Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 39 (2006) 600–611
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Using statistical and knowledge-based approaches for literature-based discovery

Meliha Yetisgen-Yildiz [a], Wanda Pratt [a,b,*]

[a] *Information School, University of Washington, Seattle, WA, USA*
[b] *Biomedical and Health Informatics, University of Washington, Seattle, WA, USA*

Received 17 June 2005
Available online 4 January 2006

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 the literature. In this paper, we present a literature-based discovery system called LitLinker that incorporates knowledge-based methodologies with a statistical method to mine the biomedical literature for new, potentially causal connections between biomedical terms. We demonstrate LitLinker's ability to capture novel and interesting connections between diseases and chemicals, drugs, genes, or molecular sequences from the published biomedical literature. We also evaluate LitLinker's performance by using the information retrieval metrics of precision and recall.

Keywords: Text mining; Literature-based discovery; Knowledge-based systems; Evaluation

1. Introduction

Information overload has become a significant problem for biomedical researchers. Scientific literature is readily available, but the sheer volume and growth rate of the literature makes it impossible for researchers to keep up with new findings outside their own narrowing fields of expertise. For example, MEDLINE, the primary bibliographic database for biomedicine, contains approximately 13 million references to journal articles and over 2000 new references are added each day [10]. Obviously, no one is able to read about advancements across this entire body of the literature. Tools are needed to help them capture and explore the new knowledge in the literature.

To address this need, we have developed a system, called LitLinker, which uses the literature-based discovery to find new connections between biomedical terms that could lead to new directions in research. Our approach incorporates knowledge-based methodologies and statistical methods to mine biomedical literature for new, potentially causal links between biomedical terms.

In this paper, we describe the architecture of LitLinker and report evaluation results. In our evaluation, we measured LitLinker's ability to capture novel and interesting connections between *diseases* and *chemicals*, *drugs*, *genes*, or *molecular sequences* from the biomedical literature. To accomplish this task, we ran LitLinker on the set of MEDLINE documents published before January 1, 2004 and compared its predictions against the new connections published between January 1, 2004 and September 30, 2005. We used the information retrieval metrics of precision and recall to evaluate the overall performance. By providing examples from the very recently published papers as

* Corresponding author. Fax: +1 206 616 3152.
E-mail addresses: melihay@u.washington.edu (M. Yetisgen-Yildiz), wpratt@u.washington.edu (W. Pratt).

1532-0464/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2005.11.010

discovery evidence, we also discuss in detail three new connections identified by LitLinker: *Alzheimer disease–endocannabinoids*, *migraine–AMPA receptors*, and *schizophrenia–secretin*.

2. Related work

Other researchers have been working in the general area of biomedical literature-based discovery for nearly fifteen years. Swanson initiated the term and was responsible for much of the earliest work in this area [19,21]. He used a combination of citation analysis and manual review in his discovery process. The former was used to determine novelty by detecting disjoint literatures. The latter was used to identify plausible new connections across disjoint biomedical literatures by examining the titles from search results. In an early example, Swanson identified a hidden connection between the disjoint literatures on *migraine* and *magnesium* [18]. He noticed this hidden connection by identifying several linking medical terms, such as *epilepsy* and *calcium channel blockers*, that occurred frequently in the titles of both the *magnesium* literature and *migraine* 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 *migraine* is correlated with *epilepsy*, and *epilepsy* is correlated with *magnesium*, then *migraine* is correlated with *magnesium*. 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 research with Smalheiser incorporates an interactive tool called Arrowsmith [20], much work still is required to setup customized lists of stop words and to sort through the many spurious connections that Arrowsmith generates.

Many other researchers replicated Swanson's approach of taking advantage of an intermediate linking literature, and we will refer to this class of work as **literature-based discovery** throughout this paper. As one example, Lindsay and Gordon [7] 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 term frequency and inverse document frequency. 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. In their more recent work, [4] attempted to show that literature-based discovery could be performed on the World Wide Web. They picked *genetic algorithm* as the starting term and used Swanson's open-ended-discovery approach to discover many potential fields of application for *genetic algorithms*. Gordon and Dumais also explored alternative techniques for identifying the linking literature [5] by using latent semantic indexing to extract close terms that occur in overlapping sets of documents. They replicated Swanson's *Raynaud's disease* and *fish 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. In previous work, we used a knowledge-based approach to identify and prune potential linking terms [3]. However, these researchers focused exclusively on evaluating their systems' ability to generate the desired linking terms, and none evaluated how easy it would be to identify the novel new target term (e.g., *magnesium*).

Weeber et al. also based their work on Swanson's approach [22]. 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. In their latest research, they used their system to investigate new potential uses for *thalidomide* with Swanson's open-discovery approach [23]. They executed the discovery process in the end of July 2000 and evaluated their hypotheses by analyzing the literature published after the execution date. Although, their system is more automated than the prior ones, it still requires a significant manual component for pruning the possible connections.

Wren applied mutual information measures to Swanson's literature-based discovery approach [24]. In his approach, if the probability of observing a term increases when another term is mentioned, then there is a correlation between these two terms. He used a joint mutual information method of ranking target terms based upon their shared associations. He evaluated his system with Swanson's *Raynaud's disease–fish oil* and *migraine–magnesium* examples.

Most recently, Srinivasan and Hristovski et al. have worked on literature-based hypothesis generation using Swanson's approach. Srinivasan developed a new text mining system called Manjal [16]. 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 term weights instead of simple term frequencies in determining the correlations among terms. The main difference between her system and the prior ones is that she used Medical Subject Headings (MeSH), keywords assigned to the document, to capture the content of the documents instead of applying natural language processing techniques. She also clustered linking term candidates under their semantic types, ranked the terms in each semantic type cluster by using an information retrieval metric based on term co-occurrences, and selected a predefined number of terms as connection terms from each semantic type cluster. Her system supports both open- and closed-discovery approaches. She has reported the results of many experiments for both discovery types, but none of the results include an overall ranking of the proposed discoveries.

In contrast to other approaches, Hristovski et al. applied association rule mining to find correlated MeSH terms in Swanson's open-discovery approach and developed a system called BITOLA [6]. For selecting the rules with correlated terms, they used an association rule metric,

support, based on MeSH term co-occurrences. They ran BITOLA on all the medical literature before the end of 2001 to extract disease-gene correlations, but they did not evaluate the correlations that BITOLA generated.

3. Approach

Like Manjal and BITOLA, our system LitLinker also uses MeSH terms to represent the documents. In the initial version of LitLinker, we used natural language processing methods to represent documents [13]. However, in our experiments, we found this method to be computationally too expensive for practical use and decided to use MeSH terms to represent documents.

One way that LitLinker differs from Manjal and BITOLA is in the approach that it uses to identify correlated terms. LitLinker applies a statistical approach that is based on the background distribution of term probabilities. LitLinker also extensively uses a medical knowledge base to prune away the uninteresting correlations.

LitLinker was designed with what Swanson calls an open-discovery approach. A high-level view of the process is illustrated in Fig. 1. Our literature-based discovery begins with a **starting term** (e.g., *migraine*), the term the researcher is interested in investigating. Next, LitLinker uses a text mining process 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** (shown as shaded circles—e.g., *epilepsy*, *calcium channel blockers*). For each of the linking terms, LitLinker then 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** (shown as shaded squares—e.g., *magnesium*). Finally, LitLinker ranks the target terms by the number of linking terms that connect the target term to the starting term. Thus, it provides an organized list of possibilities for this open-discovery process. In the figure, for both linking and target terms each color shade maps to a distinct term.

The text mining process is shown in Fig. 2. LitLinker uses a biomedical knowledge base, the Unified Medical Language System (UMLS), as an integral component throughout the text mining process [11]. This knowledge base was created by the National Library of Medicine (NLM) and contains over 975,000 biomedical concepts as well as 2.3 million concept names. The system was created 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 concepts. UMLS helps LitLinker to limit the search space by pruning away unhelpful terms. LitLinker's data mining component, as will be explained in Section 3.2, plays a key role in determining which terms are correlated with each other.

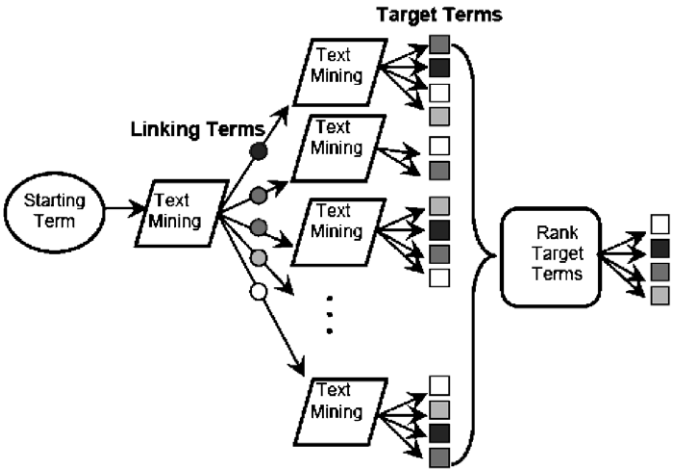

Fig. 1. The discovery process in LitLinker.

For the term that is provided, LitLinker identifies all documents in MEDLINE that contain that term and gathers all the MeSH terms used in that collection of documents. The terms are pruned and correlated terms are identified using the process described in Section 3.2.

In the following sections, we describe in detail each of the major steps of the text mining process.

3.1. Searching the literature

For the literature search, we created our own local MEDLINE database with the data leased from the NLM. LitLinker searches this local database for collecting the literatures. We constrained LitLinker's literature queries through two parameters. The first parameter is the Medical Subject Headings (MeSH) category names. MEDLINE documents are manually categorized under 22,568 MeSH category names by the experts from the NLM [9]. On the average, each document is categorized under 12 MeSH terms. LitLinker uses MeSH terms as the representation of the content of the documents and performs searches on them to collect the literatures.

The second parameter is the publication type. MEDLINE includes documents from various publication types, but we found that the documents with some of publication types (e.g., comment, biography, dictionary, and lectures) not very useful for our discovery process because they did not contain research results. We manually selected the publication types to exclude, and LitLinker eliminated documents with those publication types from its search space. Numerically, out of 12,421,396 documents available in the 2004 MEDLINE baseline 11,493,866 of them (95% coverage) are in LitLinker's search space.

In the rest of the paper, we use the term **literature** to define the set of documents categorized under a given MeSH term with valid publication types.

3.2. Finding correlations

A key part of our text mining approach is the process of identifying associated or correlated terms. This process produces both linking terms and target terms.

An easy solution to the problem of finding correlations would be to calculate the term frequencies in the literature

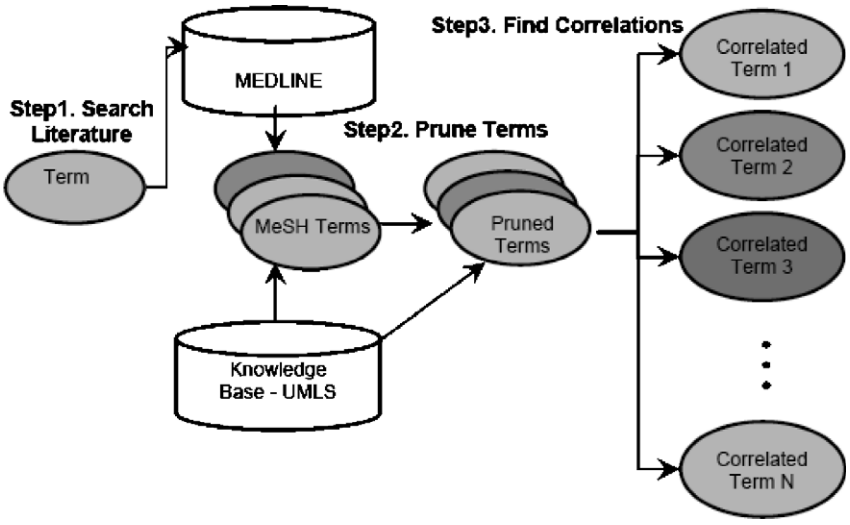

Fig. 2. The text mining process in LitLinker.

of a starting or linking term and pick the terms with high frequencies as the correlated ones. The main problem with this approach was that term frequencies indicate strong but not necessarily interesting correlations. For example, the correlation between *migraine* and *pain* was a strong one, because *pain* appeared in 301 documents of the *migraine* literature. However, it was not a very interesting connection because pain was a fairly generic term which appeared commonly with many other terms besides just migraine. In contrast, *spreading cortical depression* appeared in 92 documents of the migraine literature. Although, the correlation between migraine and spreading cortical depression was weaker than the correlation between migraine and pain, intuitively it was a more interesting correlation.

To address this problem, we focused on term probabilities rather than simple term frequencies. 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 probability distributions of the terms in interesting correlations were more diverse than those of the terms in not interesting correlations. As an example, the probabilities of *pain* in other literatures were quite similar to its probability in the *migraine* literature. *Pain* appeared in 11,926 literatures with a smooth probability distribution (StdDev = 0.002). In contrast, the probability of spreading cortical depression in *migraine* literature was much higher than its probabilities in the other literatures. *Spreading cortical depression* appeared in 1569 literatures with a more diverse probability distribution (StdDev = 0.03). Starting from this observation, we designed a statistical approach based on the background probability distribution of terms in the MEDLINE database to find interesting correlations.

As described in the previous sections, LitLinker represents the documents with MeSH terms. In the 2004 MEDLINE baseline, there are 22,568 MeSH terms, which means there are 22,568 different literatures. On the average, each document is represented with 12 MeSH terms. To calculate the term probability distribution in these literatures, we first defined the probability, P, of a MeSH term, m, appearing in a literature, l, as

$$P_l^m = \frac{F_l^m}{D_l}, \tag{1}$$

where F_l^m is the number of documents with the MeSH term m in the literature l and D_l the total number of documents in the literature l. By using this probability definition, we calculated the mean probability of the MeSH term m in the background literatures with the following formula:

$$\overline{P^m} = \frac{\sum_{l=1}^{N^m} P_l^m}{N^m}, \tag{2}$$

where N^m is the total number of literatures that contain MeSH term m.

The mean probability of a MeSH term provides a sense of whether the term is a highly frequent one in the entire MEDLINE literature, but it does not tell us whether the term is strongly associated with any particular literatures. The combination of the mean probability with the deviation of the term probability distribution is more indicative than only mean probability.

We used the standard deviation definition to calculate the deviation of term probability distribution in the background literatures. From (1) and (2) we calculated deviation of the term probability distribution for MeSH term m as

$$\sigma^m = \sqrt{\frac{1}{N^m - 1} \sum_{l=1}^{N^m} (P_l^m - \overline{P^m})^2}. \tag{3}$$

Linker calculates and stores the mean probability and the standard deviation of term distribution for each of the MeSH terms available in MEDLINE. To find which

MeSH terms are correlated to either the starting or the linking term, LitLinker first finds the list of terms that appears in the starting or linking term literature and prunes them as will be described in the following section. For the remaining terms, it calculates their z-score. The z-score of a MeSH term m in the starting or linking term literature l can be calculated from (1)–(3) as

$$z_l^m = \frac{P_l^m - \overline{P^m}}{\sigma^m}. \quad (4)$$

This score provides the distance between the probability of a MeSH term in a specific literature and the general distribution of this MeSH term in the background set of literatures. LitLinker marks the terms with z-scores larger than a predefined threshold as the correlated terms to the starting or linking term.

Although, the end goals are quite different, our statistical approach of finding correlated terms shows great similarities with Andrade and Valencia's work on automatic extraction of keywords from medical text [2]. They explored the possibility of extracting biologically significant words related to protein functions directly from MEDLINE abstracts with a similar statistical technique. However, they used a small set of terms composed of only 71 proteins to create their dataset; whereas, we applied our statistical technique to find correlations among 22,568 MeSH terms with a 95% MEDLINE coverage.

3.3. Pruning literature terms

The number of MeSH terms that appear in a literature is usually very high. For example, the literature retrieved by LitLinker for the starting term *migraine* was composed of 9919 documents. The total number of distinct MeSH terms that appeared in the migraine literature was 5095, but only a subset of those terms would make medically plausible linking terms.

While investigating the possible ways to prune non-interesting linking and target terms, we found three classes of problems: (1) some terms were too broad (e.g., medicine, disease, and human) to be target terms; (2) some terms were too closely related to the starting term to be linking terms (e.g., headache for the starting term migraine); and (3) some terms just did not make sense as plausible connections for the purposes of the discovery.

We used the MeSH hierarchy that is available in the UMLS to solve the first problem of pruning broad target terms. In the MeSH hierarchy, terms are ordered from general to specific. We assumed that if a specific term was a known connection then its more general ancestors were also known connections. LitLinker eliminates the target terms that are more general than one or more linking terms. The portion of the MeSH hierarchy for migraine is given in Fig. 3. Suppose migraine is a linking term. It can be observed from the figure that the scope of terms at the highest levels (e.g., diseases and nervous system diseases) of the migraine hierarchy is too broad to be helpful in the discovery process. By pruning the ancestors of linking terms LitLinker eliminates many of such broad target terms.

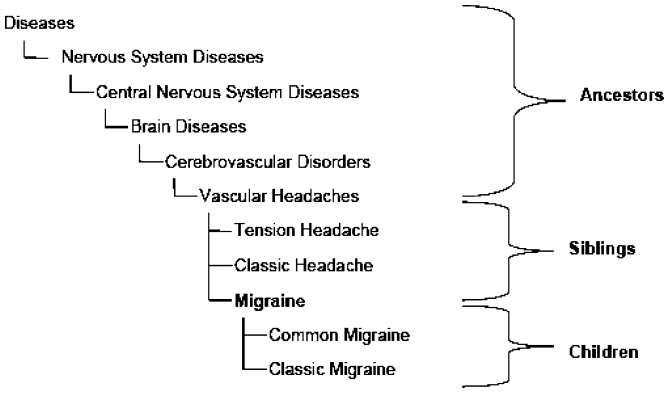

Fig. 3. MeSH hierarchy for migraine.

We again used the MeSH hierarchy to solve the second problem of pruning too closely related linking terms to the starting term. If migraine is a starting term, it can be observed from Fig. 3 that all terms in the migraine hierarchy especially the terms in the immediate family (e.g., grandparents, parents, siblings, and children) are very closely related to migraine. If selected as linking terms, such closely related terms would not lead to interesting and novel target terms for the starting term migraine. To eliminate closely related terms, LitLinker prunes all ancestors, siblings, and children of the starting term from the list of potential linking terms. For the migraine example, given in Fig. 3, LitLinker prunes all the ten terms from the list of potential linking terms.

For the third problem, the challenge was to generate an automated, generalizable approach to pruning away those implausible and uninteresting terms. We used UMLS to solve this problem. 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*. For example, the term *migraine* has a semantic type of *disease and syndrome*, and the term *magnesium* has two semantic types of *biologically active substance* and *element*, *ion*, or *isotope*. LitLinker takes two sets of semantic types as input: one for linking term selection and one for target term selection. It eliminates any terms that do not match the corresponding semantic type criteria.

Although, semantic types are very helpful in identifying the discovery domain, users of LitLinker need to select the types manually from a set of 135 semantic types for each discovery task. Recently, a group of researcher from NLM grouped semantic types under 15 more general headings that they call *semantic groups* [8]. Just as the semantic types provide a good summary for the biomedical terms, semantic groups provide a good summary for the semantic types. For example, one of the semantic types of *magne-*

Table 1
Semantic groups selected for our experiments

Linking term selection	Target term selection
Chemicals and drugs	Chemicals and drugs
Disorders	Genes and molecular sequence
Genes and molecular sequence	
Physiology	
Anatomy	

sium is *element*, *ion*, or *isotope* and the semantic group of *element*, *ion*, or *isotope* is *chemicals and drugs*. We used semantic groups as a guide to decrease the manual effort in selecting the semantic types for the experiments presented in this paper. Our goal was to find semantic types that were plausible for terms that could be correlated with a disease or a medical condition and a potential treatment. To accomplish this goal, we first selected the semantic groups listed in Table 1 and used all semantic types classified under these semantic groups in LitLinker's discovery process.

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. It also retains the linking terms that connect that target to the starting term. Second, because we are only interested in novel connections, LitLinker must prune previously known connections from the list of target terms. We decided that any co-occurrence with the starting term constituted a known connection. Thus, LitLinker checks each candidate target term against the entire set of MeSH terms that were extracted from the starting term literature. If a candidate target term is an element of this set, LitLinker eliminates it as a potential target term.

The final result of the process is to order the target terms. LitLinker ranks the target terms according to the number of linking terms that connect that target term the original starting term and prunes the ones with fewer linking terms than a previously selected linking term count threshold. Such a list should provide enough information to help researchers evaluate and explore these possible correlations to determine the ones which seem worthy for further investigation.

4. 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, and 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. They have reported overall success if one of the correlations generated by their systems was same as Swanson's discovery without evaluating the rest of the correlations. In contrast, we used a different evaluation approach that enabled us to evaluate all correlations that LitLinker generated. In our evaluation, starting with the same starting terms that Swanson used in his discoveries, we measured whether LitLinker leads us to new discoveries in the more recently published medical literature. To accomplish this goal, we divided MEDLINE into two parts: (1) a baseline literature including only publications before a selected date, and (2) a test literature including only publications between the baseline date and another later date. We ran LitLinker on the baseline literature that was published before the selected date and checked the generated connections in the test literature published after the selected date.

For our experiments, we ran LitLinker for the starting terms; *Alzheimer disease*, *migraine*, and *schizophrenia* on a MEDLINE 2004 baseline, which includes only documents published before January 1, 2004. We limited our results to only those terms in a semantic groups listed in Table 1 because the goal of our experiments was to find novel connections between the selected *diseases* and *chemicals*, *drugs*, *genes*, or *molecular sequences*. We checked the existence of target terms generated by LitLinker in the most recent starting term literatures, which we called test literatures. For our experiments, test literatures were composed of only articles included the starting terms of the experiments and published between January 1, 2004, which is the ending date of MEDLINE 2004 baseline, and September 30, 2005 (21 months).

We also used the information retrieval metrics, recall and precision, to gain a quantitative understanding for how well our system performed. To calculate the results, we first retrieved the MeSH terms that appeared in the test literatures but not in the starting term literatures from the MEDLINE 2004 baseline. Then, we filtered the retrieved lists of MeSH terms by using the semantic groups that we used for target term selections to find the ones that were chemicals, drugs, genes, or molecular sequences. We assumed that the MeSH terms in the remaining list would be new potential disease to gene or disease to drug treatment discoveries and used them as the gold standard for our precision and recall calculations. The formulas for precision and recall calculations are:

$$\text{Precision: } P_i = \frac{||T_i \cap G_i||}{||T_i||}, \tag{5}$$

$$\text{Recall: } R_i = \frac{||T_i \cap G_i||}{||G_i||}, \tag{6}$$

where T_i is the set of target terms generated by LitLinker for the starting term i, and G_i is the set of terms in the gold standard created from the test literature of starting term i.

We also drew *precision versus time* and *recall versus time* graphs to see how precision and recall values changed through time (Figs. 4–9). To accomplish this goal, we created subsets of the test literatures according to time and calculated precision and recall values for each subset. There were 21 time points on the x-axis of these graphs and each

time point represented the ending date of the corresponding subset of the test literature. For example, the first time point resulted in a subset of the test literature composed of documents published between January 1 and January 31, 2004 (1 month) and the last time point resulted in a subset of the test literature composed of documents published between January 1, 2004 and September 30, 2005 (21 months).

As in many data or text mining systems, there is no easy recipe for selecting the thresholds in our system. For the experiments, we presented in this paper, we set the z-score threshold to zero. We chose this threshold setting because we wanted LitLinker to select a term as a linking or target term only if its probability in the starting or linking literature is greater than or equal to its mean probability in the background literatures. Increasing the z-score would result in fewer but more strongly correlated linking and target terms. If we set the threshold too high, LitLinker might lose some interesting correlations. In addition to z-score threshold, we set another threshold called linking term count threshold (**LT**), to pick target terms with at least this many linking terms connecting it to the starting term. We used two different linking term count thresholds, three and five for further evaluation.

The following sections include detailed information about the results of the experiments. For each experiment, we first summarize the quantitative results, than we explain one of the connections that LitLinker identified and that also appeared as a potential discovery in recently published literature.

4.1. Alzheimer's disease

With the configuration described above, LitLinker identified 212 linking terms directly correlated with the starting term, Alzheimer disease. For the two different linking term count thresholds, LitLinker identified 600 target terms when LT = 3 and 250 target terms when LT = 5. We created the gold standard from 3983 documents that were both

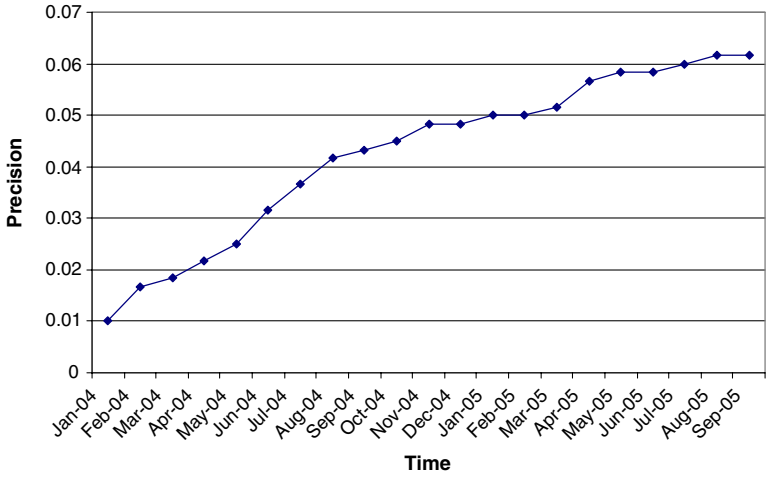

Fig. 4. Precision–time graph for Alzheimer's disease data (LT = 3). Precision values remained small but showed a positive trend by increasing as new discoveries appeared in the literature over time.

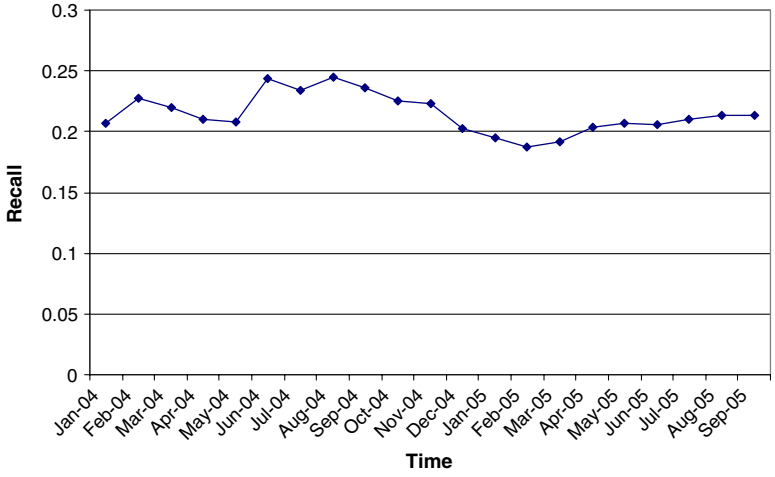

Fig. 5. Recall–time graph for Alzheimer's disease data (LT = 3). The recall values fluctuated as new discoveries appeared in the literature over time.

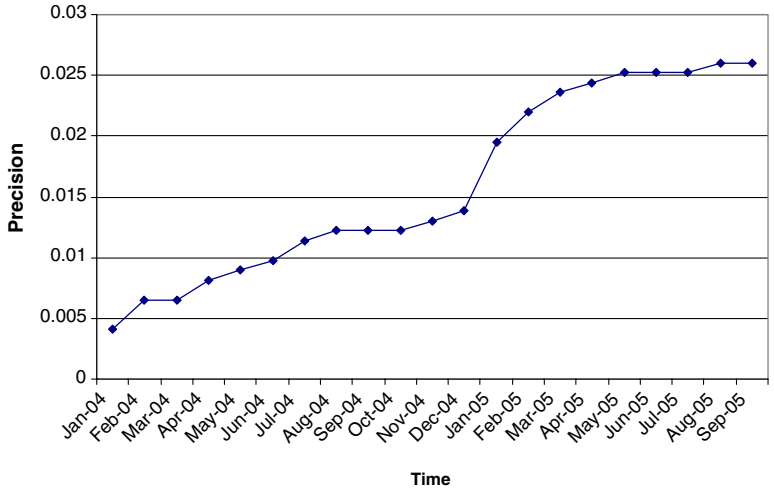

Fig. 6. Precision–time graph for migraine (LT = 3).

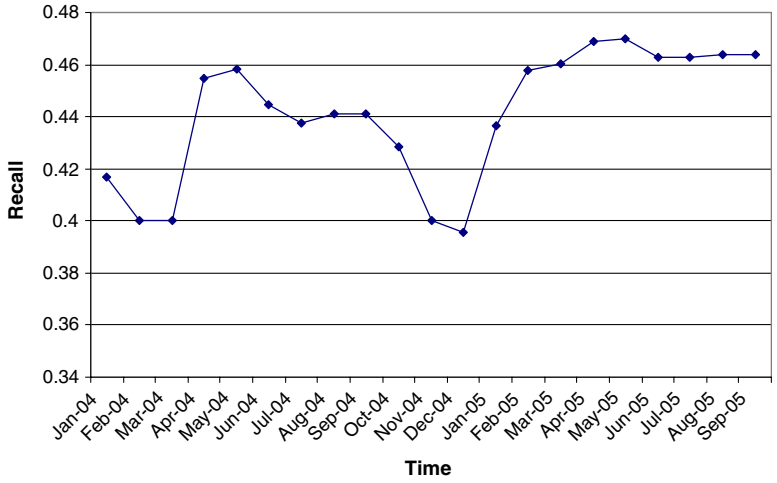

Fig. 7. Recall–time graph for migraine (LT = 3).

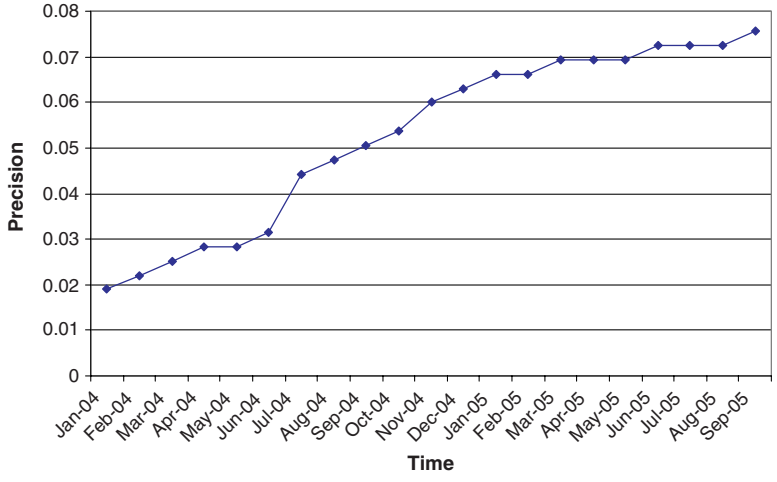

Fig. 8. Precision–time graph for schizophrenia (LT = 3).

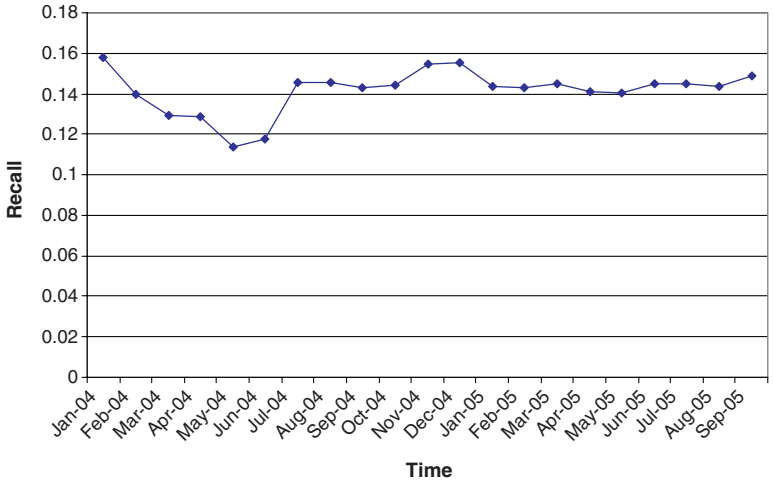

Fig. 9. Recall–time graph for schizophrenia (LT = 3).

published between January 1, 2004 and September 30, 2005 and contained the starting MeSH term Alzheimer disease. There were 143 MeSH terms in the gold standard.

Recall and precision values calculated for this experiment are given in Table 2. As can be observed from the table, the total number of target terms and the number of target terms in the gold standard decreased as the value of LT changed from three to five. Recall and precision are usually inversely correlated, and as expected, this threshold change affected the recall in a negative way and precision in a positive way.

The precision values were low but still promising. They showed LitLinker's ability to detect meaningful patterns validated by a very small test literature composed of documents published on Alzheimer disease in the last 21 months. We also do not know whether the other non-validated target terms are meaningful patterns or not yet. To see how precision changed through time, we drew a precision versus time graph for the data points collected for LT = 3 (Fig. 4). As we moved through time, the number of terms in the gold standard (G) and the number of target terms that appeared in the gold standard ($T \cap G$) increased and as a result of this the precision values increased. From this graph, it can be inferred that some of the non-validated target terms will be published in the future as new discoveries. We plan to conduct similar evaluations in the following months to check the state of non-validated target terms.

Table 2
Summary of results for Alzheimer's disease

	Linking term count threshold (LT)	
	LT = 3	LT = 5
No. of linking terms	212	212
No. of target terms (T)	600	250
No. of terms in gold standard (G)	173	173
No. of target terms in gold standard ($T \cap G$)	37	16
Recall	**0.214**	**0.092**
Precision	**0.062**	**0.064**

We also plan to work with domain experts to get their ideas on the usefulness of the patterns LitLinker generated.

We also drew the recall–time graph to see how recall changed through time (Fig. 5). Different from the increasing trend of the precision values in the precision–time graph, the recall values in the recall–time graph fluctuated between 0.245 and 0.19 with an overall average of 0.214. This occurred due to the parallel increase in the number of terms in the gold standard (G) and in the number of target terms that appeared in the gold standard ($T \cap G$) as we moved through time.

Discovery example: The MeSH term *endocannabinoids* was one of the target terms identified by LitLinker for the starting term Alzheimer disease. It was connected to Alzheimer disease via the following nine linking terms:

- *Basal Ganglia*[*1]
- Huntington disease
- Methoxyhydroxyphenylglycol
- Muscarinic agonists
- Neocortex
- *Neuroprotective agents*[*]
- Neurotransmitters
- Piperidines
- *Pyramidal cells*[*]

Endocannabinoids are defined as Marijuana-like substances act at specific receptors on the blood vessel wall to produce vasodilation. A research group from Spain recently published a paper about the possible role of endocannabinoid system in Alzheimer disease [12]. Pazos, et al., have reported the existence of profound changes in the location and density of several elements of this system in Alzheimer disease tissue samples. They investigated possible roles of cannabinoid receptor CB1, cannabinoid recep-

[1] Linking terms marked with * appear in the descriptions of the discoveries.

tor CB2, and endogenous cannabinoids which might open a new perspective for Alzheimer disease research.

Although, there has been no prior published work on the potential connection between endocannabinoids and Alzheimer disease, LitLinker could identify it by analyzing existing connections in the medical literature. Also, three of the linking terms that LitLinker identified appear in the researchers' description of the connections.

4.2. Migraine

A summary of the results generated by LitLinker for the starting term migraine is listed in Table 3. We created the gold standard from 1105 documents published between January 1, 2004 and September 30, 2005 and classified under migraine.

Precision–time and recall–time graphs for the starting term migraine are presented in Figs. 6 and 7.

Discovery example: LitLinker identified the MeSH term *AMPA receptors* as one of the target terms for the starting term migraine. It was connected to migraine via the following 12 linking terms.

- Anticonvulsants
- Benzocycloheptenes
- Brain ischemia
- Cerebral cortex
- *GABA agents**
- Neurogenic inflammation
- Neurotransmitters
- Nociceptors
- *Oxazoles**
- Piperidines
- Receptor, serotonin, and 5-HT1D
- Seizures

AMPA receptors are cell surface proteins that bind glutamate and directly gate ion channels in cell membranes. They are common mediators of fast excitatory synaptic transmission in the central nervous system. A recently published paper by Sang, et al., reports a potential role of LY293558, an AMPA receptor antagonist, in treating migraine [14]. In their study, they designed a randomized, triple-blinded, placebo-controlled, parallel group trial of 6 mg SC sumatriptan, 1.2 mg/kg IV LY293558, or placebo. From the statistical analysis of the data collected from 44 patients they concluded that LY293558 is promising for migraine treatment.

Starting from the existing body of migraine literature, LitLinker identified the potential correlation between migraine and AMPA receptors. Sang et al.'s study proves that such a connection exists for one type of AMPA receptors and this connection may lead to new directions in migraine treatment research.

4.3. Schizophrenia

The results generated by LitLinker for the starting term schizophrenia are summarized in Table 4. We created the gold standard from 3671 documents published between January 1, 2004 and September 30, 2005 and classified under schizophrenia.

Precision–time and recall–time graphs for the starting term schizophrenia are presented in Figs. 8 and 9.

Discovery example: *Secretin* was one of the target terms identified by LitLinker for the starting term Schizophrenia via the following four linking terms:

- *Autistic disorder**
- Flupenthixol
- Pirenzepine
- Sulpiride

Secretin is a peptide that simulates excretion of water and bicarbonate from the pancreas and biliary tree and secretion of digestive enzymes from the pancreas. Alamy and Sheitman et al. reported a possible connection between secretin and schizophrenia [1,15]. They formulated their hypothesis in a way that is similar to how LitLinker works. They knew that secretin might offer therapeutic benefit in autism and that autistic features can also be present in schizophrenia. They conducted a small pilot study of a single dose of porcine secretin for the treatment of refractory schizophrenia [15]. They were unable to demonstrate a significant statistical difference from the placebo treatment. However, they reported that several patients who received secretin infusions experienced clinically relevant improvements in symptoms. In another paper, the same group pre-

Table 3
Summary of results for migraine

	Linking term count threshold (LT)	
	LT = 3	LT = 5
No. of linking terms	250	250
No. of target terms (T)	1230	734
No. of terms in gold standard (G)	69	69
No. of target terms in gold standard ($T \cap G$)	32	21
Recall	**0.464**	**0.304**
Precision	**0.026**	**0.029**

Table 4
Summary of results for schizophrenia

	Linking term count threshold (LT)	
	LT = 3	LT = 5
No. of linking terms	211	211
No. of target terms (T)	317	124
No. of terms in gold standard (G)	161	161
No. of target terms in gold standard ($T \cap G$)	24	8
Recall	**0.149**	**0.05**
Precision	**0.076**	**0.064**

sented a case report about a 43 year-old male patient that supported their hypothesis on positive effects of using secretin for schizophrenia treatment [1].

For centuries, researchers have tried to find an effective cure for schizophrenia. As a result of these efforts, there is a huge body of literature on schizophrenia available in MEDLINE. Starting from this literature, LitLinker could automatically identify the potential connection between secretin and schizophrenia. One of the linking terms LitLinker identified was autistic disorder, which played a key role in Alamy and Sheitmans's work.

5. Conclusion

With the explosion of the scientific literature, literature-based discovery systems such as LitLinker will become critical for helping researchers discover connections across distinct portions of biomedicine. The main contribution of our research is our text mining architecture and our evaluation technique.

In this paper, we have shown that our mixed architecture of a statistical method based on word probability distributions and a knowledge-based approach can be incorporated into an effective system. By providing examples from the recently published papers as evidence, we have explained in detail three of the disease–protein correlations identified by LitLinker: *Alzheimer disease–endocannabinoids*, *migraine–AMPA receptors*, and *schizophrenia–secretin*. These discoveries have proven that LitLinker is capable of identifying novel and meaningful correlations between diseases and chemicals, drugs, genes, or molecular sequences in the published body of medical knowledge.

Another contribution of our research is the novel evaluation approach that we used to measure the performance of LitLinker. 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. Thus, we evaluated LitLinker with the latest correlations about the three selected diseases published in the last 21 months. LitLinker could successfully identify many of these recently discovered correlations in a pure open-ended-discovery process. Although, LitLinker's precision values were low, those values represent lower bounds because some of the non-validated correlations might be published in the future as new discoveries. Our experiments demonstrated that the precision values do increase over time as these new discoveries are published. Other researchers have focused mainly on replicating the linking terms in Swanson's examples, and none has provided quantitative evaluations for the other patterns their systems generated. Although, we used Swanson's starting terms to evaluate LitLinker, we are not restricted to only Swanson's discoveries with our evaluation approach. We can automatically evaluate our system for any starting term.

Using three example discoveries, we have shown that LitLinker provides a new and effective type of knowledge discovery approach. Unlike the information retrieval tools currently available to medical researchers, such as PubMed, LitLinker generates results about possible new connections between medical terms. LitLinker also provides an interactive web interface to display the identified correlations in an effective way [17]. This new discovery system will help medical researchers capture, and explore new connections in the vast biomedical literature to help them identify new research directions.

Acknowledgments

The National Science Foundation, award #IIS-0133973, funded this work.

References

[1] Alamy SS, Jarskog LF, Sheitman BB, Lieberman JA. Secretin in a patient with treatment-resistant schizophrenia and prominent autistic features. Schizophr Res 2004;66(2–3):183–6.
[2] Andrade MA, Valencia A. Automatic extraction of keyword from scientific text: application to the knowledge domain of protein families. Bioinformatics 1998;14(7):600–7.
[3] Blake C, Pratt W. Automatically identifying candidate treatments from existing medical literature. Proceedings of AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases, Palo Alto, CA, 2002.
[4] Gordon M, Lindsay RK, Fan W. Literature-based discovery on the World Wide Web. ACM Trans Internet Technol 2002;2(4):261–75.
[5] Gordon MD, Dumais S. Using latent semantic indexing for literature based discovery. J Am Soc Inf Sci 1998;49(8):674–85.
[6] Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Improving literature based discovery support by genetic knowledge integration. Stud Health Technol Inf 2003;95:68–73.
[7] Lindsay RK, Gordon MD. Literature-based discovery by lexical statistics. J Am Soc Inf Sci 1999;50(7):574–87.
[8] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Proceedings of Medinfo, San Francisco, 2001.
[9] National Library of Medicine. Medical Subject Headings (MeSH) Fact Sheet. Available from: <http://www.nlm.nih.gov/pubs/factsheets/mesh.html/>, 2005.
[10] National Library of Medicine. MEDLINE Fact Sheet. Available from: <http://www.nlm.nih.gov/pubs/factsheets/medline.html/>, 2005.
[11] National Library of Medicine. Unified Medical Language System Fact Sheet. Available from: <http://www.nlm.nih.gov/pubs/factsheets/umls.html/>, 2005.
[12] Pazos MR, Nunez E, Benito C, Tolon RM, Romero J. Role of the endocannabionoid system in Alzheimer's disease: new perspectives. Life Sci 2004;75(16):1907–15.
[13] Pratt W, Yetisgen-Yildiz M. LitLinker: capturing connections across the biomedical literature. Proceedings of the International Conference on Knowledge Capture (K-CAP'03), FL, 2003.
[14] Sang CN, Ramadan NM, Wallihan RG, Chappell AS, Freitag FG, Smith TR, et al. LY293558, a novel AMPA/GluR5 antagonist, is efficacious and well-tolerated in acute migraine. Cephalalgia 2004;24(7):596–602.
[15] Sheitman BB, Knable MB, Jarskog LF, Chakos M, Boyce LH, Early J, et al. Secretin for refractory schizophrenia. Schizophr Res 2004;66(2–3):177–81.
[16] Srinivasan P. Generating hypotheses from MEDLINE. J Am Soc Inf Sci Technol 2004;55(5):396–413.
[17] Skeels M, Henning K, Yetisgen-Yildiz, M, Pratt W. Interaction design for literature-based discovery. Proceedings of the International Conference for Human–Computer Interaction (CHI'05), Portland, 2005.

[18] Swanson DR. Migraine and magnesium: eleven neglected connections. Perspect Biol Med 1988;31:526–57.
[19] Swanson DR. Medical literature as potential source of new knowledge. Bull Med Libr Assoc 1990;78(1):29–37.
[20] Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif Intell 1997;91:183–203.
[21] Swanson DR. Online search for logically related non-interactive medical literatures: A systematic trial-and-error strategy. J Am Soc Inf Sci 1989;40(5):356–8.
[22] Weeber M, Klein H, de Jong-van den Berg LTW. Using concepts in literature based discovery: simulating Swanson's Raynaud–fish oil and migraine–magnesium examples. J Am Soc Inf Sci Technol 2001;52(7):548–57.
[23] Weeber M, Vos R, Klein H, de Jong-van den Berg LTW, Aronson AR, Molema G. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. J Am Med Inf Assoc 2003;10(3):252–9.
[24] Wren JD. Extending the mutual information measure to rank inferred literature relationships. BMC Informatics 2004;5(1):145–58.