow specializations. While researchers formulate new hypotheses to test, it is very important for them to identify connections to their work from other parts of the literature. However, the current volume of information has become a great barrier for this task, and new automated tools are needed to help researchers identify new knowledge that bridges gaps across distinct sections of literature. In this dissertation, we propose a knowledge discovery system called LitLinker that incorporates knowledge based methodologies, statistical methods and information extraction algorithms to mine the biomedical literature for new, potentially causal connections between biomedical terms. In this paper, we will discuss the main design challenges with our proposed solutions and the evaluation plan we designed to measure the overall performance of LitLinker.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Medical Information Systems.
H.3.3 [**Information Search and Retrieval**]
H.2.8 [**Database Applications**]: Data mining.

General Terms

Algorithms.

Keywords

Text-data mining, evaluation, knowledge-based systems.

1. INTRODUCTION

Information overload has become a significant problem for biomedical researchers. Scientific literature is readily available, but the sheer volume and the growth rate of the literature make it impossible for researchers to keep up with new findings outside their own narrowing fields of expertise. MEDLINE, the primary bibliographic database for biomedicine contains approximately 15 million references to journal articles and 3500 new references are added each day [8]. The most widely used public MEDLINE search interface, PubMed, only provides Boolean search with limited sorting options (publication date, first author name, journal name). Suppose a medical researcher wants to learn about the history of advancements in *Alzheimer's Disease* and searches MEDLINE through PubMed. For the query *Alzheimer's Disease*, PubMed returns more than 47,000 documents sorted with publication date (default option). Even if the researcher decides to review only the documents published in the last 12 months, she needs to go over more that 4,700 documents. Considering the amount of information available in MEDLINE, obviously no one is able to read about advancements across this entire body of literature.

SIGIR'06, August 6–11, 2006, Seattle, WA, USA.

The main goal of this dissertation is to propose a new knowledge retrieval approach to help medical researchers capture and explore new knowledge in the literature. Our system, LitLinker, uses literature based discovery to find new connections between biomedical terms that could lead to new directions in research. Our approach incorporates knowledge based methodologies, statistical methods and information extraction algorithms to mine biomedical literature for new, potentially causal links between biomedical terms.

2. RELATED WORK

Literature based discovery can be described as a way to examine a collection of documents and discover information not contained in any individual document in the collection. Literature based discovery research has been initiated by Swanson in the late 1980s. He proposed that combining existing, though not connected, bibliographic knowledge results in new knowledge. In an early example, he identified a hidden connection between the disjoint literatures on *Raynaud's Disease* and *Fish Oil* [15]. He noticed this connection by identifying three linking medical terms (*blood viscosity*, *platelet aggregation* and *vascular reactivity*) that occurred frequently in the titles of both the *Raynaud's Disease* literature and *fish oil* literature. The key to his approach was to assume that one level of transitivity held between correlated terms. In other words, the assumption is that if *Raynaud's Disease* is correlated with *blood viscosity*, and *blood viscosity* is correlated with *fish oil* than *Raynaud's Disease* is correlated with *fish oil*. Swanson's work introduced seminal ideas for literature based discovery; however, a limiting factor for his approach was the large amount of manual intervention required. Although his more recent work with Smalheiser incorporates an interactive tool called Arrowsmith [16], much work still is required to setup a customized list of stopwords and to sort through many spurious connections that Arrowsmith generates.

Many other researchers replicated Swanson's approach of taking advantage of an intermediate linking literature. As one example, Lindsay and Gordon [6] developed a process that followed the same basic architecture with Arrowsmith, but they added a variety of techniques to weigh terms using information retrieval methods

such as TF-IDF. They evaluated the performance, in terms of *precision* and *recall*, for generating the linking terms, where Swanson's identified linking terms served as the gold standard. Gordon and Dumais explored alternative techniques for identifying the linking literature [2] by using latent semantic indexing to extract close terms that occur in overlapping sets of documents. They replicated Swanson's *Raynaud's Disease* and f*ish oil* example to compare the performance of latent semantic indexing with the performance of term frequency and inverse document frequency methods used by Lindsay and Gordon.

Weeber et. al. also based their work on Swanson's approach [18]. They added both a natural language processing component to identify biomedical terms and a knowledge-based approach to help prune spurious connections based on the semantic type of the connection term. Although their system is more automated than the prior ones, it still requires a significant manual component for pruning the possible connections.

In contrast to other approaches, Hristovski et. al. applied association rule mining to find correlated terms in Swanson's open discovery approach and developed a system called BITOLA [3]. The main difference between their system and the prior ones is that they used Medical Subject Headings (MeSH), keywords assigned to the document, to capture the content of the documents instead of applying natural language processing techniques. For selecting the rules with correlated terms, they used an association rule metric, *support*, based on MeSH term co-occurrences.

Most recently, Srinivasan and Hu et. al. have worked on literature-based hypothesis generation using Swanson's approach. Srinivasan developed a new text mining system called Manjal [14]. As in Weeber et. al.'s system, she used a knowledge base for filtering terms according to their semantic types and like Lindsay and Gordon's approach, she used TF-IDF weights instead of simple term frequencies in determining the correlations among the MeSH terms. Hu et. al. focused on automatic selection of the semantic types for pruning the linking and target terms and developed a prototype system called Bio-SbKDS [4]. They used plain MeSH term co-occurrences to identify the term correlations and used UMLS Semantic Network to automatically select the semantic types to filter the correlated terms. Both Srinivasan and Hu et. al. evaluated their systems by replicating Swanson's discoveries.

3. LITLINKER

LitLinker was designed with what Swanson calls an open discovery approach. A high level view of the system architecture is represented in Figure 1. Our literature based discovery begins with a starting term (e.g., *Raynaud's Disease*), the term the researcher interested in investigating. Next, LitLinker uses a text mining process (Figure 2) to find a set of terms that are directly correlated with the starting term. We refer to this first set of correlated terms as the **linking terms** (e.g., *blood viscosity*). For each of the linking terms, LitLinker uses the same text mining process to identify a set of terms that are correlated with each linking term. We call these final terms **target terms** (e.g., *Fish Oil*). Finally, LitLinker ranks the target terns by the number of linking terms and the average correlation strength that connect the target term to the starting term. In the following sections, we will describe the main design challenges and present our proposed solutions.

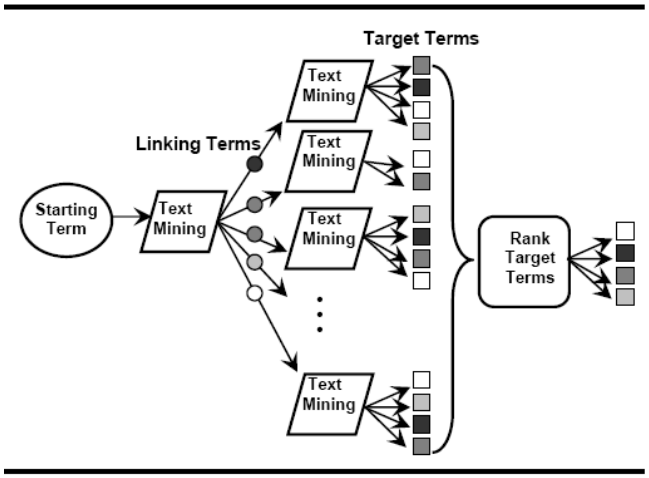

Figure 1. The discovery process in LitLinker.

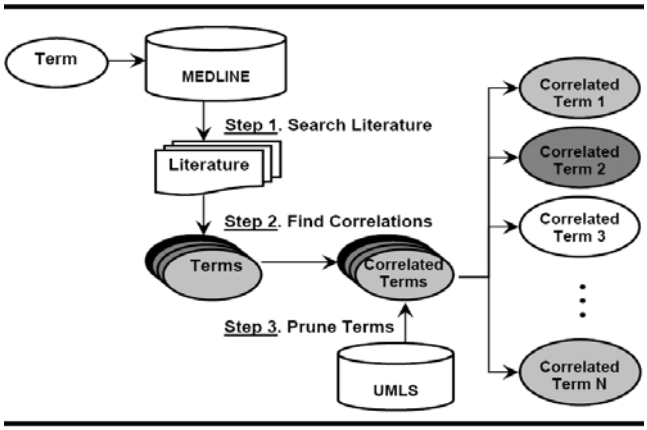

Figure 2. The text mining process in LitLinker.

3.1 Representing and Searching the Literature

We define the term **literature** as the collection of documents that include a given medical term (e.g., starting/linking term) in its representation. For the literature search, we created our own local MEDLINE database with the data leased from the National Library of Medicine (NLM). LitLinker searches this local database for collecting the literatures of a given starting or linking term.

In the context of literature based discovery, text representation can be defined as transforming text data into a representation that is suitable for literature retrieval and correlation mining algorithms to operate on. The two most common text representation approaches used by the current literature based discovery tools are *word representation* and *index term representation*. The word representation approach used in Arrowsmith [16] has two challenges. First, the percentage of long phrases, such as *calcium channel blockers*, in medical vocabulary is very high. This high prevalence of phrases represents a problem for text mining. Second, synonymy is a very common characteristic among the medical phrases.

As an alternative to word representation, many researchers preferred to use the index representations in their systems

[3,4,14,17]. MeSH is NLM's controlled vocabulary used for indexing MEDLINE articles. MEDLINE documents are manually categorized under 22,568 MeSH terms by the domain experts [7]. By representing documents with the MeSH terms, the researchers assume that the MeSH terms represent precisely what a document is about. But each MEDLINE document is indexed on the average with 12 descriptive MeSH terms. These tags can only capture just a few of the main important topics of the document and certainly miss most of the important entities mentioned inside the document.

To solve the problems with word and index representations, we have designed a knowledge-based, natural language processing approach to represent the medical text. A key part of our approach is to use a knowledge base to help identify domain-specific terms. The biomedical domain already has a large, publicly available knowledge base called the Unified Medical Language System (UMLS) [9]. NLM created this knowledge base by unifying hundreds of other medical knowledge bases and vocabularies to create an extensive resource that provides synonymy links as well as parent-child relationships among the medical concepts. We use MetaMap, a tool created by NLM that maps free text to biomedical concepts from the UMLS. In previous work, we compared MetaMap's performance against that of medical experts and reported good overall concept identification performance for MetaMap [11]. We divide each MEDLINE document into sentences, process the sentences with MetaMap, combine synonymous concepts by using UMLS unique identifiers, and store those identifiers in a database along with document and sentence identifiers. We keep sentence identifiers to be able to search within sentences.

In previous work, we both used index representation [17] and medical phrase representation (only titles) [10] in our experiments. We have not completed to process all MEDLINE abstracts yet. After we represent all MEDLINE abstracts with medical phrases, we plan to make a comparison between these two approaches and report our results.

3.2 Finding Interesting Correlations

A common approach applied to solve the problem of finding interesting correlations is to calculate the term frequencies in the literature of a starting or linking term and pick the terms with high frequencies as the correlated ones [3,4]. The main problem with term frequency approach is that term frequencies indicate strong but not necessarily interesting correlations. For example, the correlation between two medical terms *migraine* and *pain* is a strong one, because the two terms appears together in 3,085 documents in MEDLINE[1]. However, it is not a very interesting connection because *pain* is a fairly generic term which appeared commonly with many other terms besides just *migraine*. In contrast, *spreading cortical depression* appears with *migraine* in 196 documents of the *migraine* literature. Although the correlation between *migraine* and *spreading cortical depression* is weaker than the correlation between *migraine* and *pain*, intuitively it is a more interesting correlation.

To overcome this problem, some researchers have used TF-IDF instead of plain term frequencies while identifying correlated terms [6,14]. Assuming the more frequent terms tend to be more general terms, IDF gives the less frequent terms an advantage in TF-IDF calculation. But for some cases, it is possible that the TF part dominates the IDF part and uninteresting terms are selected as correlated terms.

To address the problems with the term frequency approach and TF-IDF, we focused on term probabilities while designing LitLinker. We calculated the probability of a term appearing in a literature by dividing the number of documents of the literature in which the term appeared by the total number documents in the literature. Using this approach, we could add the literature sizes into the process of finding correlations. From the term probabilities, we observed that the terms in both interesting and not interesting correlations had consistent probability distribution patterns. As an example, the probabilities of *pain* in other literatures were quite similar to its probability in *migraine* literature. In contrast, the probability of *spreading cortical depression* in *migraine* literature was much higher than its probabilities in the other literatures. Starting from this observation, we designed a statistical approach, called Z-Score, based on the background probability distribution of terms in the MEDLINE database to find interesting correlations [17].

Statistical co-occurrence approaches have been widely used to identify protein-protein interactions in medical text. Mutual information measures [1] and hypergeometric distribution [12] are among the most popular ones. We currently work on a paper that compares our Z-Score approach with the mutual information measures and hypergeometric distribution in the context of literature based discovery.

3.3 Pruning Correlations

Even with our statistical correlation mining algorithm, the number of correlated terms identified in a literature is usually very high. But only a subset of those terms would make medically plausible correlations. While investigating the possible ways to prune non-interesting terms, we found three classes of problems: (1) some terms were too broad (e.g., *medicine*, *disease*, *human*) to be linking or target terms; (2) some terms were too closely related to the starting term (e.g., *common migraine* for the starting term *migraine*) to be linking or target terms; and (3) some terms just did not make sense as plausible connections for the purposes of the discovery.

We use the concept hierarchies that are available in the UMLS to solve the first two classes of problems [17]. For the third class of problem, we use UMLS Semantic Network. In the UMLS, each term is connected through an *isa* link to one or more semantic types from a set of 135 general medical terms that the NLM calls the Semantic Network [9]. For example, the term *migraine* has a semantic type of *disease and syndrome*, and the term *magnesium* has a semantic type of *biologically active substance*. LitLinker takes two sets of semantic types as input: one for linking term selection and one for target term selection. It prunes any terms that does not match the corresponding semantic type criteria.

3.4 Ranking Target Terms

Ranking target terms from all the linking terms requires multiple processing steps. First, LitLinker merges the lists of correlated terms from each of the linking terms. Second, because we are

[1] PubMed search was conducted on March 7th , 2006.

only interested in novel connections, LitLinker must prune previously known connections from the list of target terms. It marks any co-occurrence with the starting term as a known connection and eliminates it from the list of target terms. LitLinker ranks the final list of target terms according to the number of linking terms that connect that target term the original starting term. Some target terms may have the same number of linking terms. LitLinker ranks those target terms according to their average Z-Score link weight. Such a ranked list should provide enough information to help researchers evaluate and explore these possible correlations and determine the ones which seems worthy for further investigation.

4. Finding the Meaning of Correlations

LitLinker generates a list of ranked terms indirectly correlated to the starting term through the linking terms as its final output. But it does not tell anything about the meaning of direct correlations that connect a starting term to the linking terms or the linking terms to the target terms.

As described in Section 3.2, direct correlations are identified via term co-occurrence statistics. Medical terms co-occur together for many different reasons. For example, possible explanations of a correlation between a symptom, *S*, and a chemical or drug, *D*, can be listed as; (1) *D* is used to cure *S*, (2) a side affect of *D* causes *S*, or (3) D was believed to cure/cause S and this information was proven to be wrong. As part of this dissertation, we plan to extract such knowledge to explain the meaning of the direct correlations identified by LitLinker and present this knowledge as supporting information to researchers who want to explore why some target terms have been suggested by LitLinker as potential discoveries for their input starting terms.

There has been a great amount of work published on medical text mining for extracting protein interactions [5]. In the domain of literature based discovery, finding the meaning of correlations can be seen as a superset of extracting protein interactions because the correlated terms may be from various different classes including the proteins. To solve this problem, we plan to use UMLS Semantic Network as a guide. As mentioned before, each medical concept in UMLS is mapped to at least one semantic type. Semantic Network also provides us a directed graph where each node is a semantic type and each link is a relation between the nodes it connects. In this graph, there are 135 semantic types and 49 different types of relations. As an example, the semantic type *Antibiotic* is connected to the semantic type *Disease or Symptoms* through six different relations, *affects*, *causes*, *complicates*, *diagnoses*, *prevents*, and *treats*. As can be observed from the list, some of the relations are compliment of each other. To select which relations hold for a given correlation, we plan to use text classification techniques. To overcome the expensive manual labeling process while creating the training sets, we will use a semi automated approach. For each (semantic type1 – relation - semantic type2) triple extracted from the graph, we will identify a number of medical concept couples that the semantic types and the relation hold. A medical expert will guide us in the concept selection process. After this step we will query our MEDLINE database to extract sentences that include the selected concept couples and label these sentences as positive examples in our training set for the (semantic type#1 – relation - semantic type#2) triple. To predict the relations for a given correlation, we will extract the semantic types of the concepts and the relations between those semantic types to identify the corresponding trained classifiers. We will use these classifiers on the sentences that include both correlation concepts and decide which relations hold for the given correlation according to the output of the classifiers. We plan to present a set of supporting sentences to the users for summarization and validation purposes.

5. Evaluation Plan

Evaluating knowledge discovery systems is a fundamentally challenging task because if the systems are successful, by definition they are capturing new knowledge that has yet to be proven useful. Evaluation will be one of the major components of this dissertation. In the following sections, we will explain how we quantitatively evaluate the performance of LitLinker and discuss our plans of involving medical researchers in the user centered evaluation.

5.1 Quantitative Evaluation

Swanson and Smalheiser have made various discoveries by applying their literature based discovery method to MEDLINE and published their results in the medical domain. Their discoveries have become gold standards for evaluation. Many researchers have measured the performance of their discovery systems by replicating Swanson's discoveries and using the literatures published before the original discovery dates [2,4,6,10,14,18]. They have reported overall success if one of the correlations generated by their systems was same as Swanson's discovery. Although the researchers have successfully replicated Swanson's discoveries, this type of evaluation is not complete because it does not inform us about the quality of the rest of the correlations identified by the systems.

While evaluating LitLinker, we use an alternative evaluation approach that enables us to evaluate all correlations LitLinker generates. In our evaluation, for a given starting term, we measure whether LitLinker leads us to new discoveries in the more recently published literature. To accomplish this goal, we select a cutoff date and divide the MEDLINE database into two sets. The first set includes the documents published before the selected date and the second set includes the rest. We run LitLinker on the first set of documents, and check the existence of the correlations generated by LitLinker in the second set. We use information retrieval techniques, precision and recall to quantify the overall performance. In [17], we have reported performance results of LitLinker with a cutoff date, 1/1/2004.

In addition to the evaluation method described above, we plan to use another method that adds the medical researchers' judgment of *usefulness* into the evaluation process. For this evaluation, we will closely work with a medical researcher. We will run LitLinker with a set of starting terms from researcher's interest area and we will ask her to label the identified correlations based on usefulness using a scale of *not at all*, *not really*, *neutral*, *mostly*, and *definitely*. We will use the frequencies of the labels to calculate the overall performance of the system.

5.2 User Centered Evaluation

Although, all these literature discovery systems have been developed as potential solutions to information overload problem, none of them have been tested for their capabilities in coping with medical researchers' information overload problem in real life situations yet.

In order to gain more insight on how medical researchers might incorporate LitLinker into their work, we designed an interactive web interface that supports researchers' interactive exploration of the identified connections[2]. We tested this interface with a usability study and based on the results we further tailored the interface to support the work processes of medical researchers [13]. In our user centered evaluation, we plan to use this web interface and ask medical researchers to use LitLinker for their own research questions. By conducting observations and interviews, we aim to learn about their experience such as the things they like or do not like about LitLinker's discovery process and how the discovery process relates to their current information seeking behavior while they seek for new hypotheses to test.

6. CONCLUSION and ISSUES FOR DISCUSION

With the explosion of the scientific literature, text mining systems such as LitLinker will become critical for helping researchers make discoveries across distinct portions of the biomedical literature. The main contribution of this dissertation is our text mining architecture. In [17], we have shown that our combination of a statistical method based on word probability distributions and a knowledge-based approach can be incorporated into an effective system with examples of recently published discoveries. Another contribution of our research will be our information extraction approach to find the meaning of direct correlations. To our knowledge our work will be the first one that extensively explores the meaning of correlations between all kinds of medical terms being not limited to protein-protein or drug-disease interactions. The final contribution of our research will be the quantitative and user-centered evaluation approaches that we design to measure the performance of LitLinker. These approaches can easily be adopted to other existing literature based discovery systems and used for further comparisons.

LitLinker provides a new and effective type of knowledge retrieval approach. This new retrieval approach will help researchers to capture and explore new connections in the vast biomedical literature and guide them while identifying new research directions. At the doctoral consortium, we hope to discuss all the aspects of our text mining approach and take comments and recommendations from the consortium committee members and other doctoral students especially on our information extraction approach and evaluation plan.

7. ACKNOWLEDGMENTS

The National Science Foundation, award #IIS-0133973, funded this work. I appreciate the support of my advisor, Wanda Pratt, and my committee, David Hendry, Sherrilynne Fuller, and Oren Etzioni.

8. REFERENCES

[1] Alako, B. T. F., Veldhoven, A., van Baal, S., Jelier, R., Verhoeven, S., Rullmann, T., Polman, J., and Jenster, G. (2005). CoPub Mapper: mining MEDLINE based on search term co-publication. BMC Bioinformatics 6: 51.

[2] Gordon, M. D., and Dumais, S. (1998). Using latent semantic indexing for literature based discovery. JASIST 49(8): 674-685.

[3] Hristovski, D., Stare, J., Peterlin, B., and Dzeroski, S. (2001). Supporting discovery in medicine by association rule mining in Medline and UMLS. Proc. of Medinfo, pp.1344-8.

[4] Hu, X., Li, G., Yoo, I., Zhang, X., and Xu, X. (2005). A Semantic-based Approach for Mining Undiscovered Public Knowledge from Biomedical Knowledge. Proceedings of Proc. of IEEE-GrC'05, pp. 163-166, Beijing.

[5] Krallinger, M., Erhardt, R.A.-A., and Valencia, A. (2005). Text Mining Approaches in Molecular Biology and Biomedicine. Drug Discovery Today 10(6): 439-445.

[6] Lindsay, R. K., and Gordon, M.D. (1999). Literature-based discovery by lexical statistics. JASIST 50(7): 574-587.

[7] NLM. (2005). MeSH Fact Sheet. Available at: http://www.nlm.nih.gov/pubs/factsheets/mesh.html

[8] NLM. (2005). MEDLINE Fact Sheet. Available at: http://www.nlm.nih.gov/pubs/factsheets/medline.html

[9] NLM. (2005). UMLS Fact Sheet. Available at: http://www.nlm.nih.gov/pubs/factsheets/umls.html

[10] Pratt, W., and Yetisgen-Yildiz, M. (2003). LitLinker: Capturing Connections across the Biomedical Literature. Proc. of K-Cap'03, pp.105-112, Florida.

[11] Pratt, W., and Yetisgen-Yildiz, M. (2003). A Study of Biomedical Concept Identification: MetaMap vs. People. Proc. of the AMIA'03, pp.529-533, Washington D.C.

[12] Ramani, A., Bunescu, R., Mooney, R., and Marcotte, E., (2005). Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. Genome Biology 6(5): R50

[13] Skeels, M. M., Henning, K., Yetisgen-Yildiz, M., and Pratt, W. (2005). Interaction Design for Literature-Based Discovery. Proc. of CHI'05, pp.1785-1788, Portland.

[14] Srinivasan, P. (2004). Generating Hypotheses from MEDLINE. JASIST 55(5): 396-413.

[15] Swanson, D. R. (1986). Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge. Perspect. Biol. Med. **31**: 526-557.

[16] Swanson, D. R., and Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. AI 91: 183-203.

[17] Yetisgen-Yildiz, M., and Pratt, W. (2006). Using Statistical and Knowledge-Based Approaches for Literature Based Discovery. Journal of Biomedical Informatics (To appear).

[18] Weeber, M., Klein, H., and de Jong - van den Berg, L.T.W. (2001). Using Concepts in Literature Based Discovery: Simulating Swanson's Raynaud-Fish Oil and Migraine-Magnesium Examples. JASIST 52(7): 548-5.

[2] Available at: http://litlinker.ischool.washington.edu/index.jsp