Journal of Biomedical Informatics 42 (2009) 633–643

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A new evaluation methodology for literature-based discovery systems

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

[a] Kiha, Inc., 100 S. King Street, Suite 320, Seattle, WA 98104, USA
[b] Information School, University of Washington, Seattle, WA, USA
[c] Biomedical & Health Informatics, University of Washington, Seattle, WA, USA

ARTICLE INFO

Article history:
Received 6 July 2008
Available online 16 December 2008

Keywords:
Literature-based discovery
System evaluation

ABSTRACT

While medical researchers formulate new hypotheses to test, they need to identify connections to their work from other parts of the medical literature. However, the current volume of information has become a great barrier for this task. Recently, many literature-based discovery (LBD) systems have been developed to help researchers identify new knowledge that bridges gaps across distinct sections of the medical literature. Each LBD system uses different methods for mining the connections from text and ranking the identified connections, but none of the currently available LBD evaluation approaches can be used to compare the effectiveness of these methods. In this paper, we present an evaluation methodology for LBD systems that allows comparisons across different systems. We demonstrate the abilities of our evaluation methodology by using it to compare the performance of different correlation-mining and ranking approaches used by existing LBD systems. This evaluation methodology should help other researchers compare approaches, make informed algorithm choices, and ultimately help to improve the performance of LBD systems overall.

1. Introduction

Scientific literature is readily available, but the sheer volume and growth rate of the literature makes it impossible for medical researchers to keep up with new findings outside their own narrowing fields of expertise. To address this need, many literature-based discovery (LBD) systems have been developed to find new connections between biomedical terms that could lead to new directions in research. Although many online LBD systems are available, they are not in routine use by medical researchers.

Evaluation plays an important role in the development of new fields such as LBD. Evaluation encourages scientific progress by supporting a systematic comparison of different techniques applied to a common problem and allowing researchers to learn from each other's successes and failures. However, current evaluation methods used in LBD research are focused on only one system and do not support system comparisons. In this paper, we address those limitations through two main aims. The first aim is to propose an alternative evaluation methodology for LBD systems that allows comparisons across different systems. We will give a detailed review of the current evaluation approaches used by other LBD researchers and describe our evaluation methodology by discussing its strengths and weaknesses. The second aim is to demonstrate the abilities of our evaluation methodology by using it to compare the performance of different correlation-mining and ranking approaches used by the existing LBD systems. To our knowledge, this paper will be the first LBD evaluation methodology paper that reports performance comparisons across many different algorithmic approaches.

2. Literature-based discovery

Swanson initiated the literature-based discovery (LBD) research and was responsible for much of the earliest work in this area [1]. 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* [2]. 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*.

By using his discovery approach, Swanson, with his close collaborator Smalheiser who is a medical doctor and neuroscientist,

[*] Corresponding author.
E-mail addresses: meliha@kiha.com, wpratt@u.washington.edu (M. Yetisgen-Yildiz).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.12.001

published several different hypotheses about causally connected medical terms in the biomedical domain including *Migraine–Magnesium* [2], *Raynaud's Disease–Fish Oil* [3], *Alzheimer's Disease–Estrogen* [4], *Alzheimer's Disease–Indomethacin* [5], *Somatomedin C–Arginine* [6], and *Schizophrenia–Calcium Independent Phospholipase* A_2 [7]. Swanson's work introduced seminal ideas for literature-based discovery. Many other researchers replicated his approach of taking advantage of an intermediate linking literature and developed LBD systems to support hypothesis generation [8–22].

The general architecture of those LBD systems that were designed with what Swanson calls an open-discovery approach is represented in Fig. 1. The discovery process begins with a **starting term** (e.g. *migraine*), the term that the researcher is interested in investigating. Next, the LBD system uses a **correlation-mining approach** to find the terms that are directly correlated with the starting term. We refer to these first correlated terms as the **linking terms** (e.g. *epilepsy* and *calcium channel blockers* for the *migraine* starting term example) because they link the starting term to target terms. For each of the linking terms, the LBD system then uses the same correlation-mining approach to identify terms that are correlated with each linking term. We call these final correlated terms **target terms** (e.g. *magnesium*). We assume that the linking terms are already known connections to the starting term, and that target terms are the potential new discoveries. Finally, the LBD system orders the target terms with a **ranking approach**.

In this paper, we will describe four different correlation-mining approaches (*Association Rules* [11], *TF-IDF* [13,16], *Z-Score* [21], and *Mutual Information Measure (MIM)* [20]) and three different target-term ranking approaches (*Average Minimum Weight* [20], *Linking Term Count* [21], and *Literature Cohesiveness* [23]) that have been previously used by researchers in their LBD systems. We will use our LBD system, called LitLinker, as a platform to evaluate those correlation-mining and ranking approaches. LitLinker uses MEDLINE as the main resource for biomedical literature and represents the content of MEDLINE documents with Medical Subject Headings (MeSH) [24]. More detailed information about LitLinker's system architecture can be found in [14,21].

3. Current evaluation approaches

Evaluating open-discovery LBD systems is a fundamentally challenging task because if LBD systems are successful, by definition they are capturing new knowledge that has yet to be proven useful. To overcome this challenge, many LBD researchers have measured system performance by replicating historical discoveries, typically using Swanson's published discoveries as the gold standards. For those evaluations, LBD researchers reported overall success if one of the correlations generated by their system matched Swanson's discovery without evaluating the rest of the generated target terms [8–10,12–14,16,17,20].

As an alternative evaluation approach, some researchers incorporated medical expert knowledge into the evaluation process of their LBD systems. Weeber et al. used their discovery system to investigate new potential uses for drug *thalidomide* with Swanson's open-discovery approach [19]. One of the researchers involved in this study was a medical researcher with a background in pharmacology and immunology. For the starting term *thalidomide*, their system generated a list of linking terms that were constrained to be immunologic factors. They manually selected the promising linking terms with the involvement of the medical researcher. For the selected linking terms, their system generated a list of target terms that were constrained to be disease or syndrome names. The medical researcher manually assessed each of the selected diseases. Based on the assessment, they compiled a list of four diseases: *chronic hepatitis C*, *myasthenia gravis*, *helicobacter pylori induced gastritis*, *acute pancreatitis* for which the researcher hypothesized that *thalidomide* could be an effective treatment.

Srinivasan and Libbus evaluated their system Manjal by using a semi-automated approach with experts as well. In their experiment, they used *turmeric,* a widely used spice in Asia, as their starting term. The aim for their experiment was to identify diseases where *turmeric* could be useful in their treatment. Libbus, who has a medical background, identified the correlated terms in the turmeric literature manually [25]. The system used those terms as the intermediate linking terms, and the resulting target terms were designated as a gold standard. They then compared the original target terms that Manjal identified against the new gold standard. This approach highly depends on the subjective decision of the medical researcher in deciding upon the linking terms as well as in evaluating the target terms.

Wren et al. also incorporated medical expert knowledge into the evaluation process [26]. Their literature-based discovery system started with the term *cardiac hypertrophy* and identified a total of 2102 linking terms and 19,718 target terms. To evaluate their system, the researchers performed laboratory experiments on one target term—the 3rd ranked target term, *chlorpromazine*. *Chlorpromazine* is a chemical that is used as an anti-psychotic and anti-emetic drug. In their lab experiments, they looked for

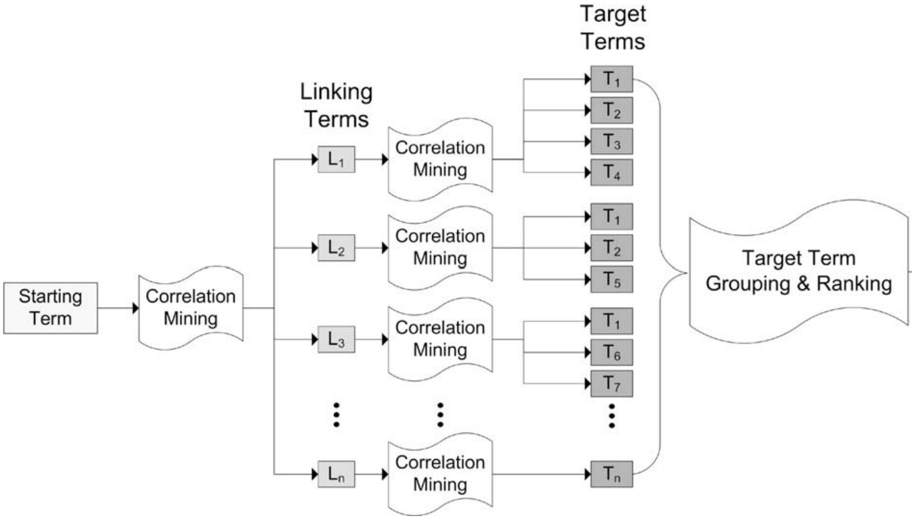

Fig. 1. General architecture of an LBD system that supports open-discovery approach.

an association between *chlorpromazine* and *cardiac hypertrophy*. They reported that *chlorpromazine* could reduce *cardiac hypertrophy* in mice.

Torvik and Smalheiser evaluated their system Arrowsmith with a gold standard created by medical experts [27]. Arrowsmith supports Swanson's closed-discovery approach, where the system identifies the linking terms that connect a user provided starting term to a user provided target term. In their approach, six domain experts selected six starting term—target term couples and they manually identified linking terms to use as gold standards in the evaluation of the quantitative modeling of Arrowsmith. Because they use a closed-discovery approach, their evaluation only assesses linking terms.

Hristovski et al. performed a statistical evaluation of their system, BITOLA [11,28]. The purpose of their evaluation was to see how many of the potential discoveries made by their system at some point in time become realized at a later time. To accomplish this goal, they ran their system for the starting term *Multiple Sclerosis* on the set of documents published between 1990 and 1995. They checked the existence of the proposed discoveries in the set of documents published between 1996 and 1999 and calculated precision and recall. They used a very limited portion of MEDLINE and reported the performance statistics of their system without comparing it to those of other systems.

In [21], we used a similar but more extensive approach to evaluate our system. We used the whole MEDLINE 2004 Baseline as the document base, and calculated information retrieval metrics as well as precision-time and recall-time curves. In this paper, we extend our evaluation approach further to compare different correlation-mining algorithms and ranking approaches. A detailed review of LBD evaluation approaches can be found in [29].

4. Principles of LBD system evaluation

Our main purpose in defining an evaluation methodology is to add a degree of standardization to the process of evaluating LBD systems. Without a systematic approach, evaluations offer little more than ad hoc evidence of the value of an LBD system. While investigating the current evaluation approaches, we identified four evaluation principles where the current approaches fail.

1. *The evaluation methodology should evaluate the quality of all target terms produced by an LBD system.* Although all the researchers who used Swanson's discoveries for evaluation have successfully replicated his discoveries, this type of evaluation does not inform us about the quality of the complete set of target terms identified by their systems. Depending on the approaches used to select the correlated terms, an LBD system might return hundreds or even thousands of target terms for a given starting term. Evaluating the whole system on only one of those target terms does not guarantee that the rest of the target terms would yield useful discoveries. As with information retrieval systems, an LBD system that returns a single helpful target term in a sea of unhelpful target terms is unlikely to be useful.
2. *The evaluation methodology should be based on the analysis of the results of a number of well-defined experiments.* The researchers who replicated Swanson's discoveries are limited in their evaluations to the small number of discoveries published by Swanson. All those researchers, except Srinivasan, used one or at most two of Swanson discoveries in their evaluations. The performance reported for one or two discovery cases cannot be used to generalize the true performance of an LBD system. Thus, the evaluation methodology must be repeatable for different starting terms.
3. *The evaluation methodology should be independent of any prior knowledge to ensure the methodology is not biased for or against the LBD system under evaluation.* In his papers, Swanson described each of his discoveries in great detail, and it would be difficult for other researchers to ignore those helpful details in designing their evaluations. In other words, the researchers who replicated Swanson's discoveries knew exactly what they were seeking as the desired target and linking terms. It would be easy to tune the parameters of their systems to identify those terms. Such an approach might result in LBD systems that perform well for the specific example cases but not well for other cases.
4. *The evaluation methodology should enable comparing the performance of different systems.* Performance comparison is the essence of evaluation. However, neither replicating Swanson's discoveries nor incorporating expert knowledge to system evaluation allow detailed comparisons among different LBD systems. Replicating Swanson's discoveries allows the researchers to say system *A* is better than another system *B* if *A* replicates a selected discovery but *B* does not. However, if both *A* and *B* successfully replicate the given discovery successfully, it becomes impossible to determine which system is superior to the other. For evaluations that incorporate expert knowledge, it becomes difficult to quantify the results, and thus hard to use to compare different LBD systems. In addition, because the aim of LBD tools is to identify novel correlations, disagreements on the interestingness of the correlations could arise if multiple medical researchers are involved in the evaluation process, and there is no easy way to resolve those disagreements.

5. Evaluation methodology

The purpose of our evaluation methodology is to provide a way to compare multiple algorithms used in LBD systems and to compare their use on the recently published medical literature. To accomplish this goal, we divide MEDLINE into two sets: (1) a **pre-cut-off set** including only publications before a selected cut-off date and (2) a **post-cut-off set** including only publications between the cut-off date and another later date. We ran LitLinker, using various algorithms used in LBD systems, on the pre-cut-off set and checked the generated connections in the post-cut-off set (Fig. 2). In other words, we define the generated connections from the pre-cut-off set to be discoveries, if they co-occur with the starting term in the post-cut-off set. We quantify the overall performance by using information retrieval evaluation methods. The steps of the evaluation methodology are listed in Fig. 3.

Our evaluation approach adheres to the four evaluation principles we listed in the previous section. (1) It evaluates the quality of all target terms produced by an LBD system by using the gold standard created from the post-cut-off set for the starting term. (2) The evaluation methodology is repeatable for as many starting terms as the researcher wants. (3) It is independent of any prior knowledge. As an example, in the evaluations we present in this paper, we used 100 different starting terms that were randomly selected from disease names available in MeSH. (4) Because the evaluation methodology quantifies the system performance by using IR metrics, it can be used to compare the performance of different systems, and in the second half of this paper, we use this methodology to compare different algorithms used in LBD systems. In the following sections, we will explain each step of our evaluation methodology in detail.

5.1. Selecting the cut-off date

Selecting the cut-off date plays an important role in our evaluation approach because it affects the amount of time a hypothesized discovery has to turn into a true discovery. For example,

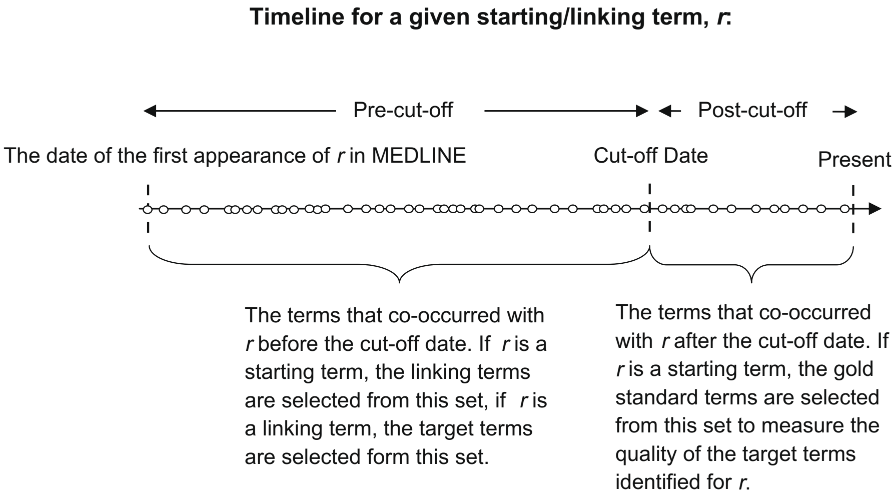
Timeline for a given starting/linking term, r:
Pre-cut-off
Post-cut-off
The date of the first appearance of r in MEDLINE
Cut-off Date
Present
The terms that co-occurred with r before the cut-off date. If r is a starting term, the linking terms are selected from this set, if r is a linking term, the target terms are selected form this set.
The terms that co-occurred with r after the cut-off date. If r is a starting term, the gold standard terms are selected from this set to measure the quality of the target terms identified for r.

Fig. 2. Timeline for a given starting/linking term *r*. The dots represent the terms that co-occur with *r* and their locations on the timeline represent their first co-occurrence dates with *r*.

For a given starting term r:

1. Select a *cut-off date* and divide MEDLINE into two sets:
   i. a *pre-cut-off set B* including only publications before a selected *cut-off date,*
   ii. a *post-cut-off set T* including only publications after the cut-off date.
2. Run the LBD system on the *pre-cut-off set B* for the *starting term r* and collect the list of *target terms T*.
3. Create a *gold standard G* for the *starting term r* from the *post-cut-off set T*.
   i. Extract all the *index terms*, $I_{post-cut-off}$, that co-occurred with the *starting term r* in the *post-cut-off set T*.
   ii. Extract all the *index terms*, $I_{pre-cut-off}$, that co-occurred with the *starting term r* in the *pre-cut-off set B*.
   iii. Take the difference of $I_{post-cut-off}$ and $I_{pre-cut-off}$ and label the remaining set of the terms as *gold standard terms* ($G = I_{post-cut-off} - I_{pre-cut-off}$).
   iv. If a pruning mechanism (i.e. pruning with semantic types) is used in the selection of the target terms, prune the *gold standard G* with the same mechanism.
4. Check the existence of target terms *T* in the *gold standard G* and use *information retrieval (IR)* evaluation metrics to report the overall performance.

Fig. 3. Steps of the evaluation methodology. Note that the set of terms in 3*i* ($I_{post-cut-off}$) includes the terms that co-occurred with the starting term in the post-cut-off set. Because we are only interested in novel discoveries, we took the difference between the two sets and created a gold standard ($G = I_{post-cut-off} - I_{pre-cut-off}$) from the terms that co-occurred with the starting term for the first time after the cut-off date. Each of the terms in *G* can be identified by LitLinker as a target term.

one of the reasons for the low overall performance in our evaluation reported in [21] was that a post-cut-off set composed of documents published in a short amount of time, in this case 21 months, was insufficient to evaluate the predicted correlations. Discoveries in the medical field, as in other scientific fields, take time. Some of the non-validated target terms from our study might be published in the future as new discoveries. To solve this problem, the post-cut-off set should span a reasonable amount of time for new discoveries to emerge. Determining that time is admittedly subjective. It is also critical to select a cut-off date that is much later than the date of the first appearance of the starting term in the medical literature. Otherwise, the pre-cut-off set would include only a very small set of documents indexed with the starting term, and with such a restricted pre-cut-off set, the LBD system would produce few linking terms. Thus, the LBD system would have too little information about the starting term to produce many target terms.

5.2. Running the LBD system

LBD systems use different correlation-mining approaches to identify the linking and target terms. All the correlation-mining approaches that we will describe in the following section require decisions to set thresholds. Those thresholds directly affect the quality and the quantity of the linking and target terms generated for a given starting term. As in many data or text mining systems, there is no easy recipe for selecting the thresholds in LBD systems. They must be selected with care to satisfy the goals of the discovery process.

5.3. Creating the gold standard

One of the most challenging components of evaluating an LBD algorithm's performance is determining the gold standard on which to compare the LBD's predictions. In this paper, we chose to define a discovery as a new co-occurrence between terms. This choice allowed us to generate a gold standard automatically from the post-cut-off set. The gold-standard set includes all the terms that (1) are found in test-set documents that contain the starting term and (2) did not co-occur with the starting term in the pre-cut-off set because those terms are presumably already known and not new discoveries. If the LBD system uses a pruning criteria to select the interesting linking and target terms, the gold-standard set could be further reduced by that criteria. The most commonly used pruning criteria to filter linking and target terms is pruning by semantic types [8,12,16,21].

5.4. Calculating the evaluation metrics

We use the information retrieval metrics, recall and precision, to gain a quantitative understanding for how well an LBD system performs. In the context of LBD, precision is defined as the fraction of target terms that are in the gold standard, and recall is defined as the fraction of gold standard terms that are retrieved. The formulas for precision and recall calculations are:

$$\text{Precision}: P_i = \frac{\|T_i \cap G_i\|}{\|T_i\|} \tag{1}$$

$$\text{Recall}: R_i = \frac{\|T_i \cap G_i\|}{\|G_i\|} \tag{2}$$

where T_i is the set of target terms generated by the LBD system 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*.

Precision and recall are effective metrics to evaluate the complete set of target terms but those metrics are set based and do not take the ranking of target terms into consideration. Because most researchers are unlikely to examine the entire set of target terms, we also use the following three evaluation metrics that consider the ranking: (1) 11-point average interpolated precision, (2) precision at *k*, and (3) mean average precision. In **11-point average interpolated precision curves** [30], for a given starting term, the target terms are ranked and individual precision values are interpolated to 11 standard recall levels (0 to 1 in increments of 0.1). The interpolated precision at standard recall level *i* is set to the maximum precision obtained for any actual recall level greater than or equal to *i*. Although precision at a recall level of 0.0 is not defined, this interpolation rule defines an interpolated precision value for recall level 0.0. If there is more than one starting term, the measured precision values for each recall level is averaged over the set of starting terms. With 11-point precision recall curves, we can observe how precision changes as recall levels increase for different LBD systems and compare the performances of those systems. The curves that are closer to the upper right corner of the graph perform better than those that are closer to the lower left corner.

We also use **precision at *k*** [30], which takes precision at the top-ranked *k* target terms as a measure of the performance of LBD systems. We assume that the target terms that are ranked above a certain threshold (*k*) are more important to users than those that are ranked below the threshold because users are likely to look at the highest ranked terms first.

The third metric we use to evaluate LBD algorithm's performance is **mean average precision** (**MAP**) [30]. In this metric, for a given starting term, the target terms are ranked with a ranking approach. Precision values are calculated after each gold standard term identified in the ranked list of target terms and MAP is calculated by taking the average of those calculated precision values. Unlike the other metrics, MAP provides a single-value summary of system performance that can easily be used to compare different algorithms.

6. Correlation-mining and ranking algorithms used in LBD systems

In the first half of the paper, we described the evaluation methodology in detail. In this section, we will summarize different correlation-mining and target-term ranking algorithms previously used by researchers in designing their LBD systems. We used the evaluation methodology to compare the performance of those algorithms. To accomplish this task, we implemented each algorithm in our LBD system LitLinker and used LitLinker as a platform in our evaluations.

6.1. Correlation-mining algorithms

We implemented four different correlation-mining algorithms for identifying linking and target terms in LitLinker. The first two algorithms, *Association rules* and *TF-IDF* are based on term co-occurrence frequencies. The last two algorithms, *Z-Score* and *Mutual Information Measure* (*MIM*), are based on term co-occurrence probabilities.

6.1.1. Association rules

Association rules were originally developed with the purpose of market basket analysis [31]. A market basket is a collection of items purchased by a customer in a single transaction. Given a set of transactions, where each transactions is a set of items, an association rule is defined as an expression $A \rightarrow B$, where *A* and *B* are sets of items. An association rule can be interpreted as if every item in *A* is purchased then it is likely that the items in *B* will also be purchased.

In the context of LBD, Hristovski et al. applied association rules to identify the correlated biomedical terms [11]. In their application, transactions are documents and items are words or descriptors used for indexing the documents. To identify the linking/target terms, MEDLINE is mined for association rules, $A \rightarrow B$, where *A* is the starting/linking term and *B* is the linking/target term. The two important measures for an association rule are:

- Support: Medical terms *A* and *B* are correlated if they co-occur together in many documents.

$$\text{Support}: s = \|D_A \cap D_B\| \tag{3}$$

- Confidence: Medical terms *A* and *B* are correlated if the probability of *B* appear in a document, given that *A* appears in the document is high.

$$\text{Confidence}: c = \frac{\|D_A \cap D_B\|}{\|D_A\|} \tag{4}$$

where D_A is the set of documents that include *A* and D_B is the set of documents that include *B*.

Thresholds for support and confidence are selected to identify interesting and useful association rules. The set of rules that pass the threshold test are used to extract the linking/target terms.

6.1.2. TF-IDF

Term Frequency-Inverse Term Frequency (**TF-IDF**) is a statistical measure used to evaluate how important a word is to a document in a collection of documents [32]. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the collection of documents.

Lindsay et al. [13] and Srinivasan [16] have adopted TF-IDF to identify correlated biomedical terms in their LBD systems. We implemented Lindsay et al.'s approach and calculated the TF-IDF value of a term couple, *l* and *m*, as:

$$TF\text{-}IDF_l^m = TF_l^m \times IDF_l = t_l^m \times \log \frac{n}{t_l} \tag{5}$$

where t_l^m is the number of documents that include both *l* and *m*, *n* is the size of the document collection and t_l is the number of documents that include *l*. For a given starting/linking term, all the terms with a TF-IDF value higher than a selected threshold are selected as correlated terms and labeled as linking/target terms.

6.1.3. Z-score

In contrast to the previous two algorithms, ***Z*-score** focuses on term probability distributions rather than term co-occurrence frequencies to mine term correlations from text [21]. Suppose *V* is the set of terms that are used to index the MEDLINE literature. In this approach, the MEDLINE literature is divided into subsets where there is one subset for each term $l \in V$, and it includes the documents indexed with the term *l*. The probability of a term appearing in a subset is calculated by dividing the number of documents of the subset in which the term appeared by the total number documents in the subset. Using this approach, the subset sizes are added into the process of finding correlations. The probability, *P*, of a term, *m*, appearing in a subset, *l*, is calculated as:

$$P_l^m = \frac{F_l^m}{D_l}, \tag{6}$$

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

$$\overline{P^m} = \frac{\sum_{i=1}^{N} P_i^m}{N}, \tag{7}$$

where *N* is the total number of subsets that contain the term *m*.

The mean probability of a 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 subsets. The combination of the mean probability with the deviation of the term probability distribution is more indicative than only mean probability. From (7) and (8) the deviation of the term probability distribution for the term *m* can be calculated as:

$$\sigma^m = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (P_i^m - \overline{P^m})^2} \tag{8}$$

The *Z*-score of a term *m* in the starting or linking term subset *l* can be calculated from (7)–(9) as:

$$Z\text{-}score(l, m) = \frac{P_l^m - \overline{P^m}}{\sigma^m} \tag{9}$$

This score provides the distance between the probability of a term in a specific subset and the general distribution of this term in the background set of subsets. The terms with *Z*-scores larger than a predefined threshold are marked as the correlated terms to the starting or linking term.

6.1.4. Mutual Information Measure

Wren used **Mutual Information Measure** (**MIM**) to identify the correlated terms [20]. MIM has been widely applied to quantify dependencies between variables, including co-occurring terms in text. The MIM score of two terms, *m* and *n*, is calculated as:

$$MIM(m, n) = \log_2 \frac{P_{mn}}{P_m p_n} \tag{10}$$

where P_{mn} is the joint probability of terms *m* and *n* appearing together in the same document, while P_m and P_n are the probabilities of observing the terms *m* or *n*, respectively, in any given document.

In the context of term co-occurrences, the ratio between P_{mn}, and the product of P_m and P_n is a measure of the degree of statistical dependence between the terms. The log of this ratio is the amount of information that we acquire about the presence of one of the terms when we observe the other. If the terms are independent, the product of P_m and P_n is equal to P_{mn} and this equality produces a MIM score equal to zero. If the probability of observing the term *m* increases when the term *n* is mentioned, then the MIM score is greater than zero. If the two terms are rarely mentioned together then the MIM score is less than zero.

Wren also applied the time-dependence property of scientific research and discovery to the calculation of probabilities, P_{mn}, P_m, and P_n. For example, the discovery of a new gene will occur at a given point in time within the history of medical publications. Regardless of the gene's overall frequency in the whole medical literature, the probability it will appear in the literature prior to its discovery is zero. Thus P_m and P_n are calculated from their time of first appearance in the medical literature and P_{mn} is calculated using the latter of these two dates as:

$$P_m = \frac{T_m}{T - S_m} \tag{11}$$

$$P_n = \frac{T_n}{T - S_n} \tag{12}$$

$$P_{mn} = \frac{T_{mn}}{T - \max(S_m, S_n)} \tag{13}$$

where *T* is the total number of documents in MEDLINE, T_m is the total number of documents with the *m*, T_n is the total number of documents with the *n*, and T_{mn} is the total number of documents with both terms *m* and *n*. S_m and S_n represent the number of documents were published before the first occurrences of terms *m* and *n* and max (S_m, S_n) is a function that returns the larger of the two values.

The terms with MIM scores larger than a predefined threshold are marked as the correlated terms to the starting or linking term.

6.2. Target-term ranking algorithms

For a given starting term, LBD systems produce a list of target terms as potential discoveries. These target terms must be ranked in a way that provides enough information to help researchers while they explore these correlations to determine the ones that seem worthy of further investigation.

Ranking target terms from all the linking terms requires multiple processing steps. First, an LBD system needs to merge the lists of correlated terms from each of the linking terms. Second, because we are only interested in novel connections, the LBD system 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 and removed such terms from the list of target terms. Because LitLinker uses MeSH terms to represent the content of documents, the co-occurrence check was based on MeSH term co-occurrences. In following sections, we will describe three different ranking algorithms to order the final set of target terms.

6.2.1. Average Minimum Weight (AMW)

Average Minimum Weight (**AMW**) proposed by Wren is based on the assumption of inferring a correlation between a starting term and a target term depends upon how much information was in the starting term—linking term association as well as the linking term—target term association, and the overall information association of the inferred association would be no greater than the least mutual information given by these two associations [20]. With this assumption, the ranking score between a starting term S and a target term T is defined as:

$$Score(S,T) = \frac{\sum_{i=1}^{n} \min(MIM(S,L_i), MIM(L_i,T))}{n} \quad (14)$$

where L_i is the ith linking term and n is the number of linking terms that connect S to T. We also adopted this score for Z-score, association rules, and TF-IDF. An example case for Z-score is as follows:

$$Score(S,T) = \frac{\sum_{i=1}^{n} \min(Z\text{-}score(S,L_i), Z\text{-}score(L_i,T))}{n} \quad (15)$$

6.2.2. Linking Term Count with Average Minimum Weight (LTC-AMW)

An alternative assumption could be that the number of linking terms that connect the target terms to the original starting term is the main indication of a strong correlation [14,21]. With this assumption, a proposed correlation with m linking terms is stronger than another with n linking terms if $m > n$. **LTC-AMW** makes this assumption and ranks the target terms according to the number of linking terms that connect the target terms to the original starting term. However, in this ranking approach, some target terms may have the same number of linking terms. Such target terms are ordered according to their average minimum weight. These secondary scores for Z-score, MIM, association rules, and TF-IDF are calculated as described in the previous section.

6.2.3. Literature Cohesiveness (COH)

Swanson et al. defined a new MeSH based measure called **Literature Cohesiveness** (**COH**) [23]. They used this measure to sort the linking terms identified by their LBD system. In this approach, for a given term, literature is defined as the set of documents returned by a MEDLINE search for the given term. For a given literature L, a threshold k is calculated as:

$$k = \text{int}(1.7 \times \ln(u) + 0.5) \quad (16)$$

where u is the number of unique MeSH terms in L and int is a function to round off k to the nearest integer. With this threshold, the cohesiveness of L is calculated with the following formula.

$$\text{COH} = \frac{top}{top + rem} \quad (17)$$

where *top* is the sum of frequencies of MeSH terms in L in the range of 2 to $k + 1$ and *rem* is the sum of remaining frequencies.

In [23], Swanson et al. did not explain the reasons that lie under the selection of multiplier 1.7 in Eq. (17). They used this measure to rank the linking terms between a given starting term and a target term. In this paper, we used their measure to sort the target terms with the same multiplier.

7. Evaluation

In the previous sections, we explained our evaluation methodology and summarized four correlation-mining algorithms and three ranking algorithms used in LBD systems. In this section, we will use our evaluation methodology to compare the performance of those correlation-mining algorithms and ranking algorithms.

7.1. Experimental setup

In our experiments, we ran our LBD system LitLinker for 100 different starting terms. The starting terms were disease names that were randomly selected from MeSH. We used MEDLINE 2008 Baseline[1] as the knowledge source for biomedicine. MEDLINE includes documents from various publication types, but we found that the documents with some publication types (e.g., comment, biography, dictionary, and lectures) were not useful for discovery because they did not contain research results. We manually determined the publication types to exclude and eliminated the documents with those publication types from LitLinker's search space. Table 1 includes the complete list of excluded publication types.

In our experiments, we set the cut-off date to January 1, 2000, which resulted in a pre-cut-off set composed of 11,695,196 million documents published before January 1, 2000 and a post-cut-off set composed of 3,961,789 documents published after January 1, 2000. For each of the 100 starting terms, we created one gold-standard set based on the post-cut-off set and used it in our comparisons. The sizes of the gold-standard sets were distributed widely (mean = 122.4, stddev = 94.13). This difference can be explained by the high variation in the number of documents that contain each starting term. For example, there were 1431 post-cut-off set documents that included the starting term *Hypercalcemia,* which resulted in 192 gold standard terms. In contrast, not many medical papers have been published about the starting term *Langer-Giedion Syndrome* in the last 96 months. There were only 33 post-cut-off set documents that included *Langer-Giedion Syndrome*, which resulted in only 15 gold standard terms.[2]

In our experiments, we were interested in finding novel connections between the 100 starting term disease names and *chemicals, drugs, genes, or molecular sequences.* As most LBD researchers have done, we used the UMLS **Semantic Network** to limit the results. In the UMLS, each medical term is categorized under at least one **semantic type** and each semantic type is categorized under one **semantic group** [33]. For example, the semantic type of *migraine* is *Disease and Syndrome* and the semantic type of *panic disorder* is *Mental or Behavioral Dysfunction*. The semantic group of both semantic types is *Disorders*. We gave LitLinker two sets of semantic groups as input: one for linking term selection and one for target term selection. LitLinker eliminated any terms that did not match the corresponding semantic group criteria. Table 2 includes the list of semantic groups that we used in our experiments.

As mentioned before, there is no easy recipe for selecting the thresholds of correlation-mining algorithms. For the experiments presented in this paper, we set the Z-score threshold to zero, which means that it would select a term as a linking/target term only if its probability in the starting/linking literature is greater than or equal to its mean probability in the background literatures. We also set the MIM threshold to zero to ensure the selected linking/target terms were statistically dependent on the starting/linking terms. For association rules, we followed what other LBD researchers did [11] and set the support and confidence thresholds to average

[1] http://www.nlm.nih.gov/bsd/licensee/2008_stats/baseline_doc.html.

[2] The post-cut-off set was created by querying MEDLINE 2008 Baseline for the 10 randomly selected starting terms and the time period, January 1, 2000–December 31, 2007.

Table 1
Excluded publication types.

Address	Directory	Letter
Bibliography	Editorial	News
Biography	Guidelines	Newspaper article
Comment	Lectures	Patient education handout
Congresses	Legal ceases	Periodical index
Dictionary	Legislation	Practical guideline

Table 2
Semantic groups selected for our experiments.

Linking Term Selection	Target Term Selection
Chemicals & Drugs	Chemicals & Drugs
Disorders	Genes & Molecular Sequence
Genes & Molecular Sequence	
Physiology	
Anatomy	

support (2.6) and average confidence (0.0055). Similarly, we set the TF-IDF threshold to average TF-IDF (142.45). The averages were taken over the all possible term couples available in the MEDLINE search space of the algorithms.

7.2. Comparison of ranking algorithms

In our evaluation, we first compared the performance of the ranking algorithms. Fig. 4 includes 11-point average interpolated precision graphs generated by the ranking algorithms for each correlation-mining algorithm. As can be seen from the graphs, for all the correlation-mining algorithms, LTC-AMW performed the best and COH performed the worst.

We also calculated the MAP values generated by three different ranking algorithms for each correlation-mining algorithm. The MAP values presented in Fig. 5 were averaged over the 100 randomly selected starting terms. As can be seen from the bar graph, for all correlation-mining algorithms, LTC-AMW produced the highest MAP value and COH produced the lowest MAP value.

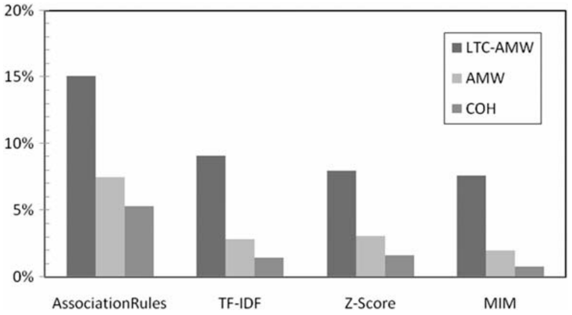

Fig. 5. *MAP* versus *correlation-mining algorithm* bar graph. LTC-AMW produced the best MAP values.

In addition to 11-point average interpolated precision and MAP, we compared the performance of ranking algorithms by calculating the precision levels for the top-ranked target terms. From the lists of target terms generated by the correlation-mining algorithms for each starting term, we selected the top 100 target terms ranked by LTC-AMW, AMW, and COH. We examined those lists of target terms starting from the top and selected 10 intervals (10–100 in increments of 10) to calculate precision (*precision at* $k \times 10$ where $1 \leqslant k \leqslant 10$). Because we have 100 different starting terms, we calculated the average precision for each interval to combine the results from each experiment. Fig. 6 includes the graphs that represent the change in precision as the target term rank interval increases. For all the correlation-mining algorithms, LTC-AMW identified more gold standard terms than either AMW or COH did and produced the best precision values.

The data presented in Figs. 4–6 was the precision averaged over the 100 starting terms. When we investigated the individual cases without averaging, we again found that LTC-AMW performed the

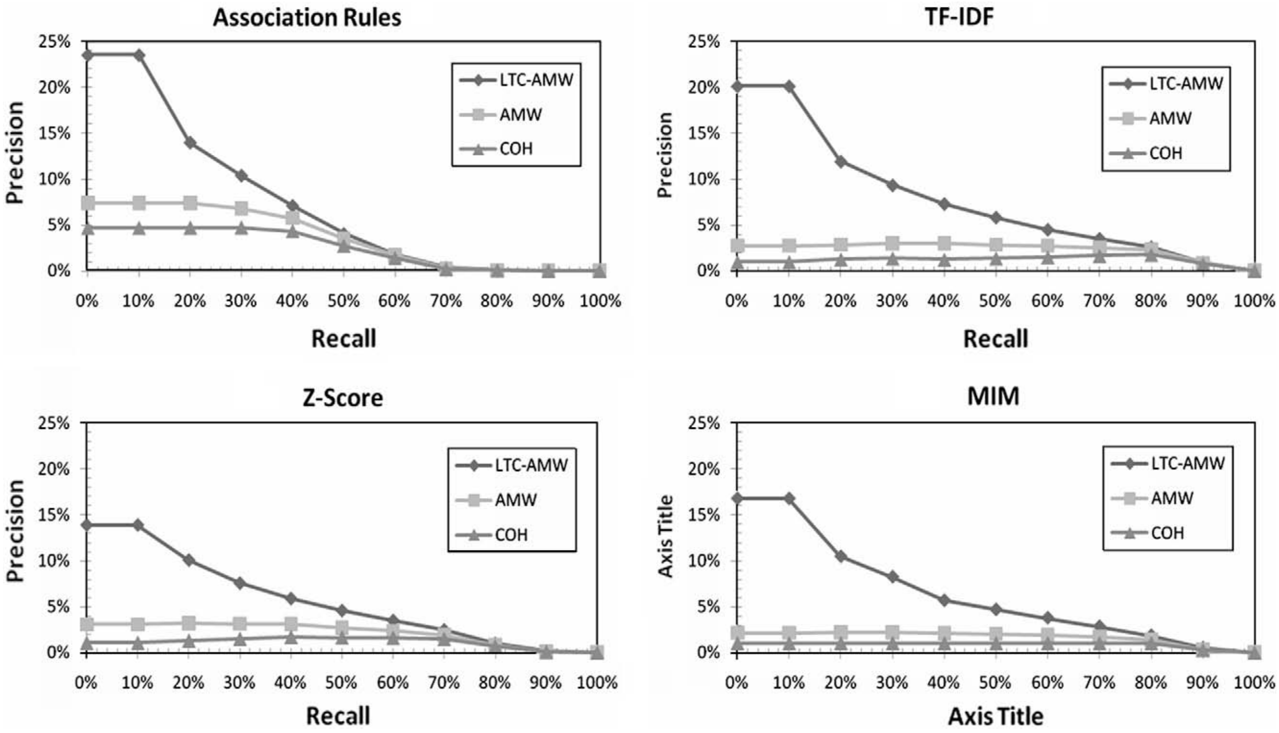

Fig. 4. Eleven-point average interpolated precision graphs. LTC-AMW produces the best and COH produces the worst precision values for all correlation-mining algorithms.

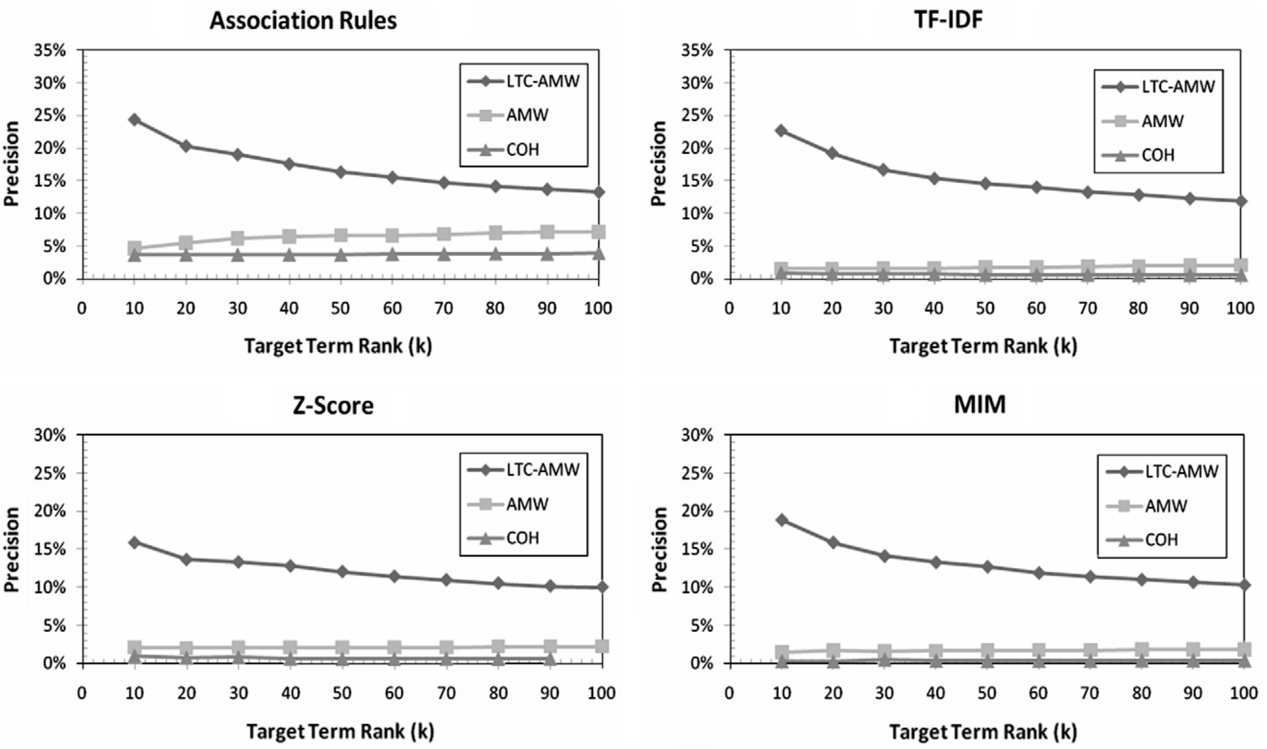

Fig. 6. Precision versus target term rank (k) graphs. LTC-AMW produces the best and COH produces the worst precision@k results for all correlation-mining algorithms.

best and COH performed the worst for all starting terms and all correlation-mining algorithms. These results provide convincing evidence, for a variety of starting terms, that LTC-AMW provides far superior results compared with either of the other two common algorithms—COH, and AMW. A possible explanation of this performance difference is that medical researchers are more likely to notice a connection between two disconnected terms when the terms are indirectly connected by many linking terms in the literature. Thus, LTC-AMW consistently did a better job in identifying the gold standard terms.

7.2.1. Comparison of correlation-mining algorithms

After comparing the performance of the ranking algorithms, we compared the performance of the correlation-mining algorithms using only the best ranking approach—LTC-AMW. Table 3 includes overall statistical information about the number of linking and target terms generated for each starting term by the correlation-mining algorithms and the number of gold standard terms identified as target terms (true positives). Table 3 also includes overall precision, recall, and MAP values averaged over the 100 starting terms.

On the average, association rules produced the lowest and MIM and TF-IDF produced the highest number of target terms. As an expected outcome of this observation, on the average, MIM and TF-IDF identified the highest and association rules identified the lowest number of gold standard terms as target terms.

Table 3
Summary of overall performance results. The values presented in the table were averaged over the 100 selected target terms. The highest precision@k values are bolded.

Correlation-mining algorithm	Linking Term Count	Target Term Count	True positives	Precision (%)	Recall (%)	MAP (%)
Association rules	97.42	871.73	48.93	**8.8**	53.76	**20.15**
TF-IDF	57.93	4479.34	**87.68**	2.29	**88.07**	10.21
Z-score	231.77	3270.75	71.66	2.88	75.58	11.83
MIM	**933.37**	**7718.49**	87.24	1.38	85.51	9.21

Consistent with the previous discussion, association rules produced the highest overall precision value and lowest overall recall value. MIM and TF-IDF produced higher recall values but lower precision values when compared to the other two algorithms. In terms of MAP, association rules performed the best and MIM performed the worst.

The number of target terms generated by a correlation-mining algorithm is directly determined by the thresholds used to distinguish the interesting correlations from the set of all possible correlations. Because there is not a standard way of selecting the thresholds for different algorithms, using overall precision, recall, or MAP can be misleading metrics when comparing the performance of correlation-mining algorithms. To overcome this problem, we compared the algorithms' ability to identify gold standard terms at high ranks. Our assumption, based on many information retrieval studies [34,35], is that researchers will only investigate the top-ranked target terms. Thus, we believe that precision at the top ranks will provide a metric corresponding more closely to usefulness than overall precision or recall would. For each algorithm and starting term, we calculated *precision at k* ($k = 10 \times i$ where $1 \leqslant i \leqslant 5$). As can be seen from Table 4, association rules produced better overall precision than the other three algorithms produced.

Table 4
Overall precision@k results. Association rules produced the best and *Z*-score produced the worst precision@k results for $k = 10 \times i$ where $1 \leqslant i \leqslant 5$. ($p$:precision, e:standard error). The highest values are bolded.

Correlation-mining algorithm	Precision@*k*									
	k = 10		*k* = 20		*k* = 30		*k* = 40		*k* = 50	
	p (%)	*e* (%)	*p* (%)	*e* (%)	*p* (%)	*e* (%)	*p* (%)	*e* (%)	*p* (%)	*e* (%)
Assoc. rules	**24.4**	1.7	**20.35**	1.5	**19.03**	1.3	**17.6**	1.2	**16.36**	1.1
TF-IDF	22.7	1.8	19.25	1.5	16.7	1.2	15.35	1.1	14.58	1.1
Z-score	15.9	1.6	13.65	1.2	13.33	1.1	12.8	1	12.06	0.9
MIM	18.8	1.6	15.85	1.3	14.07	1.1	13.25	1.1	12.68	0.9

For each algorithm and each starting term, we also calculated the MAP values for the top-ranked *k* gold standard terms identified as target terms by the algorithm for the starting term. We could have calculated MAP values for the top-ranked *k* target terms but such an approach has a major drawback. Suppose we are comparing two algorithms *A* and *B* by calculating the MAP values for the top-ranked 100 target terms. There is only one gold standard term in *A*'s top 10 target term list, and that term is ranked 1st. On the other hand, all the terms in *B*'s top 10 target term list are gold standard terms. For both *A* and *B*, the MAP value for the top 10 target terms is equal to 100% even though *B*'s performance is clearly superior to *A*'s performance. Because of this drawback, we decided to calculate MAP values for the top-ranked *k* gold standard terms (MAP@k). This approach allowed us to quantify the correlation-mining algorithms' ability to identify a certain number of gold standard terms at higher ranks. To summarize the results for each algorithm, we averaged the MAP@k values over the set of starting terms. Table 5 includes MAP@k values for $k = 10 \times i$ where $1 \leqslant i \leqslant 5$. As can be seen from the table, association rules performed better than the other three algorithms in identifying gold standard terms at higher ranks. In contrast, *Z*-score performed the worst. To understand the significance of our results, we ran Student *t*-tests (2-tailed, $\alpha = 0.05$) and found that the performance difference between association rules and the other three algorithms is statistically significant.

8. Limitations

Evaluating LBD algorithms or systems remains challenging, particularly because even the definition of a "discovery" can be controversial [36,37]. In this paper, we assume that a connection between two terms was undiscovered if the terms had not co-occurred in the MEDLINE literature. Correspondingly, we used an objective definition of discovery as taking place when two terms co-occur in the literature. Although others have used subjective measures based on expert opinion, evaluating the potential discoveries corresponding to even one starting term would require experts to assess thousands of potential discoveries [19,26]. Thus, the expert-opinion based approach is clearly infeasible for any large-scale evaluations, such as our evaluation of 100 starting terms using several different algorithms. In addition, for all our evaluations, we represented the content of the MEDLINE documents with MeSH terms. The main problem with this type of representation is the content coverage. In the MEDLINE 2006 baseline, each document averaged 12 descriptive MeSH terms. These tags can capture the main topics, but clearly do not represent all terms used in a document. Although the researchers who proposed the evaluated LBD algorithms used different approaches to represent the content of documents in their LBD systems, to make a fair comparison among the algorithms, we needed to use the same text representation approach for all algorithms.

Table 5
Overall *MAP*@k results. Association rules produced the best and *Z*-score produced the worst MAP@k results for $k = 10 \times i$ where $1 \leqslant i \leqslant 5$ (*m*: *MAP*, *e*: standard error). The highest MAP@k results are bolded.

Correlation-mining algorithm	MAP@k									
	k = 10		*k* = 20		*k* = 30		*k* = 40		*k* = 50	
	m (%)	*e* (%)	*m* (%)	*e* (%)	*m* (%)	*e* (%)	*m* (%)	*e* (%)	*m* (%)	*e* (%)
Assoc. Rules	**27.31**	1.92	**21.72**	1.62	**18.95**	1.42	**17.67**	1.29	**16.86**	1.21
TF-IDF	23.74	1.84	18.23	1.47	15.53	1.3	13.83	1.18	12.71	1.09
Z-score	18.08	1.68	14.11	1.34	12.22	1.17	11.1	1.07	10.41	0.99
MIM	20.64	1.79	15.66	1.43	13.36	1.26	11.96	1.15	11.02	1.07

There are trade-offs with using either the MeSH term representation or one based on natural language processing (NLP) methods. In our previous research, we used an approach based on NLP to represent the content of the MEDLINE documents [14]. Although using an NLP approach produced more index terms, the indexing process was computationally expensive. In addition, that approach only indexes the title and the abstract (if available). In contrast, the MeSH terms are selected by the human indexers based on the full text of the documents. Thus, for all our experiments, we decided to represent the documents using MeSH terms. Future researchers could use our approach to evaluate the effectiveness of their algorithms with alternative representation approaches.

9. Conclusion

The research described in this paper provides two main contributions. The first contribution is our evaluation methodology for LBD systems that allows comparisons across different systems. The methodology is fully automated and can be repeated multiple times. The second contribution is that we demonstrated the abilities of our evaluation methodology by using it to compare the performance of four different correlation-mining algorithms and three ranking algorithms that are used by the existing LBD systems. In our experiments, we found that LTC-AMW performed consistently better than AMW and COH in ranking the target terms. We also found that, overall, association rules performed the best in terms of precision and MAP; whereas, MIM and TF-IDF performed better in terms of recall when compared to the other correlation-mining algorithms. Because overall precision, recall, and MAP values were highly dependent on the selection of the thresholds of the correlation-mining algorithms, we also compared the performance of correlation-mining algorithms based on their ability to identify gold standard terms at higher ranks. Based on average precision at k and MAP@k, Association Rules performed the best in identifying gold standard terms at higher ranks.

To our knowledge, this paper is the first LBD evaluation methodology paper that reports performance comparisons across many different algorithmic approaches. In addition, the methodology that we described will allow LBD researchers to conduct further studies and to evaluate new potential algorithms both for finding correlations in the literature and for ranking proposed target terms. Thus, the methods and results from this paper should help advance and improve LBD systems, and ultimately lead to improved use of the literature in the research discovery process.

Acknowledgment

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

References

[1] Swanson DR. Two medical literatures that are logically but not bibliographically connected. J. Am. Soc. Inf. Sci. 1987;38:228–33.
[2] Swanson DR. Migraine and magnesium: eleven neglected connections. Perspect Biol Med 1988;31:526–57.
[3] Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 1986;30:7–15.
[4] Smalheiser NR, Swanson D. Linking estrogen to Alzheimer's disease: an informatics approach. Neurology 1996;47:809–10.
[5] Smalheiser NR, Swanson D. Indomethacin and Alzheimer's disease. Neurology 1996;46:583.
[6] Swanson DR. Somatomedin C and arginine: implicit connections between mutually isolated literatures. Perspect Biol Med 1990;33:157–86.
[7] Smalheiser NR, Swanson D. Calcium-independent phospholipase A2 and schizophrenia. Arch Gen Psychiatry 1998;55:752–3.
[8] Blake C, Pratt W. Automatically identifying candidate treatments from existing medical literature. In: Proceedings of AAAI spring symposium on mining answers from texts and knowledge bases, California; 2002.

[9] Gordon MD, Dumais S. Using latent semantic indexing for literature based discovery. J Am Soc Inf Sci 1998;49:674–85.

[10] Gordon MD, Lindsay RK. Toward discovery support systems: a replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. J Am Soc Inf Sci 1996;47:116–27.

[11] Hristovski D, Stare J, Peterlin B, Dzeroski S. Supporting discovery in medicine by association rule mining in Medline and UMLS. In: Proceedings of Medinfo; 2001.

[12] Hu X, Li G, Yoo I, Zhang X, Xu X. A semantic-based approach for mining undiscovered public knowledge from biomedical knowledge. In: Proceedings of IEEE international conference on granular computing, Beijing; 2005. p. 22–8.

[13] Lindsay RK, Gordon MD. Literature based discovery by lexical statistics. J Am Soc Inf Sci 1999;49:674–85.

[14] Pratt W, Yetisgen-Yildiz M. LitLinker: capturing connections across the biomedical literature. In: Proceedings of the international conference on knowledge capture (K-Cap'03); 2003. p. 105–12.

[15] Smalheiser NR, Torvik VI, Bischoff-Grethe A, Burhans LB, Gabriel M, Homayouni R. Collaborative development of the Arrowsmith two node search interface designed for laboratory investigators. J Biomed Discov Collaboration 2006;1.

[16] Srinivasan P. Generating hypotheses from MEDLINE. J Am Soc Inf Sci 2004;55:396–413.

[17] 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 2001;52:548–57.

[18] Weeber M, Kors JA, Mons B. Online tools to support literature-based discovery in the life sciences. Brief Bioinform 2005;6:277–86.

[19] Weeber M, Vos R, Klein H, de Jong-van den Berg LTW, Aronson AR. 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 Inform Assoc 2003;10:252–9.

[20] Wren JD. Extending the mutual information measure to rank inferred literature relationship. BMC Bioinform 2004;5:145.

[21] Yetisgen-Yildiz M, Pratt W. Using statistical and knowledge-based approaches for literature based discovery. J Biomed Inform 2006;39:600–11.

[22] Bekhuis T. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. Biomed Digit Libr 2006;3(2).

[23] Swanson DR, Smalheiser NR, Torvik VI. Ranking indirect connections in literature-based discovery: the role of medical subject headings. J Am Med Inform Assoc 2006;57:1427–39.

[24] NLM, MeSH Fact Sheet; 2005. Available from: http://www.nlm.nih.gov/pubs/factsheets/mesh.html.

[25] Srinivasan P, Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics 2004;20:I290–6.

[26] Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR. Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics 2004;20:389–98.

[27] Torvik VI, Smalheiser NR. A quantitative model for linking two disparate sets of articles in Medline. Bioinformatics 2007;23:1658–65.

[28] Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting semantic relations for literature-based discovery. In: Proceedings of the American medical informatics association fall symposium (AMIA'06), Washington, DC; 2006.

[29] Yetisgen-Yildiz M, Pratt W. Evaluation of literature-based discovery systems. In: Bruza PD, Weeber M, editors. Literature-based discovery. Springer; 2008.

[30] Voorhees EM, Harman DK, editors. TREC: experiment and evaluation in information retrieval. MIT Press; 2006.

[31] Agrawal R, Imielinski T, Swami A. Mining associations between sets of items in massive databases. In: Proceedings of the ACM-SIGMOD 1993 international conference on management of data, Washington, DC; 1993.

[32] Spark Jones K. A statistical interpretation of term specificity and its application in retrieval. J Doc 1972;28:111–21.

[33] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. In: Proceedings of Medinfo, San Francisco; 2001.

[34] Jansen BJ, Spink A. An analysis of web documents retrieved and viewed. In: Proceedings of the international conference on internet computing, Las Vegas, NV; 2003. p. 65–9.

[35] Joachims T, Granka L, Pan B, Hembrooke H, Gay G. Accurately interpreting click through data as implicit feedback. In: Proceedings of SIGIR, Salvador, Brazil; 2005. p. 154–61.

[36] Kostoff RN. Validating discovery in literature-based discovery. J Biomed Inform 2007;40(4):448–50.

[37] Pratt W, Yetisgen-Yildiz M. Response to 'Validating discovery in literature-based discovery'. J Biomed Inform 2007;40(4):450–2.