Evaluation of Literature-based Discovery Systems

Meliha Yetisgen-Yildiz[1] and Wanda Pratt[1,2]

[1] The Information School, University of Washington, Seattle, USA.
[2] Biomedical and Health Informatics, School of Medicine, University of Washington, Seattle, USA.
{melihay,wpratt}@u.washington.edu

Abstract. Evaluating discovery systems is a fundamentally challenging task because if they are successful, by definition they are capturing new knowledge that has yet to be proven useful. To overcome this difficulty, many researchers in literature-based discovery (LBD) replicated Swanson's discoveries to evaluate the performance of their systems. They reported overall success if one of the discoveries generated by their system was the same as Swanson's discovery. This type of evaluation is powerful yet incomplete because it does not inform us about the quality of the rest of the discoveries identified by the system nor does it test the generalizability of the results. Recently, alternative evaluation methods have been designed to provide more information on the overall performance of the systems. The purpose of this chapter is to review and analyze the current evaluation methods for LBD systems and to discuss potential ways to use these evaluation methods for comparing performance of different systems, rather than reporting the performance of only one system. We will also summarize the current approaches used to evaluate the graphical user interfaces of LBD systems.

1 Introduction

Evaluation plays an important role in the development of new fields such as literature-based discovery (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. In this chapter, we will give an overview of the current state of evaluation in literature-based discovery research and discuss potential ways for future evaluations.

2 Evaluation Metrics

When developing an LBD system, it is critical to know how reliable the results are likely to be. Measuring the reliability of a prediction system requires two main components: a gold standard and an evaluation metric to measure the system’s performance with respect to the gold standard. For a given starting term, which

Swanson called *C-Term*, a typical LBD system produces two sets of terms; *linking terms* and *target terms*. The linking terms, which Swanson called *B-Terms*, directly connect a given starting term to the target terms, which Swanson called *A-Terms*. The gold standards used to evaluate those two sets of terms are different from each other, and the gold standard creation methods depend on which of the evaluation methods listed in Section 3 is used. We will describe how the gold standards for linking/target terms are created for certain types of evaluation methods in Section 3. For now, we will define the gold standards for linking/target terms as the two sets of terms that are known to be directly/indirectly connected to a given starting term. In this section, we will summarize the metrics used to measure the performance of LBD systems.

2.1 Information Retrieval Metrics

The main purpose of evaluation in information retrieval research (IR) is to measure IR systems' performance in returning the relevant documents and in not returning the non-relevant documents to user queries. In IR evaluation, the gold standard is the set of relevant documents and two most popular IR metrics used to measure system performance are *precision* and *recall* [1]. For a given query and an IR system, precision can be defined as the proportion of relevant documents in the set of documents returned by the system and recall can be defined as the proportion of the relevant documents retrieved by the system from the gold standard.

In contrast to IR systems, LBD systems return terms instead of documents. Thus, precision and recall are mainly used to measure the effectiveness of an LBD system in returning linking and target terms for a given starting term, rather than the effectiveness of an IR system in returning documents for a given query. Precision and recall for the LBD system evaluation are calculated with the following formulas:

$$\text{Precision:} \qquad P_i = \frac{\|T_i \cap G_i\|}{\|T_i\|} \qquad (1)$$

$$\text{Recall:} \qquad R_i = \frac{\|T_i \cap G_i\|}{\|G_i\|} \qquad (2)$$

where T_i is the set of linking/target terms generated by the LBD system for the starting term *i*, and G_i is the set of terms in the linking/target term gold standard that the LBD system created for the starting term *i*.

As with IR system evaluation, one challenge in interpreting precision and recall is that there is a trade-off between the two metrics. Usually a system that aims to achieve high precision will result in low recall and vice versa. To solve this problem, some information retrieval researchers invented a new measure called *F-Measure* which is a combined version of precision and recall. F-Measure is calculated with the following formula:

$$\text{F-Measure:} \qquad F = \frac{(1+\beta^2)\times R \times P}{(\beta^2 \times P) + R} \qquad (3)$$

where *R* is the recall, *P* is the precision, and β is the relative value of the precision. The most commonly used case $\beta = 1$ assigns equal emphasis on precision and recall, whereas a lower value assigns a higher emphasis on precision and a higher value assigns a higher emphasis on recall.

Another common method to combine precision and recall is to draw a precision-recall curve. In this curve, the x-axis corresponds to recall and the y-axis corresponds to precision. Because of the trade-off between precision and recall, precision-recall graphs usually have a concave shape. Trying to increase recall typically introduces more false positives (target terms that are not in the gold standard), and thereby reduces precision. Trying to increase precision typically reduces recall by decreasing the number of true positives (target terms that are in the gold standard). An ideal goal of a prediction system is to increase both precision and recall by making improvements to the system. In other words, the entire curve must move up and out to the right so that both recall and precision are higher at every point along the curve. The most common use of precision-recall curves is for system comparisons.

2.2 Receiver Operating Characteristics (ROC) Curve

Receiver Operating Characteristics (ROC) curve provides a graphical representation of the relationship between the true positive and false positive rate of a prediction system [2]. These curves are used frequently in comparing the effectiveness of different medical diagnostic tests. The y-axis corresponds to the *sensitivity* of the system. Sensitivity measures the performance of the system in predicting the true positives. The x-axis corresponds to the *specificity* (expressed as *1-specificity* in the graph). Specificity represents the ability of the system in identifying true negatives. The sensitivity and the specificity of a LBD system can be calculated as:

$$\text{Sensitivity:} \qquad Y_i = \frac{TP_i}{TP_i + FN_i} \tag{4}$$

$$\text{1- Specificity:} \qquad X_i = 1 - \frac{TN_i}{TN_i + FP_i} \tag{5}$$

where for the starting term *i*; TP_i is the number of true positives (the target terms that are in the gold standard), FN_i is the number of false negatives (the gold standard terms that are not identified as target terms), FP_i is the number of false positives (the target terms that are not in the gold standard), and TN_i is the number of true negatives (the terms that are both not selected as target terms and not in the gold standard).

The ROC curves show the performance as a trade off between specificity and sensitivity of the prediction system. The area under the ROC is a convenient way of comparing different prediction systems. A random system has an area of 0.5, while and ideal one has an area of 1.

2.3 Probabilistic Approaches

Because the purpose of LBD systems is to predict novel connections between medical terms, it is also important to compare their prediction performance with that of pure random prediction. One way to accomplish this objective is to calculate the probability of randomly achieving the performance of a given LBD system. This probability can be modeled with *hypergeometric distribution*. Suppose for a given starting term, an LBD system returns *k* target terms where *i* of the target terms that are in the gold standard, there are *n* terms in the gold standard and there are *m* terms in the search space of the system. The probability of having *i* gold standard terms in randomly selected *k* target terms is calculated with the following formula:

$$p(x=i)=\frac{\binom{n}{i}\binom{m-n}{k-i}}{\binom{m}{k}} \qquad (6)$$

If the value of *p* is close to zero, achieving the performance of the LBD system by randomly selecting the target terms is highly unlikely. If the value of p is close to 1, the prediction of mechanism of the LBD system needs to be improved because random selection of the terms gives almost the same performance.

3 Current Evaluation Approaches

Evaluating the performance of LBD systems is a fundamentally challenging task because if these systems are successful, by definition, they are capturing new knowledge that has yet to be proven useful. After a detailed analysis of the existing literature on LBD systems, we identified the following four different approaches used to evaluate LBD systems; replicating Swanson's discoveries, using statistical evaluation approaches, incorporating expert knowledge, and publishing in the medical domain. In this section, we will explain each evaluation approach in detail and discuss their advantages and disadvantages.

3.1 Replicating Swanson's Experiments

Even though the LBD systems are designed to produce new knowledge, measuring their performance by replicating the historical discoveries has been seen an effective evaluation approach by many LBD researchers. Swanson and Smalheiser published several different hypotheses about causally connected medical terms in the biomedical domain including *Migraine – Magnesium* [3], *Raynaud's Disease - Fish Oil* [4], *Alzheimer's Disease – Estrogen* [5], *Alzheimer's Disease – Indomethacin* [6], *Somatomedin C – Arginine* [7], and *Schizophrenia – Calcium Independent Phospholipase* A_2 [8]. Their discoveries have become gold standards for evaluation, and LBD researchers have measured the performance of their discovery systems by replicating Swanson's discoveries using the literature published before the original discovery dates. They have run their systems with Swanson's starting terms on the literature published prior to the discovery dates and reported overall success if one of the correlations generated by their systems matched Swanson's discovery.

Several researchers have used this strategy to evaluate the linking terms generated by their systems. Lindsay and Gordon [9] developed a process that followed the Swanson's discovery approach. They evaluated the performance of their process, in terms of precision and recall, for generating the linking terms, where Swanson's identified linking terms for *Migraine-Magnesium* example served as the gold standard. Gordon and Dumais applied latent semantic indexing to Swanson's discovery process [10]. They demonstrated the performance of their approach by replicating Swanson's *Raynaud's Disease* and *Fish Oil* discovery. Blake and Pratt applied a knowledge-based approach to identify and prune potential linking terms [11]. They replicated Swanson's *Migraine-Magnesium* example to evaluate their approach. However, all of these researchers focused on evaluating the linking terms by using Swanson's linking terms as the gold standard, and none pursued or evaluated how easy it would be identify the novel target term (e.g., *magnesium*), which is the main goal of LBD systems..

Weeber et. al. also based their work on Swanson's approach [12]. They evaluated their literature-based discovery tool DAD by simulating Swanson's *Raynaud's Disease-Fish Oil* and *Migraine-Magnesium* examples. Their system supported both open and closed discovery approaches. In the open discovery approach, DAD first identified the linking terms that are directly connected to the starting terms, *Raynaud's Disease* and *Migraine*, and then identified the target terms that are connected to the linking terms identified in the first step. They reported which of the Swanson's linking terms DAD could identify and the ranks of *Fish Oil* and *Magnesium* in the final lists of target terms. In the closed discovery approach, they analyzed the starting term literature and the target term literature separately and identified the overlapping terms. They compared those terms with Swanson's linking terms and reported the results.

The most extensive evaluation of this type was done by Srinivasan [13]. She developed a literature based discovery system called Manjal. As Weeber et. al.'s system, Manjal supports both open and closed discovery approaches. To evaluate her system, Srinivasan successfully replicated five of Swanson's discoveries including *Raynaud's Disease-Fish Oil, Migraine-Magnesium, Alzheimer's Disease-Indomethacin, Somatomedin C-Arginine,* and *Schizophrenia-Calcium Independent Phospholipase A2*. For each discovery, she reported the rank of the desired target term in the list of target terms generated by Manjal with the open discovery approach. She also reported the ranks of the desired linking terms identified by Manjal with the closed discovery approach.

Most recently, Hu et. al. developed a prototype system called Bio-SbKDS based on Swanson's discovery approach [14]. They replicated Swanson's *Migraine-Magnesium* and *Raynaud Disease-Fish Oil* discoveries for evaluation purposes. He used *Migraine* and *Raynaud's Disease* as starting terms. They reported which of Swanson's linking terms their system could identify as linking terms and the ranks of *Magnesium* and *Fish Oil* in the final lists of target terms generated by their system.

In previous research, we also replicated Swanson's *Migraine-Magnesium* discovery to evaluate the capabilities of our system LitLinker [15]. As other researchers, we compared our linking terms with Swanson's linking terms and reported the rank of *Magnesium* in the final list of target terms.

The main advantage of this type of evaluation is the ease of designing it. In his papers, Swanson described each of his discoveries in great detail. The researchers use the information provided in those papers as a guide in designing their evaluations. For each discovery, the publication date of the corresponding paper serves as the original discovery date and the list of medical terms he used as links between his starting term and target term serves as a linking term gold standard.

Although all the researchers mentioned in this section have successfully replicated Swanson's discoveries, this type of evaluation is not complete because it does not inform us about the quality of the rest of the target terms identified by their systems. Depending on the approaches used to select the correlated terms, a literature-based discovery system might return hundreds or even thousands of terms as the 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 also provide information with similar quality. 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.

Another disadvantage of this approach is that the researchers are limited in their evaluations to the small number of discoveries published by Swanson. His discoveries mostly focused on diseases and their potential new treatments. Nevertheless, LBD tools can be used for various other tasks, such as identifying novel protein-protein interactions. Because the researchers know exactly what they are seeking as the desired target and linking terms in this limited set of discoveries, they can tune the parameters of their systems to be able to identify those terms. Such an approach might result in systems that perform well for the specific example cases but not well for other cases.

In addition, comparing the performance of different systems is one of the main objectives of system evaluation. However, replicating Swanson's discoveries does not allow detailed comparisons between different LBD systems. This evaluation method allows the researchers to say a system *A* is better than another system *B* if *A* simulates a selected discovery but *B* does not. However, if both *A* and *B* successfully simulate the given discovery successfully, it becomes impossible to determine which system is superior to the other.

3.2 Using Statistical Evaluation Methods

To overcome the drawbacks of the previous approach, some researchers have applied statistical evaluation methods to measure the overall performance of literature-based discovery systems for multiple target terms. As an example, Hristovski et.al. performed a statistical evaluation of their system, BITOLA [16]. The purpose of their evaluation was to see how many of the potential discoveries made by their system at a specified point in time become realized at a later time. To accomplish this goal, they ran their system for the starting term *Multiple Seclerosis* 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 the medical literature and reported the performance statistics of their system without comparing it to those of other systems.

To evaluate our system LitLinker, we used a similar but more extensive approach than Hristovski et.al.'s approach; this approach enabled us to evaluate all correlations LitLinker generated. In our evaluation, for a given starting term, 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 sets: (1) a baseline set including only publications before a selected cut-off date, and (2) a test set including only publications between the cut-off date and another later date. We ran LitLinker on the baseline set and checked the generated connections in the test set.

As an evaluation example, in [17], we ran LitLinker for the starting terms; *Alzheimer Disease*, *Migraine*, and *Schizophrenia* on a baseline set, which included only documents published before January 1, 2004 (cut-off date). We limited the linking terms and the target terms to only those terms in a semantic group 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 test set that was composed of articles published between January 1, 2004 and September 30, 2005 (21 months).

Table 1. 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	

To calculate precision and recall, for each starting term, we first retrieved the terms that co-occurred with the starting term in the test set but did not co-occur with the starting term in the baseline set. Then, we filtered the retrieved list of 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 terms in the remaining list would be new potential disease to gene or disease to drug treatment discoveries and used them as the target term gold standard for our precision and recall calculations.

In our current research, we used our evaluation approach to compare two different methods for identifying linking or target terms based on a starting term, Z-Score [17] and MIM [18]. To accomplish this task, we first implemented the methods within our LitLinker framework. In our experiments, for each method, we ran LitLinker for 10 randomly selected disease names on a baseline set, which includes only documents published before January 1, 2004. We created a target term gold standard for each disease from the test set documents published between January 1, 2004 and July, 31, 2006 (31 months).

We calculated precision and recall of both methods for each disease and ran statistical significance tests to measure the significance of the performance differences. We also used precision-recall graphs to compare different correlation methods. To draw precision-recall graphs, we used the ranked list of target terms generated by the two methods. We examined these lists of target terms starting from the top and selected intervals to calculate precision and recall with the formulas (1) and (2). Because we had 10 different starting terms, to combine the results from each experiment, we calculated the average precision and recall for each interval. We also compared the prediction performances of both methods with that of pure random prediction with hypergeometric distribution as described in Section 2.3.

The main advantages of this type of evaluation are that the evaluation is fully automated, can be repeated for multiple starting terms, and enables comparison among different systems. On the other hand, its main drawback is that the calculated precision for target terms is the lower bound. The target term gold standard only includes the new correlations that are published between the cut-off date and the date of the experiment. It cannot include the correlations that will appear in the future. As a result, some of the target terms identified by the LBD system might become legitimate discoveries in the future but are considered incorrect target terms now. Another disadvantage is that this approach only evaluates the target terms without providing any information about the linking terms.

3.3 Incorporating Expert Opinion

As an alternative to the previous approaches, some researchers incorporated medical expert knowledge to 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 on 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. In the assessment process, they tried to find additional bibliographic and other evidence for the linking terms between the *thalidomide* and the diseases identified as target terms. To accomplish this goal, for each disease, they first extracted the list of linking terms that connect the disease to *thalidomide*. Next, they extracted the sentences that included *thalidomide* and the extracted linking terms and the sentences that included the linking terms and the disease. They provided those sentences to the medical expert for assessment. Based on the assessment, they compiled a list of four diseases; c*hronic hepatitis C*, *myasthenia gravis*, *helicobacter pylori induced gastritis*, *acute pancreatitis* for which the researchers hypothesized that *thalidomide* could be an effective treatment.

Srinivasan and Libbus evaluated their system Manjal by using a semi-automated approach with experts. In their experiment, they used *turmeric,* a widely used spice in Asia, as their starting term. The aim of their experiment was to identify diseases where *turmeric* could be useful in the treating them. They ran Manjal for the starting term *turmeric,* and, with the selected thresholds, Manjal identified 26 terms as the linking terms, L_1. To evaluate the linking terms in L_1, a medical researcher identified a second set of linking terms, L_2, after reading the documents about *turmeric*. There were 27 terms in L_2. They used this manually created list as the linking term gold standard. They compared L_1 with L_2 and calculated recall and precision with the following formulas:

$$\text{Precision:} \qquad P = \frac{\|L_1 \cap L_2\|}{\|L_1\|} \qquad (7)$$

$$\text{Recall:} \qquad R = \frac{\|L_1 \cap L_2\|}{\|L_2\|} \qquad (8)$$

Manjal generated two sets of target terms; one from the automatically generated linking terms and one from the manually selected linking terms. They used the second set as the target term gold standard to evaluate the first set and reported precision and recall. In addition to reporting precision and recall, they did a detailed citation analysis and described the potential use of *turmeric* in the treatment of *retinal diseases*, *Crohn's disease*, and *spinal cord injuries*. In contrast to the statistical approach described in the previous section, the advantage of Srinivasan and Libbus's approach is that it allows us to evaluate the linking terms in addition to the target terms. However, the evaluation highly depends on the subjective decision of the medical researcher in deciding which terms are correlated with the starting term. This decision is crucial because it also directly effects the selection of the terms in the target term gold standard. It is also unclear whether the gold standard set of target terms reflects a true gold standard because no checking has been done on those target terms.

Wren et.al. also incorporated medical expert knowledge into the evaluation process [20]. The researchers who contributed to this study had a medical background. They ran their literature-based discovery approach for the starting term *cardiac hypertrophy* and identified a total of 2102 linking terms and 19718 target terms. To evaluate their approach, they performed laboratory tests for 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 an association between *chlorpromazine* and *cardiac hypertrophy*. They gave 20mg/kg/day *isoproterenol* by osmotic minipump to two groups of mice, with one group additionally receiving 10mg/kg/day *chlorpromazine*. Their results showed that the amount of *cardiac hypertrophy* was significantly reduced in the *isoproterenol* plus *chlorpromazine* treated mice in comparison to the control group only given *isoproterenol*. They reported that *chlorpromazine* could reduce *cardiac hypertrophy* by showing their experimental results with mice as evidence. Their work is an excellent example of how literature-based discovery tools can be integrated to medical researcher's real-life research activities.

The main advantage of this type of evaluation is the involvement of the medical researchers, who are the real users of the LBD systems into the evaluation process. To identify what medical researchers find interesting or not interesting could inform LBD system designers while they upgrade the algorithms or the other approaches they use in the discovery process. The downside is the high cost of evaluation. Weeber et. al. reported that their manual effort while evaluating the output of their system consisted of several one hour sessions during a two-week period. Such an evaluation is also hard to quantify, and thus hard to use to compare different LBD systems. 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.

3.4 Publishing in the Medical Domain

Another approach that is used to evaluate LBD systems is publishing the discoveries in medical journals or presenting them in the medical domain. This evaluation approach is a very powerful yet a very challenging one. Publishing in the medical domain requires the flexibility to write for the medical audience, but the overall benefit is clear: validation of work, impact on the science, external visibility for LBD research, and the chance to gain new collaborators. This type of evaluation is not commonly used in LBD research. Among all LBD researchers, Swanson is the only researcher who could publish his discoveries in the medical journals. In addition to Swanson's personal interest in medicine, his close collaboration with Smalheiser who is a medical doctor and neuroscientist, resulted in various publications [3-8, 21].

4 User Interface Evaluation

The success of an LBD system in facilitating new discoveries depends on its interface's ability to inform and engage its users as they attempt to interpret and evaluate the proposed connections. The amount of data produced by an LBD system is usually immense. As an example, when LitLinker replicated Swanson's *Migraine-Magnesium* discovery, it processed over 4 million documents. It generated 349 linking terms and 545 target terms with 57,622 possible starting term-linking term and linking term-target term combinations. To be able to handle the amount and complexity of the output data, one of the primary objectives of an LBD system interface must be to promote user comprehension of numerous complex relationships among the terms involved in each proposed connection in an effective way. The interface must also provide flexible navigation and a level of detail appropriate to the scope of each view without obscuring data necessarily for evaluation purposes. And most importantly, the interface should help researchers incorporate the LBD system's results into their own research discovery process. To accomplish those objectives requires the involvement of real users into the interface design process. One way to involve users is by conducting usability evaluations and changing the interface design according to the feedback collected from the participants of the evaluation.

We designed a web-based graphical interface for LitLinker[1]. Our aim in developing an interface was to allow researchers to carefully assess the potential connections generated by LitLinker. We first developed a prototype interface and conducted a usability evaluation with ten participants, including nine graduate students and one faculty member [22]. The evaluation consisted of three parts: a general introduction, a task-based questionnaire, and an interview. The participants used LitLinker with *Migraine* as the starting term, to complete a task-based questionnaire. The tasks were designed to evaluate each participant's ability to find specific data, to navigate the interface, and to compare the strengths of connections. Participants were asked to talk aloud and as they completed the tasks. The interviewer observed without answering questions and noted any difficulties the participants experienced. After participants completed the questionnaire, we interviewed them to discover aspects of the interface that were confusing or were particularly helpful. We identified many design problems during this usability evaluation and modified our interface to increase its usability.

Similarly, Smalheiser et. al. evaluated their LBD system, Arrowsmith as part of a five year neuroscience project at University of Illinois-Chicago [23]. The goal of their evaluation study included making scientific discoveries, publishing papers, and identifying new research directions. In contrast to our study, they did not recruit human subjects or study their behavior on standardized tasks. Rather, the medical researchers who participated in the study chose the search topics and observed the outcomes. Each participant was given an electronic notebook to record opportunities for conducting Arrowsmith searches, whether they arose from laboratory experiments, from attending conferences, or from discussions with other researchers, and to record the details of completed Arrowsmith searches. Participants sent the notebook entries via e-mail to the researchers and the researchers called the participants every week to monitor the course of their scientific work, to learn more about the completed searches, to receive suggestions for improving the interface, and to document the follow-up of completed searches. Based on the input they received from the participants, they updated the Arrowsmith interface. They also focused on information seeking needs and strategies of medical researchers as they formulate new hypotheses.

5 Conclusion

LBD systems have great promise for improving medical researchers' efficiency while they seek information in the vast amount of literature available to them. Although many online LBD systems are available, they are not in routine use. For a wider usage of LBD systems, effective evaluation is essential. Evaluation will not only help to identify which algorithmic approaches work best for LBD, but also provide

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

information about how discovery systems can best enhance the real-life work processes of medical researchers. In this chapter, we summarized the current evaluation approaches used to evaluate LBD systems and their interfaces, but more research on evaluation methods that standardize system comparisons and explore user behavior is needed.

6 Acknowledgements

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

7 References

1. Baeza-Yates, R., and Ribeiro-Neto, B., *Modern Information Retrieval*. 1999: ACM Press, Addison-Wesley.
2. Bradley, A.P., *The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms.* Pattern Recognition, 1997. **30**(7): p. 1145-1159.
3. Swanson, D.R., *Migraine and Magnesium: Eleven Neglected Connections.* Perspect. Biol. Med., 1988. **31**: p. 526-557.
4. Swanson, D.R., *Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge.* Perspect. Biol. Med., 1986. **30**(1): p. 7-15.
5. Smalheiser, N.R., and Swanson, D., *Linking Estrogen to Alzheimer's Disease: An Informatics Approach.* Neurology, 1996. **47**(3): p. 809-810.
6. Smalheiser, N.R., and Swanson, D., *Indomethacin and Alzheimer's Disease.* Neurology, 1996. **46**(2): p. 583.
7. Swanson, D.R., *Somatomedin C and Arginine: Implicit Connections between Mutually Isolated Literatures.* Perspectives in Biology and Medicine, 1990. **33**(2): p. 157-186.
8. Smalheiser, N.R., and Swanson, D., *Calcium-Independent Phospholipase A2 and Schizophrenia.* Arch Gen Psychiatry, 1998. **55**: p. 752-753.
9. Lindsay, R.K., and Gordon, M.D., *Literature based discovery by lexical statistics.* Journal of American Society for Information Science, 1999. **49**(8): p. 674-685.
10. Gordon, M.D., and Dumais, S., *Using latent semantic indexing for literature based discovery.* Journal of American Society for Information Science, 1998. **49**(8): p. 674-685.
11. Blake, C., and Pratt, W. *Automatically Identifying Candidate Treatments from Existing Medical Literature.* . in *Proceedings of AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*. 2002. California.
12. Weeber, M., Klein, H., and de Jong - van den Berg, L.T.W., *Using Concepts in Literature Based Discovery: Simulating Swanson's Raynaud-Fish Oil and Migraine-Magnesium Examples.* Journal of American Society for Information Science, 2001. **52**(7): p. 548-557.
13. Srinivasan, P., *Generating Hypotheses from MEDLINE.* Journal of American Society for Information Science, 2004. **55**(5): p. 396-413.
14. Hu, X., Li, G., Yoo, I., Zhang, X., and Xu, X. *A Semantic-based Approach for Mining Undiscovered Public Knowledge from Biomedical Knowledge*. in *Proceedings of IEEE International Conference on Granular Computing*. 2005. Beijing.
15. Pratt, W., and Yetisgen-Yildiz, M. *LitLinker: Capturing Connections across the Biomedical Literature*. in *Proceedings of International Conference on Knowledge Capture (K-Cap'03)*. 2003. Florida.

16. Hristovski, D., Stare, J., Peterlin, B., and Dzeroski, S. *Supporting discovery in medicine by association rule mining in Medline and UMLS*. in *Proceedings of Medinfo*. 2001.
17. Yetisgen-Yildiz, M., and Pratt, W., *Using Statistical and Knowledge-Based Approaches for Literature Based Discovery.* Journal of Biomedical Informatics (To appear), 2006.
18. Wren, J.D., *Extending the mutual information measure to rank inferred literature relationship.* BMC Bioinformatics, 2004. **5**(1): p. 145.
19. Weeber, M., Vos, R., Klein, H., de Jong - van den Berg, L.T.W., and Aronson, A.R., *Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide.* Journal of American Medical Informatics Association 2003. **10**(3): p. 252-259.
20. Wren, J.D., Bekeredjian, R., Stewart, J.A., Shohet, R.V., and Garner, H.R., *Knowledge discovery by automated identification and ranking of implicit relationships.* Bioinformatics, 2004. **20**(3): p. 389-398.
21. Swanson, D.R., *Atrial fibrillation in athletes: implicit literature-based connections suggest that overtraining and subsequent inflammation may be a contributory mechanism.* Medical Hypotheses, 2006. **66**(6): p. 1085-92.
22. Skeels, M.M., Henning, K., Yetisgen-Yildiz, M., and Pratt, W. *Interaction Design for Literature-Based Discovery*. in *Proceedings of the International Conference for Human-Computer Interaction (CHI'05)*. 2005. Portland, WA.
23. Smalheiser, N.R., et. al., *Collaborative development of the Arrowsmith two node search interface designed for laboratory investigators.* Journal of Biomedical Discovery and Collaboration, 2006. **1**(8).