The Knowledge Engineering Review, Vol. 32, e12, 1–35.
doi:10.1017/S0269888917000042

Emerging approaches in literature-based discovery: techniques and performance review

YAKUB SEBASTIAN[1], EU-GENE SIEW[2] and SYLVESTER O. ORIMAYE[3]

[1]*Faculty of Engineering, Computing and Science, Swinburne University of Technology Sarawak Campus, Jalan Simpang Tiga, 93350 Kuching, Sarawak, Malaysia;*
e-mail: ysebastian@swinburne.edu.my;
[2]*School of Business, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, 47500 Subang Jaya, Selangor, Malaysia;*
e-mail: siew.eu-gene@monash.edu;
[3]*School of Information Technology, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, 47500 Subang Jaya, Selangor, Malaysia;*
e-mail: sylvester.orimaye@monash.edu

Abstract

Literature-based discovery systems aim at discovering valuable latent connections between previously disparate research areas. This is achieved by analyzing the contents of their respective literatures with the help of various intelligent computational techniques. In this paper, we review the progress of literature-based discovery research, focusing on understanding their technical features and evaluating their performance. The present literature-based discovery techniques can be divided into two general approaches: the traditional approach and the emerging approach. The traditional approach, which dominate the current research landscape, comprises mainly of techniques that rely on utilizing lexical statistics, knowledge-based and visualization methods in order to address literature-based discovery problems. On the other hand, we have also observed the births of new trends and unprecedented paradigm shifts among the recently emerging literature-based discovery approach. These trends are likely to shape the future trajectory of the next generation literature-based discovery systems.

1 Introduction

Literature-based discovery (LBD) encompasses various computational approaches that aim at discovering previously unknown associations between pieces of existing knowledge by analyzing their relevant literatures (Swanson, 2008). Due to the explosion in the number of scientific literatures being published today, it has become increasingly challenging to keep track of the developments in all research areas, which in turn may leave many valuable logical connections between disparate bodies of knowledge remain unnoticed (Swanson, 1986b). In view of this challenge, LBD aims at exploring algorithmic approaches to finding hidden links between previously disjoint groups of research papers, either in automatic or semi-automatic fashion (Smalheiser, 2012). This paper provides a comprehensive review and performance evaluation of the existing LBD techniques. In addition to presenting a useful classification of most LBD methods, it places a special emphasis on evaluating the performance of recently emerging techniques.

The development of LBD as a research field can be traced to Swanson's serendipitous discovery of the potential benefits of dietary fish oil on the treatment of Raynaud's syndrome (Swanson, 1986a; DiGiacomo *et al*., 1989). Unlike other typical laboratory-situated discoveries, Swanson's discovery was groundbreaking in that it was generated using a content analysis technique applied to two seemingly unrelated sets of literatures.

Several review papers have been previously published in this field, yet with some limitations. Davies (1989) provided the earliest, but limited coverage on Swanson's pioneering work on LBD.

Subsequent reviews by Weeber *et al*. (2005) and Bekhuis (2006) fell short of giving sufficient technical depths as they were intended for non-technical audience, especially biomedical researchers and digital librarians.

Other reviews are limited in terms of their topical coverage. In an attempt to encourage more work toward alternative discovery models, Smalheiser (2012) focused only on criticizing the limitations of Swanson's ABC discovery model. In the same way, the review by Yetisgen-Yildiz and Pratt (2009) was limited to discussing the limitations of the present LBD evaluation methodologies. Jensen *et al*. (2006), Hahn *et al*. (2012) and Li *et al*. (2014) only discussed highly specialized LBD systems for biology and pharmacogenomics.

There are more comprehensive reviews. Ganiz *et al*. (2005) provided detailed discussions on various LBD algorithms, whereas Kostoff *et al*. (2009) focused their attention on critically evaluating the quality of discoveries produced by the existing LBD systems. Unfortunately, both reviews did not cover many newer LBD techniques.

Unlike the previous review papers, the current paper affords more in-depth technical discussions on a wide range of existing LBD algorithms. More importantly, we offer insightful performance evaluations on some of the most recently emerging LBD approaches, with important implications for future research. Further, the present review will benefit both novice and experienced researchers. Novice researchers and less-technical readers who are new to LBD may use this paper as an introduction to LBD. To further assist these readers, we supply many highly informative diagrams to help understand the complex methodological features of various LBD approaches being described. For more experienced LBD researchers, this review provides valuable insights into the field's technical developments, critical performance evaluations, unaddressed research challenges, and the predicted future research directions.

The rest of this paper is organized as follow. Section 2 defines LBD. It explains the basic model underlying most LBD approaches and argues for its significance within the contemporary scientific communities. The section also highlights the evolution of LBD approaches over time and suggests a taxonomic structure for categorizing different types of LBD techniques. The subsequent sections look at each LBD category in more detail, starting with Section 3 which broadly reviews various techniques under the traditional LBD approach. It is followed by Section 4, where we present the main contributions of this paper. Here, we conduct the performance evaluations of notable examples of the emerging LBD approaches. We also discuss their methodological characteristics, strengths and limitations at great length, before closing this section by anticipating several future trends among these emerging techniques. Section 5 discusses a number of pertinent research problems that may shape the future trajectory of LBD research. Section 6 concludes this paper.

2 Literature-based discovery

2.1 Overview

LBD can be defined as a systematic computational approach to combining distinct and previously disconnected pieces of knowledge found in the existing literature in order to infer novel and interesting knowledge (Ganiz *et al*., 2005; Swanson, 2008). The main goal of LBD is to produce interesting and novel knowledge, where 'novelty' refers to any fact or finding that has never been publicly published in scientific literatures (Swanson, 2008). Although some authors have viewed LBD as a subset of the largest text mining field (Berry & Castellanos, 2004; Smalheiser, 2012), many LBD systems incorporated non-text mining methods that utilize structured databases, such as STRING[1] and OMIM[2], as well as semi-structured information sources, such as in-text citations, images, tables, bibliographic metadata, and citation links.

The majority of LBD techniques build upon a fundamental premise known as the *ABC discovery model* (Swanson, 1987; Weeber *et al*., 2005; Bekhuis, 2006; Smalheiser, 2012). This model is intuitive, easy to understand, and versatile (Smalheiser, 2012). It operates based on a simple syllogistic reasoning which assumes that if an object *A* is associated with another object *B* and that object *B* is associated with yet another object *C*, then it can be inferred that object *A* is eventually associated with object *C*.

[1] http://string-db.org/
[2] http://www.omim.org/

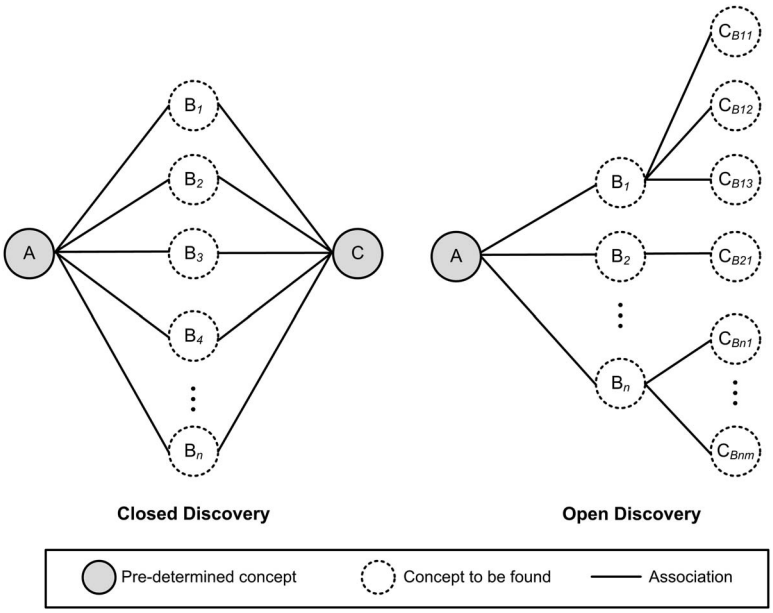

Figure 1 The closed discovery model (left) and the open discovery model (right)

Two variants of the ABC model exist, as depicted in Figure 1 (Ganiz *et al*., 2005; Weeber *et al*., 2005; Kostoff *et al*., 2009). The first model, *closed discovery*, requires that objects *A* and *C* be predetermined by the user. In this case, the goal of LBD is to find novel associations $A-B$ and $B-C$ that facilitate the implicit association between *A* and *C*. The second model, *open discovery*, assumes that only *A* is known to the user and that the main objective is to find the possible links to an unknown object *C*, given one or more intermediate objects *B*. Hence, while the first model focuses on finding objects *B*, the second model emphasizes on looking for both *B* and *C* objects.

2.2 Motivation

The development of a research field is often influenced by various external factors. As previously suggested, LBD research initially started in response to the need to find previously unknown logical associations between the increasingly fragmented pieces of knowledge in the midst of currently explosive growth of scientific literatures (Swanson, 1979, 1986b, 1993; Larsen & Von Ins, 2010). Apart from this primary motivation, we also observe other key factors fueling the ongoing interests in LBD research: the growing number of evidence advocating more interdisciplinary approaches to research, and the accumulative evidence of the existing LBD systems' effectiveness for addressing real world problems.

First, a number of recent research findings have suggested the potential merit of the interdisciplinary research approach. A report by the US National Research Council of the National Academies (Feller & Stern, 2007) observed that many scientific discoveries often involve drawing novel connections between various scientific domains, which coincidentally resembles the LBD model. Likewise, a study by Chen *et al*. (2009) reported that many important scientific breakthroughs could be characterized by the presence of unprecedented co-citation links between previously disjoint groups of research papers.

What matters more is that the trend above seems to generalize to most scientific publications. For instance, Uzzi *et al*. (2013) studied 17.9 million records in Thomson Reuter's *Web of Science*, where they found that the highest-impact scientific papers often involved making unusual combinations between commonly known knowledge. Similarly, in a study on US Patent records in years between 1790 and 2010, Youn *et al*. (2015) discovered that many previously reported inventions featured new combinations between existing techniques and technologies artifact. Once again, these findings appear to be consistent with the underlying idea behind LBD.

Second, recent advances in LBD research may have also been motivated by the increasing number of studies that report the successful applications of LBD methods in a wide range of knowledge discovery scenarios.

For instance, Swanson's hypothesis on the relationship between fish oil and Raynaud's syndrome (Swanson, 1986a) was successfully validated via an actual clinical trial study (DiGiacomo *et al*., 1989). Not only that, there have also been evidence of LBD techniques' being used to discover previously unknown associations between various biomedical entities, including between *migraine* and *magnesium* (Swanson, 1988), *estrogen* and *Alzheimer's disease* (Smalheiser & Swanson, 1996b), *indomethacin* and *Alzheimer's disease* (Smalheiser & Swanson, 1996a), *nordihydroguaiaretic acid* and *breast cancer* (Sneed, 2003), *curcumin longa* and *retinal disease* (Srinivasan *et al*., 2004), *chlorpromazine* and *cardiac hyperthropy* (Wren *et al*., 2004), *Parkinson's disease* (PD) and *Crohn's disease* (CD) (Kostoff, 2014), and between *hypogonadism* and *sleep quality* (Miller *et al*., 2012).

Besides, LBD techniques have also been used to suggest possible new treatments for existing diseases, such as *cataracts* (Kostoff, 2008), PD (Kostoff & Briggs, 2008), *multiple sclerosis* (Kostoff *et al.*, 2008), and *breast cancer* (Li *et al*., 2010). They have also been applied to find new therapeutic uses of existing drugs such as *thalidomide* (Weeber *et al*., 2003) and *Metformin* (Ding *et al*., 2013). Finally, special purpose LBD systems have been created to elucidate novel *gene–disease associations* (Perez-Iratxeta *et al*., 2005), *protein–protein interactions* (van Haagen *et al*., 2009, 2011), *adverse drug reactions* (Shang *et al*., 2014) and *drug repositioning* (Andronis *et al*., 2012; Wei *et al*., 2014).

2.3 Classification of literature-based discovery techniques

The progress of LBD research has resulted in the evolving approaches and techniques over time, as depicted in Figure 2. The diagram suggests a general trend where LBD techniques tend to become increasingly automated, relying on the usage of richer knowledge representations, and better equipped to tackle more complex discovery problems. This trend has given rise to a set of distinct LBD approaches, which we have indicated in the form of shaded ellipses in the diagram. The overlapping sections of these ellipses point to hybrid LBD approaches.

This review proposes a taxonomy that divides the current LBD techniques into two broad approaches, namely the *traditional approach* and the *emerging approach*. Each approach is further divided into several categories. The traditional approach consists of three subcategories, that is, the *statistical approach*, *knowledge-based approach*, and *visualization approach* (Ganiz *et al*., 2005; Bekhuis, 2006; Smalheiser, 2012), as illustrated in Figure 3. Likewise, the emerging LBD approach is divided into distinct approaches, which include the *context-driven subgraph model* (Cameron *et al*., 2015), *bibliographic coupling model* (Kostoff, 2014), *cluster similarity model* (Fujita, 2012), *entitymetrics* (Ding *et al*., 2013), and *heterogeneous graph models* (Eronen & Toivonen, 2012; Sebastian *et al*., 2015).

As we emphasized at the beginning of this paper, the primary contribution of this review is in our detailed performance evaluation of the emerging LBD approach. Unlike its traditional counterparts, the emerging approach demonstrates better performance in terms of the ability to infer complex associations within the literature, on top of their higher scalability and accuracy (Cameron, 2014). Nonetheless, in order

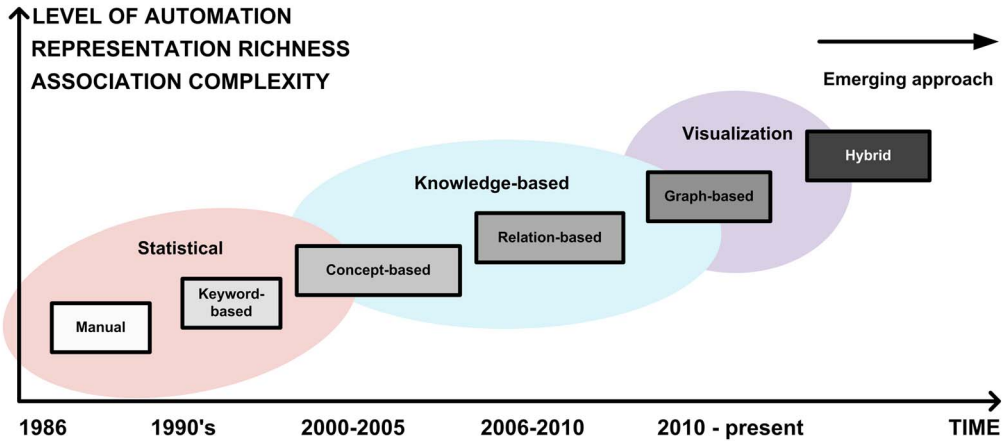

Figure 2 The technical evolution of literature-based discovery approaches (adapted from Cameron, 2014)

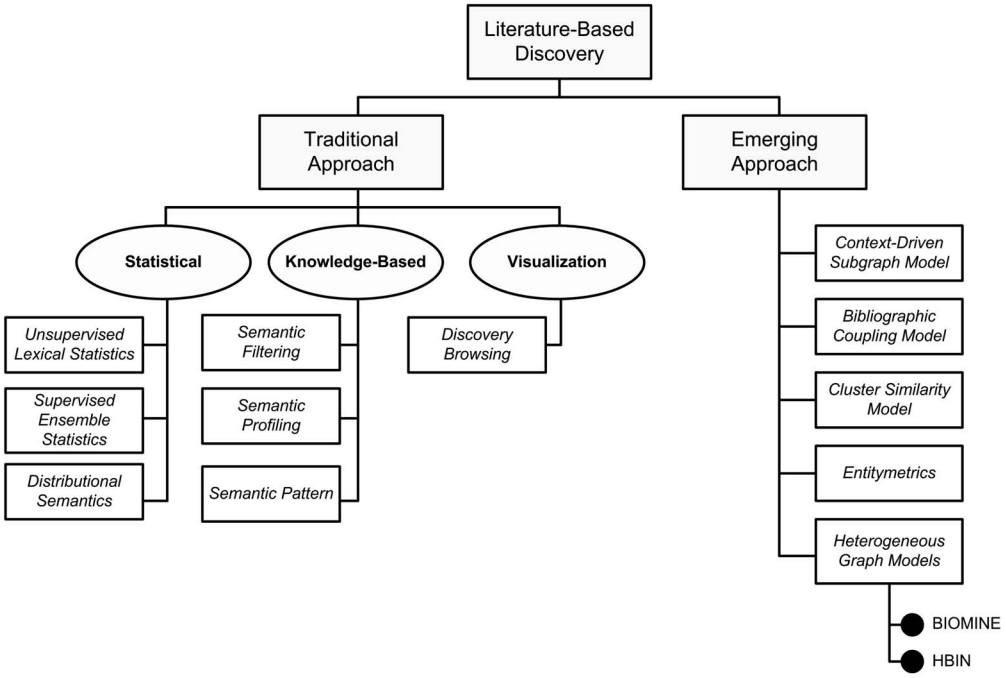

Figure 3 Classification of literature-based discovery techniques. HBIN = heterogeneous bibliographic information network

to fully understand and appreciate the merits of this newer approach, it is important to revisit several representative traditional LBD techniques. We cover these in the following sections.

3 Traditional literature-based discovery approaches

3.1 Statistical approach

The statistical LBD approach establishes links between disjoint knowledge by looking for the most frequently co-occurring terms or concepts in the existing literature. This is primarily achieved by computing the statistical distributions and frequencies of the terms, without further considerations for their semantics (Lindsay & Gordon, 1999). Assuming the closed discovery model and given source term *A* and target term *C*, this approach employs various statistical measures to determine which intermediate terms *B* could meaningfully connect *A* to *C* (Gordon & Lindsay, 1996).

For instance, to replicate Swanson's previously mentioned dietary fish oil–Raynaud's syndrome (DFORS) discovery (Swanson, 1986a), Swanson and Smalheiser (1997) retrieved all articles containing a source term 'fish oil' in their titles. Other terms that appeared frequently with this source term were selected as the possible intermediate terms, such as the term 'blood viscosity'. The intermediate terms were then ranked based on the relative frequency of their co-occurrences with the source term, allowing terms that exceeded a predetermined threshold value to be shortlisted. In the next phase, each shortlisted intermediate term was treated as if it were a new source term and the same procedure above was repeated. The final result was a list of possible target terms (e.g. 'Raynaud's syndrome').

We further divide the statistical LBD approach into three categories of techniques: *unsupervised lexical statistics*, *supervised ensemble statistics*, and *distributional semantics*. Each category is discussed in the sections below.

3.1.1 Unsupervised lexical statistics techniques

Considered to be one of the earliest LBD techniques, the *unsupervised* lexical statistics techniques (Gordon & Lindsay, 1996; Lindsay & Gordon, 1999; Gordon *et al*., 2002) rely on word count statistics, as illustrated in Figure 4. To perform the ABC discovery model, the steps on the gray plate are repeated

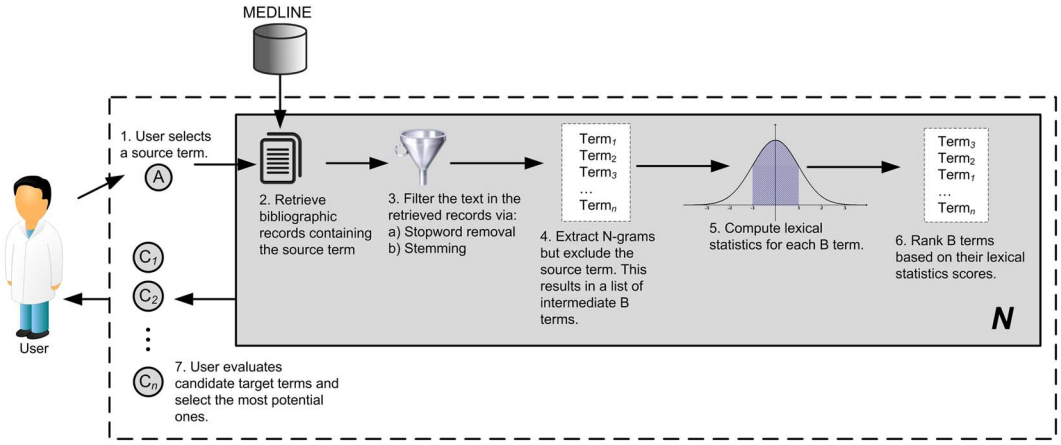

Figure 4 The general methodology of the unsupervised lexical statistics technique proposed by Gordon and Lindsay (1996) and Lindsay and Gordon (1999). The dotted line marks the system's boundary.

N times, where *N* denotes the number of intermediate term *B* connecting the source term *A* to target term *C*. The effectiveness of this technique is determined by the extent to which the desired intermediate terms and target terms (assumed to be known in advance) rank highly in the list of candidate terms.

To replicate Swanson's magnesium–migraine (MM) discovery (Swanson, 1988), the technique first downloaded bibliographic records containing the source term 'migraine'. The records included titles, abstracts, and subject descriptors. *N*-grams[3] were subsequently extracted from these records, following prior word-stemming and stopwords removal[4]. Note that these *N*-grams should not be words that typically characterize the source literature (e.g. the word 'migraine' or other closely associated terms), nor should they include very general terms. Finally, for each extracted *N*-gram, a number of lexical statistics scores were calculated in order to determine and rank its importance in bridging *A* and *C*. The used scores included *token frequency* (*tf*), *document frequency*, *relative frequency*, and *frequency* × *inverse document frequency* (*tf* × *idf*) (Gordon & Lindsay, 1996; Lindsay & Gordon, 1999).

The unsupervised lexical statistic techniques offer some benefits in terms of their high intuitiveness and ease of computation. However, given their reliance on using lexical statistical measures that tend to favor frequently co-occurring terms, they may miss other interesting associations originating from less frequent terms (Kostoff *et al.*, 2009; Petrič *et al.*, 2010). Not only that, the success of this technique greatly depends on the user's prior knowledge and inherent bias. For example, in the MM discovery, Lindsay & Gordon (1999) removed a number of highly ranked intermediate terms given the authors' foreknowledge that the terms would not eventually lead to the desired target term 'magnesium'.

3.1.2 Supervised ensemble statistics technique

The second category of the statistical approach is the supervised ensemble statistics technique. This technique derives a single score by learning a number of different weighted features from a text corpus. As shown in Figure 5, this learning step is facilitated with the help of logistics regression algorithm, after which the learned score is used for deciding the relevance and the interestingness of various intermediate terms that connect source term *A* and target term *C* (Torvik & Smalheiser, 2007).

In contrast to the previous unsupervised lexical statistics approach, this technique utilizes a machine learning technique to determine the most relevant intermediate terms. Hence, it requires less human intervention and therefore less prone to the user bias. There are limitations, however. Since the choice of the learned features was arbitrary, there is also the possibility that other useful yet unexplored features might have been overlooked (Torvik & Smalheiser, 2007). More importantly, this technique requires that a

[3] *N*-gram is the continuous *n*-term or *n*-word sequence in a text.

[4] Stemmed words are words whose inflections have been removed by a stemming algorithm so that only their base or root forms are retained. The goal is to reduce level of noise in text. Refer to `http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html` for more details.

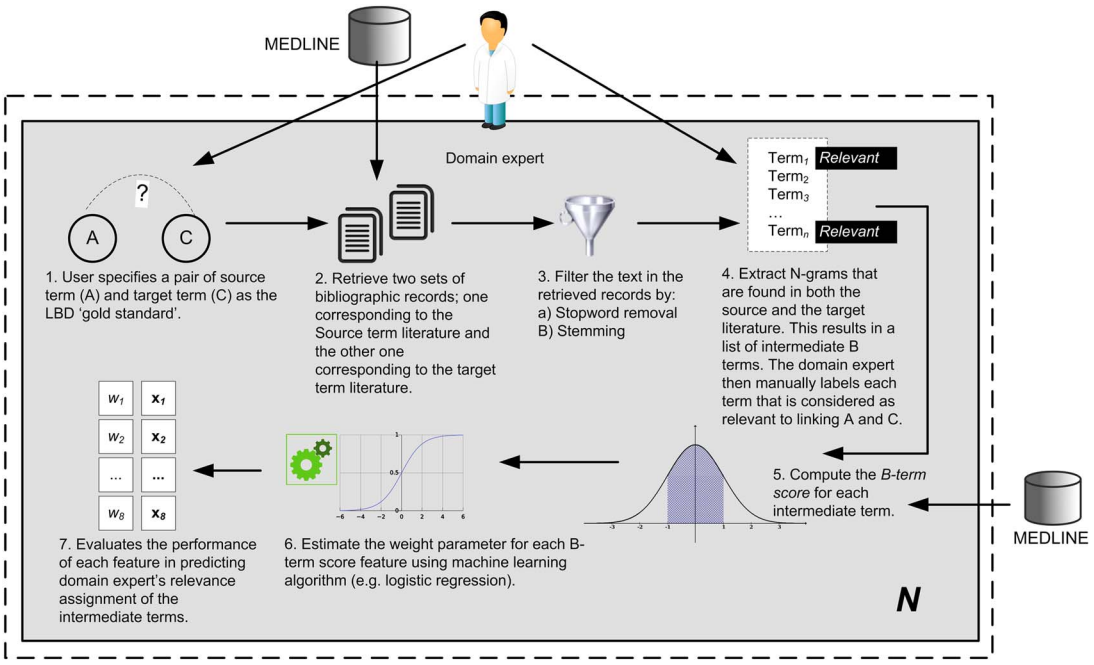

Figure 5 The procedure for learning feature weights using the supervised ensemble statistics technique (Torvik & Smalheiser, 2007). LBD = literature-based discovery

large number of high-quality gold standards be available to adequately train the model, which is still a challenging research problem (Smalheiser, 2012).

3.1.3 Distributional semantics techniques

Unlike the previous two approaches, distributional semantic technique uses scalable algorithms to build representation of terms based on the patterns of their occurrence in natural language texts (Symonds *et al*., 2014). At the heart of this technique is the assumption that two terms are semantically related if they appear in a *similar context* (Salton & McGill, 1986; Symonds *et al*., 2014). The context of a term is defined as a vector representation of other terms that co-occur with it. As a result, two semantically related terms are expected to exhibit similar vector representations. Figure 6 depicts the overall methodology of this technique.

Gordon and Dumais (1998) applied the *Latent Semantic Indexing* (LSI) algorithm (Deerwester *et al*., 1990) that is capable of directly computing the semantic similarity between a source term and a target term even if both terms never appear together. The technique also eliminates the need to generate and evaluate large numbers of intermediate terms. Related to Gordon and Dumais (1998), Cohen *et al*. (2010) introduced a distributional semantics technique called the *Reflective Random Indexing* (RRI), which scaled better than LSI for processing very large corpora.

Besides its ability to eliminate the need to generate and evaluate a large number of intermediate terms from the ABC discovery model, the distributional semantics technique has better scalability, up to a million MEDLINE records (Cohen *et al*., 2010). Furthermore, these techniques can be applied to perform LBD on non-English documents (Symonds *et al*., 2014).

There are limitations. LSI experienced difficulties in identifying the desired target terms when evaluated on Swanson's DFORS hypothesis (Gordon & Dumais, 1998). Although the performance reportedly improved when the RRI model was used (Cohen *et al*., 2010), it required that certain intermediate terms (e.g. 'platelet') be used to guide the search for the correct target term. Moreover, when used to analyze MEDLINE records, the techniques' specificity and precision only improved under the condition that the Medical Subject Headings (MeSH) terms were supplied to the algorithm (Cohen *et al*., 2010, 2012). Lastly, since documents are modeled as bags-of-words, this technique is oblivious to the positional and sectional information of terms, making it difficult to interpret the LBD results (Cohen & Hersh, 2005).

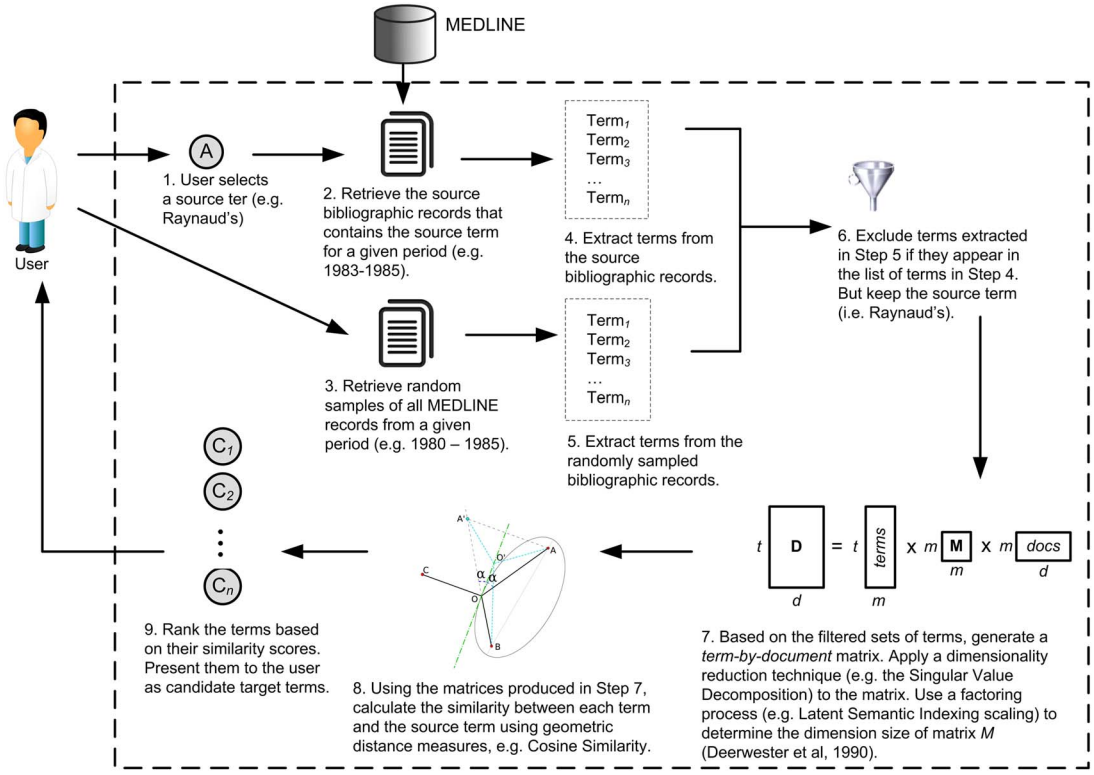

Figure 6 The methodology of the distributional semantics technique (Gordon & Dumais, 1998)

In addition to the three subcategories of statistical LBD approach discussed above, other types of statistical measures have been explored. They include *term frequency-based measures* (Gordon & Lindsay, 1996; Lindsay & Gordon, 1999), *association rules* (Hristovski *et al.*, 2000; Pratt & Yetisgen-Yildiz, 2003), *fuzzy binary relations* (Perez-Iratxeta *et al.*, 2002), *mutual information measure* (Wren, 2004), *z-score* (Yetisgen-Yildiz, 2006; Yetisgen-Yildiz & Pratt, 2006), and *R-scaled score* (Frijters *et al.*, 2010).

3.2 Knowledge-based approach

The second main subset of the traditional LBD approach is the knowledge-based approach, which owes its performance to the ability to mine external domain-specific knowledge-based resources (e.g. ontologies and biomedical databases), instead of using lexical statistics to achieve the accuracy in detecting meaningful hidden relations between disconnected literatures (Weeber *et al.*, 2005). It is common to find natural language processing (NLP) and information extraction algorithms supplying the reasoning capabilities required of these LBD systems (Weeber *et al.*, 2005; Bekhuis, 2006; Smalheiser, 2012).

The knowledge-based LBD approach usually prioritizes intermediate terms and target terms according to a set of predetermined semantic types. For example, Lytras *et al.* (2005) and Hu *et al.* (2005) removed concept terms that belonged to very general semantic types and prioritized only the source and target terms that satisfied a set of predefined semantic relations in a biomedical ontology. There are three types of knowledge-based LBD techniques in the literatures: the *semantic filtering* technique, *semantic profiling* techniques, and *semantic pattern* techniques.

3.2.1 Semantic filtering technique

The semantic filtering technique maps natural language text to specific biomedical concepts using NLP tools, such as *MetaMap*[5]. The Figure 7 illustrates the steps in this technique. The aim is to obtain

[5] `http://metamap.nlm.nih.gov/`

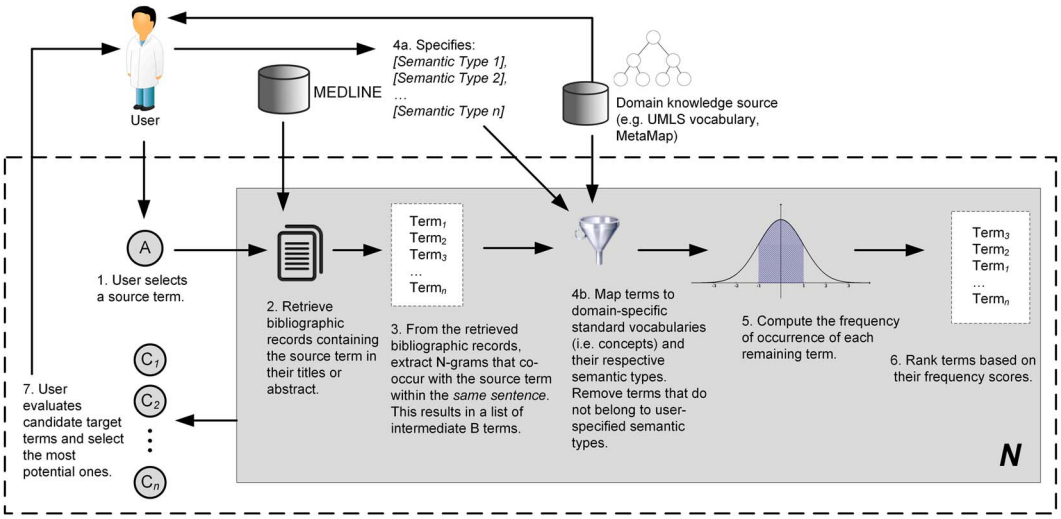

Figure 7 The steps involved in the semantic filtering technique (Weeber *et al.*, 2001). The dotted line marks the system boundary

intermediate terms which satisfy a set of user's predefined semantic types, thus reducing the total number of intermediate concepts that need to be evaluated by the user.

Unlike previously described lexical statistics technique, this technique relies on semantic type filtering to provide its accuracy. Users significantly influence the discovery process through the formulation of appropriate semantic filters. But the technique has a drawback in that it requires non-trivial amount of domain knowledge to select the most effective semantic filters. Furthermore, it assumes that the actual target terms are known in advance. Unfortunately, for many real world scenarios it may not be possible to determine these target terms from the outset of a discovery process (Weeber *et al.*, 2001).

3.2.2 Semantic profiling technique

Srinivasan (2004) introduced *topic profile* (*TP*), which is a vector of semantic type vectors. Figure 8 describes in detail the technique's algorithm for solving the open discovery problem.

A semantic type vector is an unordered set of weighted MeSH terms that belong to a specific semantic type. Since there are 134 semantic types in the UMLS (Unified Medical Language System) ontology, a TP may be made up of up to 134 semantic type vectors. Equation (1) illustrates how a TP_i may look like. In this formula, w_i,134,1 corresponds to the weight of the MeSH term m. The MeSH term $m_{134,1}$, in turn, belongs to the 134th semantic type (Srinivasan, 2004):

$$TP_i = \left\{ \left\{ w_{i,1,1} m_{1,1}, w_{i,1,2} m_{1,2}, \ldots \right\}, \ldots, \left\{ w_{i,134,1} m_{134,1}, w_{i,134,2} m_{134,2}, \ldots \right\} \right\} \quad (1)$$

In this technique, the MeSH weight is calculated as a $tf - idf$ score (Salton & McGill, 1986), such that a TP represents the relative importance of various semantic types in the text based on the frequency of occurrence of their component MeSH terms. Similar to the TP technique, van Haagen *et al.* (2009) later proposed *concept profile*, which was capable of measuring the similarity between two types protein concepts in the literature.

In contrast to the semantic filtering technique (Weeber *et al.*, 2001), the advantage of semantic profiling technique is obvious: it provides a way to assign the relative importance of a semantic type based on the accumulative weights of its component MeSH terms (Srinivasan, 2004). Depending on the nature of an LBD task, this allows one to prioritize certain semantic types over the others. Unfortunately, as suggested in Figure 8, this technique is computationally expensive as it requires generating and evaluating every single intermediate term in order to arrive at the correct target terms. Like many other knowledge-based techniques, the success of this technique also relies on incorporating the information from various domain-specific knowledge bases, for example, MeSH biomedical vocabularies.

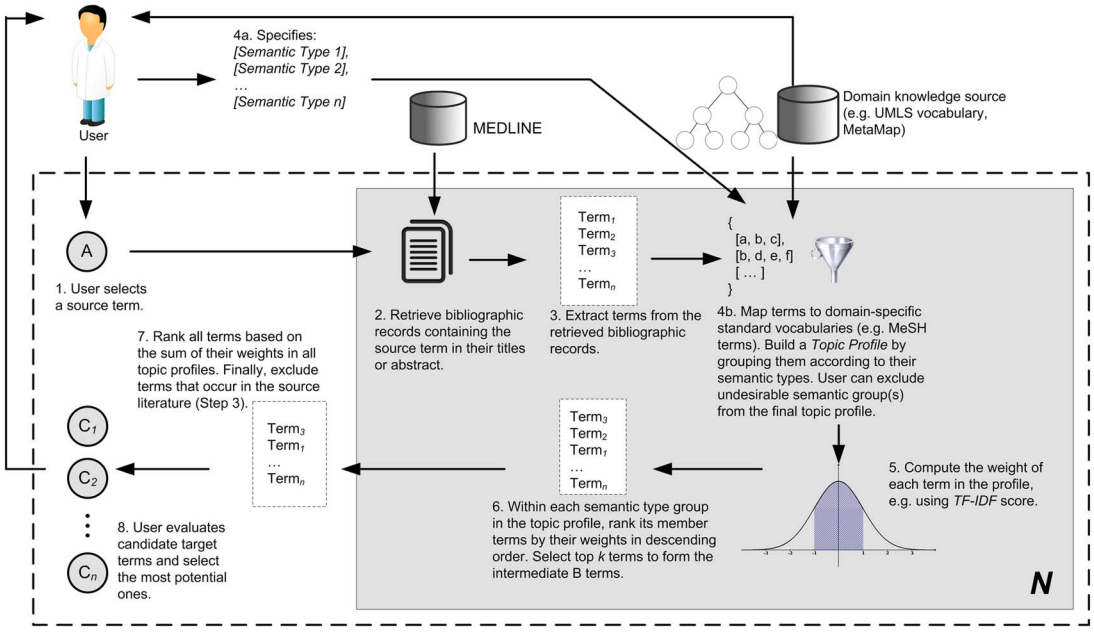

Figure 8 The steps used by the profiling technique (Srinivasan, 2004). MeSH = Medical Subject Headings; tf = token frequency; idf = inverse document frequency

3.2.3 Semantic pattern techniques

These techniques use semantic predications extracted from natural language texts. Hristovski *et al*. (2006) introduced *discovery pattern*, a set of conditions in the form of a subject–predicate–object-like structure known as *predication*. These pattern-like conditions guide the evaluation of candidate novel associations generated by the LBD system. Figure 9 illustrates the process.

To replicate Swanson's DFORS hypothesis, Hristovski *et al.* searched for complementary predications in MEDLINE texts that conform to a certain discovery pattern. For example, the sentence 'local increase of blood viscosity during cold-induced Raynaud's phenomenon' was parsed to produce predication ASSOCIATED_WITH (*Raynaud's*, *blood viscosity*, *increase*). In the same manner, the sentence 'a statistically significant reduction in whole blood viscosity was observed at seven weeks in those patients receiving the eicosapentaenoic acid rich oil' was parsed to obtain predication ASSOCIATED_WITH (*eicosapentaenoic acid*, *blood viscosity*, *decrease*). By assembling these two complementary predications, the user may hypothesize that, since eicosapentaenoic acid (i.e. dietary fish oil) decreases blood viscosity and high blood viscosity is observed among most Raynaud's patients, it can then be inferred that regularly consuming eicosapentaenoic acid may alleviate Raynaud's syndrome.

Developing from Hristovski *et al.*'s discovery pattern model, the *Predication-Based Semantic Indexing* model represented concepts and their relationships as vectors in a hyperdimensional space (Cohen *et al.*, 2012). In this case, finding a discovery pattern is viewed as a geometrical function in a hyperdimensional space that is solved by tracing certain predication pathways.

The main strength of the semantic pattern techniques is their ability to better interpret the nature of the associations among concepts in the literature (Hristovski *et al.*, 2006). As a result, they may be used to figure out more complex hidden associations better than the lexical statistics approach (Kraines *et al.*, 2010). And given that the associations among semantic types are explicitly represented in the form of predications, it is easy for the user to verify the plausibility of associations between a source and a target concept.

There are drawbacks, though. First, scalability remains an issue. The technique requires two distinct stages in order to extract the semantic predications: an initial stage where the semantic predications are extracted for the source concept and each of the intermediate concepts, and the second stage where semantic predications must be extracted for each intermediate concept and the target concept. Since the number of intermediate concepts tends to grow exponentially (Wren, 2008), these two-staged procedures almost certainly become a bottleneck in the algorithm (Cameron *et al.*, 2013). Other limitations include the

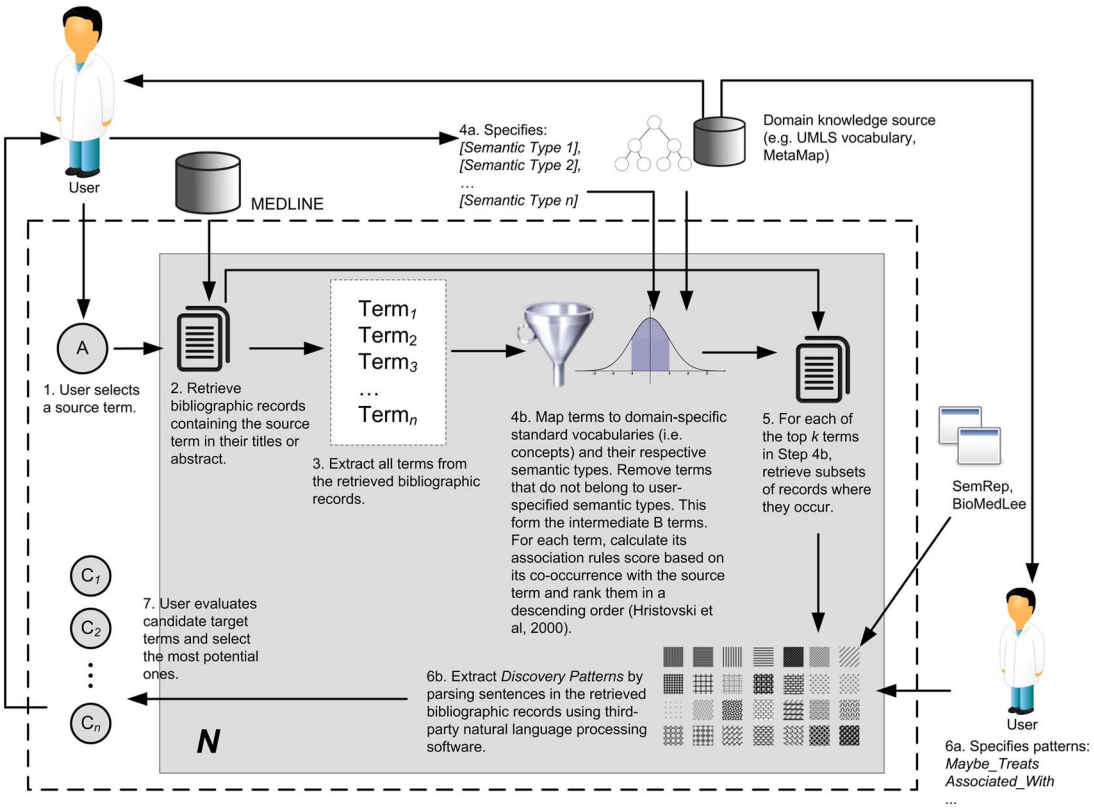

Figure 9 The steps involved in the semantic pattern technique (Hristovski *et al.*, 2006)

technique's requirement that the user be able to select the most promising intermediate and target concepts, as well as its reliance on the availability of third-party NLP software to enable the extraction of predications from natural language texts (Hristovski *et al.*, 2006).

3.3 Visualization approach

This last category of the traditional LBD approach harnesses graph visualizations. Graphs and networks provide a versatile representation and visualization of structured and unstructured information (Juršič *et al.*, 2012; Chen *et al.*, 2013; Ding *et al.*, 2013). For LBD, graph representations can be used to visualize the relationships between terms or concepts in text. For instance, a syllogistic association between two disconnected objects x and y can be identified when a graph shows that x is connected to object z, and object z is subsequently connected to object y (Narayanasamy *et al.*, 2004).

van Mulligen *et al.* (2002) introduced two-dimensional graphs in which strongly correlated concepts were plotted close to each other on a graphical space, known as the *Associative Concept Space*. The implicit connections between a source term and a target term were inferred by automatically looking for pairs of complementary links between a source term and an intermediate term, as well as the links between the intermediate term and a target term link pairs in the graph.

Following van Mulligen *et al.*, a number of other visualization techniques were studied. Narayanasamy *et al.* (2004) explored the effectiveness of an association graph for modeling transitive associations. Wilkowski *et al.* (2011) proposed *discovery browsing* technique, which plotted the subject and object of predications as nodes and their predicates as links in a semantic graph. This algorithm has the capability of capable of automatically suggesting interesting paths along the graph. We further discuss this technique below.

3.3.1 Discovery browsing techniques

The *discovery browsing* technique allows users to navigate through the complex relationships among biomedical concepts, using a combination of semantic predications and graph-based methods

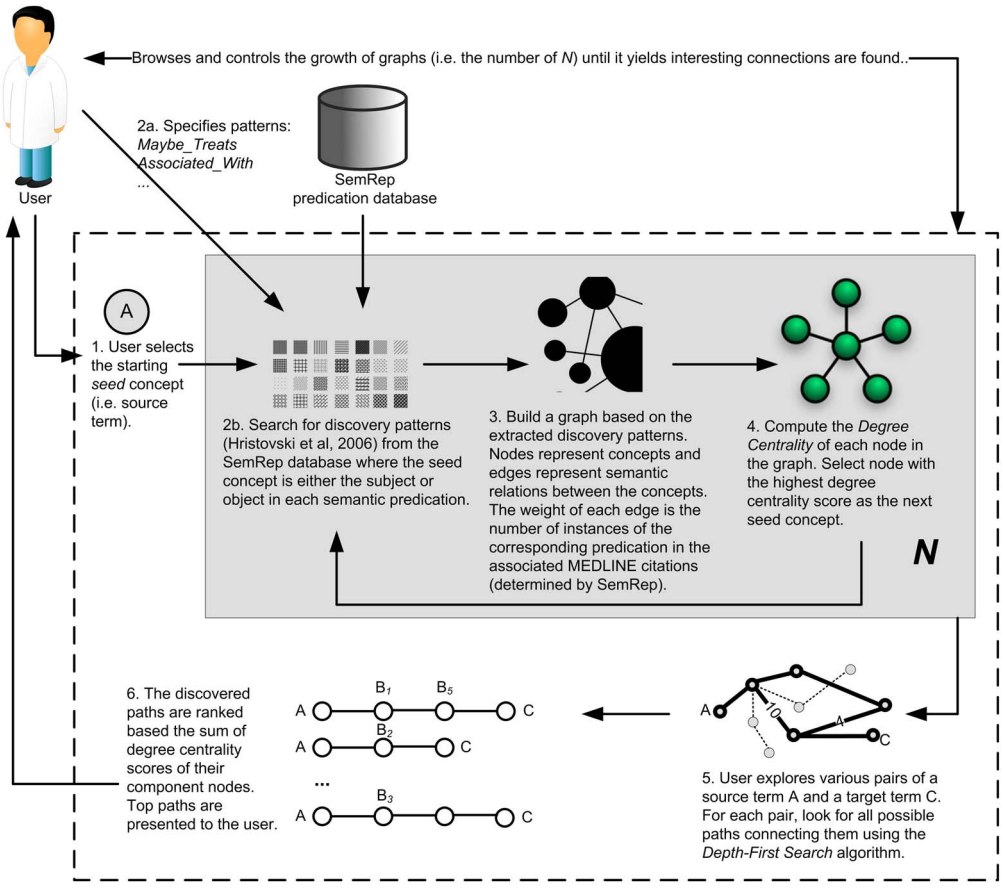

Figure 10 The steps involved in the discovery browsing technique (Wilkowski *et al.*, 2011)

(Wilkowski *et al.*, 2011; Goodwin *et al.*, 2012). It extended the discovery pattern technique proposed earlier by Hristovski *et al.* (2006) by allowing a serial chain of multiple intermediate B terms between a source concept A to a target concept C, such that $A \leftrightarrow (B_1 \leftrightarrow B_2 \leftrightarrow B_3 \leftrightarrow \ldots B_n) \leftrightarrow C$. This contrasts with the simplistic ABC discovery model where only a single intermediate concept B is assumed between A and C.

At the initial stage, the discovery browsing technique generated a semantic predication graph. Concepts were modeled as nodes and the semantic relations between them were represented as the edges between the nodes. The graph was then built iteratively, using a user-specified concept as *seed*. From here, the algorithm searched a biomedical predication database for all predications containing the given seed concept as their subject and/or object. Once found, these predications were subsequently used to further grow the graph until the user noticed interesting patterns in the graph. Figure 10 explains this methodology.

We provide an example of paths extracted from a semantic predication graph in Figure 11. An edge between two nodes may represent one or more semantic relations between two concepts, denoted by the number label appearing on each edge. Each path is ranked according to its *degree centrality* score, which is the sum of degree centrality scores of all nodes that make up the path. Paths with high scores are considered interesting on the assumption that they connect many important nodes. In this example, a top ranking path of length 4 is the [Melatonin] $\leftrightarrow$ [Interleukin-1β] $\leftrightarrow$ [Glutamate] $\leftrightarrow$ [CLOCK] $\leftrightarrow$ [Serotonin] path, which suggests an association between the sleep hormone *Melatonin* and *Serotonin*. Serotonin is a neurotransmitter known to regulate human mood (Wilkowski *et al.*, 2011).

The main strength of the discovery browsing technique is its ability to allow users to control the growth of the semantic predication graphs (Wilkowski *et al.*, 2011), such as fine-tuning the discovery process to avoid overwhelming computing resources. The technique also leverages the existing graph-based algorithms to automatically measure the interestingness of discovery patterns, exemplified by the usage of a node's degree centrality (Freeman, 1978). Unfortunately, this technique favored two concepts that

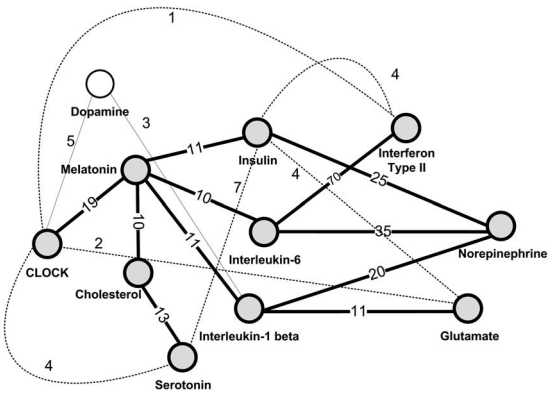

Figure 11 A partial representation of the discovery browsing graph. The image is adapted from Wilkowski *et al.* (2011)

co-occur 10 times or more in MEDLINE records during the process of extracting the semantic predication. Although applying such threshold may be useful for controlling the maximum number of predications being generated, it may also result in an LBD system incapable of finding rarer yet valuable associations.

4 Emerging literature-based discovery approaches

Having broadly considered the traditional LBD approaches, we now direct our attention to providing detailed review on various emerging LBD techniques. By 'emerging' we refer to a new generation of LBD methods, algorithms, or techniques which adopt non-traditional paradigms in solving the LBD problems. The definition emphasizes on the fundamental changes observed in these emerging techniques' approaches and is not primarily concerned with the time of their publications. Therefore, we do not consider recent works such as Preiss *et al*. (2015) and Preiss and Stevenson (2016) among the emerging LBD techniques. Although these works gave interesting studies on the effects of word sense disambiguation on the quality of LBD results, their general approach toward LBD is not very different from the traditional knowledge-based approach. Likewise, we do not consider the semi-supervised approach to learning closed chained relations proposed by Seki (2015) as an emerging approach because its approach closely resembles the discovery pattern technique (Hristovski *et al*., 2006), except that it does not require predefined semantic relations. Figure 12 shows our classification of the emerging LBD approach.

Two trends characterize the emerging LBD approach. First is the convergence of traditional statistical, knowledge-based, and visualization approaches into an integrated LBD solution (Cameron, 2014). This provides the emerging LBD techniques with a better capability in finding various latent associations in the literatures that may be too complex to be modeled using any stand-alone approach (Cameron *et al*., 2013).

The second trend is the incorporation of techniques borrowed from other research fields, such as scientometrics (Small, 2010; Kostoff, 2014), link prediction (Getoor & Diehl, 2005), machine learning (Piatetsky-Shapiro *et al*., 2006; Chang & Blei, 2010), and community detection (Newman, 2001). These new techniques offer fresh perspectives on how the LBD problem can be addressed. For example, previous research in the field of scientometrics and information retrieval have suggested that the latent relationships between documents can be inferred by studying their link structures. It is well-established that the network structure between interlinked Web pages is a rich source of information about their content and quality (Kleinberg, 1999; Brin & Page, 2012). Not only that, according to the *structural variation theory* (Chen *et al*., 2009), past instances of transformative discovery are often characterized by the formation of a number of new citation relationships between two previously disconnected groups of papers. This idea coincides with the notion of LBD (Swanson, 1990). Therefore, it is possible to ultimately view LBD as a link prediction problem between previously disjoint clusters of papers (Sebastian, 2014; Sebastian *et al*., 2015).

In the following sections, we review the techniques and performance of various emerging LBD techniques. We begin by discussing the context-driven subgraph model.

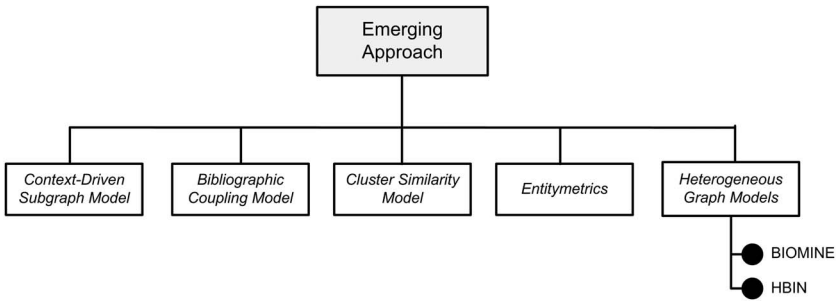

Figure 12 Classification of the emerging literature-based discovery approach. HBIN = heterogeneous bibliographic information network

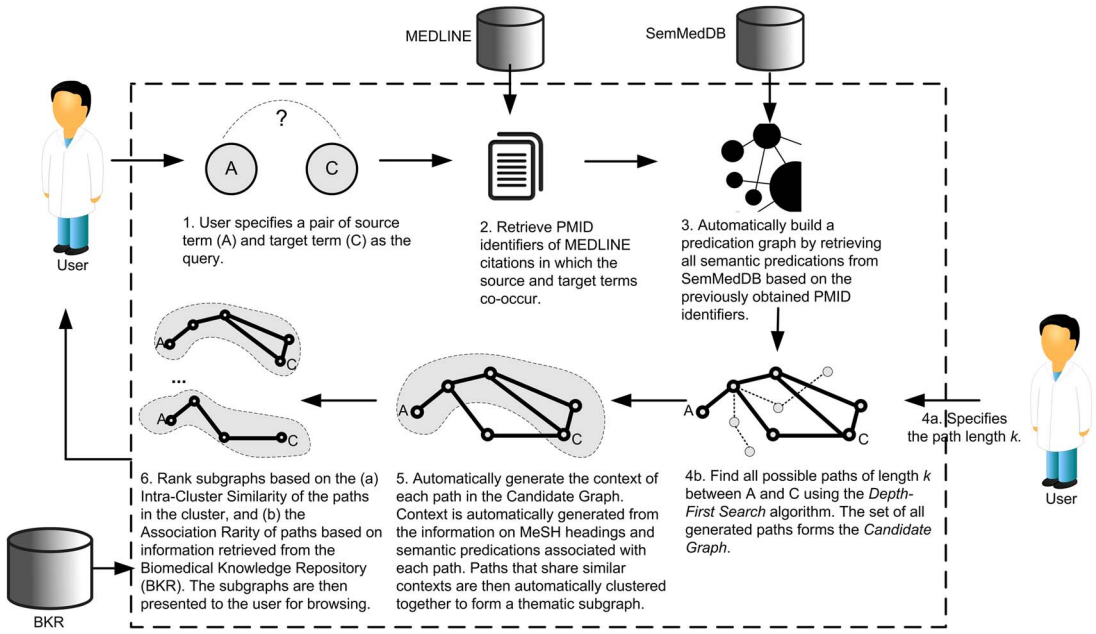

Figure 13 The steps involved in the context-driven subgraph model (Cameron *et al.*, 2015). MeSH = Medical Subject Headings

4.1 Context-driven subgraph model

The *context-driven subgraph model* combines the elements of statistical, knowledge-based, and visualization approaches into a semi-automatic LBD technique that allows users to visualize the contexts in which previously disjoint source and target terms may be connected. These contexts are modeled as semantic predication subgraphs containing thematically similar paths (Cameron *et al.*, 2013). Figure 13 shows its overall methodology.

In the beginning, the technique searched for the relevant literature to be analyzed. It then extracted the semantic predications mentioned in these literature with the help of the SemMedDB predication database[6]. These semantic predications were subsequently used to generate a directed, labeled predication graph. The subjects and objects in the extracted predications constitute the nodes in the graph, meanwhile the predicates serve as the edges between the nodes. From this predication graph and using the *Depth-First Search* algorithm (Tarjan, 1972), semantic predications connecting the source term 'dietary fish oil' to the target concept 'Raynaud's Syndrome' were identified in an attempt to reconstruct Swanson's DFORS hypothesis. Appropriate subgraphs were subsequently formed by grouping relevant predications, such that

[6] http://skr3.nlm.nih.gov/SemMedDB/

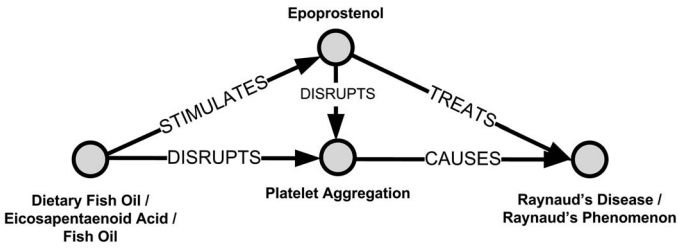

Figure 14 An example of a subgraph that portrays one aspect of the relationship between fish oil and Raynaud's syndrome (Cameron *et al.*, 2013)

each group of predications eventually represented a coherent and logical chain of associations between the source and the target concepts being studied.

A later version of the context-driven subgraph model enabled an automatic subgraph generation by clustering the semantic predication paths following a specific, predefined relatedness threshold value (Cameron *et al*., 2015). The degree of path relatedness was computed based on the information derived from the MeSH descriptors of the relevant articles where the paths originated from. Interactive visualizations of these subgraphs are made accessible via Web application *OBVIO*[7]. It is worth mentioning that recent semantic relation analysis techniques, such as *SemPathFinder* (Song *et al*., 2015), may be used to improve the efficacy of the current context-driven model by providing a more accurate semantic path analysis technique.

Subgraphs is an elegant way to represent complex relationships among entities in the literature. For example, the claim that dietary fish oil may treat Raynaud's syndrome by means of inhibiting platelet aggregation (Swanson, 1986a) can be modeled by a subgraph in Figure 14. The paths in this subgraph suggest that dietary fish oil can treat Raynaud's by stimulating the production of *Epoprostenol*. Epoprostenol, in turn, inhibits platelet aggregation, which is the primary cause of Raynaud's syndrome. From this subgraph model, one may also observe that the biological mechanism connecting dietary fish oil to Raynaud's syndrome can be quite complex. There are multiple pathways connecting both entities and the intermediate terms are connected to one another in the form of $A - B_n - C$ relation (Smalheiser & Swanson, 1996a; Wilkowski *et al*., 2011; Cameron *et al*., 2013).

4.1.1 Performance evaluation

The current technique was first applied to replicate Swanson's DFORS discovery in the closed discovery model (Cameron *et al*., 2013). Out of a total of 2124 associations in which the concept 'fish oil' served as the root of the paths, it found that 14 associations containing the 'Raynaud's' concept as their terminal. On the other hand, among the 17 848 associations where the 'Eicosapentaenoic acid' concept served as the root concept, 172 associations had the 'Raynaud's' concept as their terminal. In both scenarios, associations that were relevant to the reconstruction of Swanson's discovery had to be manually selected by the user.

Improving the previous technique, Cameron *et al*. (2015) introduced an automatic way to extract the relevant paths from a predication graph and construct the corresponding context-based subgraphs from these paths. When used to find relevant intermediate terms connecting fish oil and Raynaud's Syndrome, the technique successfully recovered blood viscosity and platelet aggregation terms but missed vascular reactivity (Cameron *et al*., 2015). This is a limitation because the traditional semantic profiling technique by Srinivasan and Libbus (2004) had previously managed to recover all three intermediate terms. Likewise, for reconstructing the migraine and magnesium hypothesis (Swanson, 1988), the technique only recovered seven out of 11 possible latent associations between migraine and magnesium (Cameron *et al*., 2015). In contrast, Srinivasan and Libbus (2004) managed to recover 10 associations.

4.1.2 Strengths and limitations

The main strength of the current technique is its ability to automatically extract thematic subgraphs. These subgraphs make it possible for users to interpret the meaning of semantic predication paths given a highly specific context. As such, the model affords much greater explanatory power compared to other previous

[7] `http://knoesis-hpco.cs.wright.edu/obvio/`

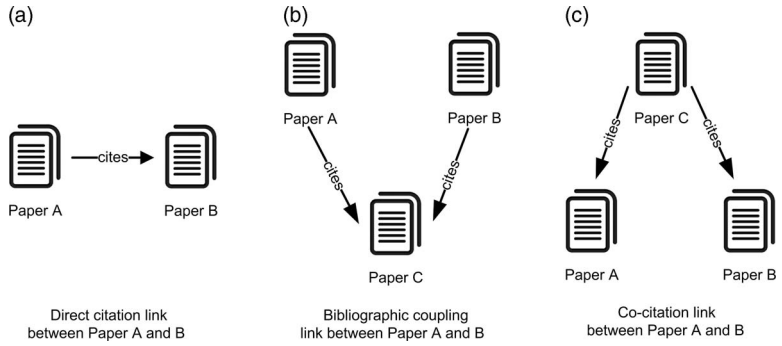

Figure 15 Various types of bibliographic link structures

LBD techniques. Although it recovered fewer number of intermediate associations than Srinivasan and Libbus (2004), this model requires less prior knowledge than the latter, making less prone to user bias.

There are some limitations. Users are still required to manually set the relatedness threshold values and the semantic predication filters. The technique also relies on the availability of MeSH descriptors in order to define the context of each subgraph. It is unclear how the model would have performed in the absence of these descriptors. Lastly, similar to Hristovski *et al*. (2006), this technique relied on third-party NLP tools to extract the semantic predications. Consequently, any lack of accuracy on the part of these tools may result in the technique's missing important latent associations (Cameron *et al*., 2015).

4.2 Bibliographic coupling technique

Since LBD is fundamentally concerned with the analysis of the content and structure of scientific literatures (Smalheiser & Torvik, 2008), the bibliographic structures of scientific literatures may reveal meaningful yet previously unknown relationships between articles, journals, authors, topics, research specialties, and even countries. Figure 15 shows three commonly studied bibliographic link structures (Boyack & Klavans, 2010). Apart from these, there are other types of bibliographic links that can be used to characterize the structures of scientific literatures, for example, author co-citation link (White & Griffith, 1981) and co-word analysis (Callon *et al*., 1983). Analyzing these bibliographic structures could help identify promising areas for generating new discoveries (Chen *et al*., 2009; Small, 2010; Nakamura *et al*., 2014).

Kostoff (2014) was the first to explore the efficacy of bibliographic coupling structures in LBD. *Bibliographic coupling* refers the sharing of references between documents (Kessler, 1963), where two documents that cite many common references are considered as strongly coupled. This technique examines these shared references between disjoint literatures to help select the most potential intermediate terms connecting the PD and CD literature (Kostoff, 2014).

Figure 16 illustrates the procedures used in this technique. It examines two groups of intermediate terms. The first group consists of common phrases from the title or abstract of disjoint literatures. The second group comprises of common phrases in the title of the *shared references* between both literatures. The research objective is to identify promising intermediate terms from the second group which are not present in the first group (Kostoff, 2014).

4.2.1 Performance evaluation

Two performance evaluations were conducted to determine the effectiveness of this technique (Kostoff, 2014). The first evaluation looks at the ability of the technique to uncover the underlying themes connecting the PD and CD literature. At first, the hierarchical clustering software CLUTO[8] was used to cluster papers that corresponded to the shared references between the PD and CD literatures. As a result, records of the same research theme were clustered together. For each research theme, a factor analysis was applied to identify significant phrases in the titles of the shared references that best represent that theme.

[8] `http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview`

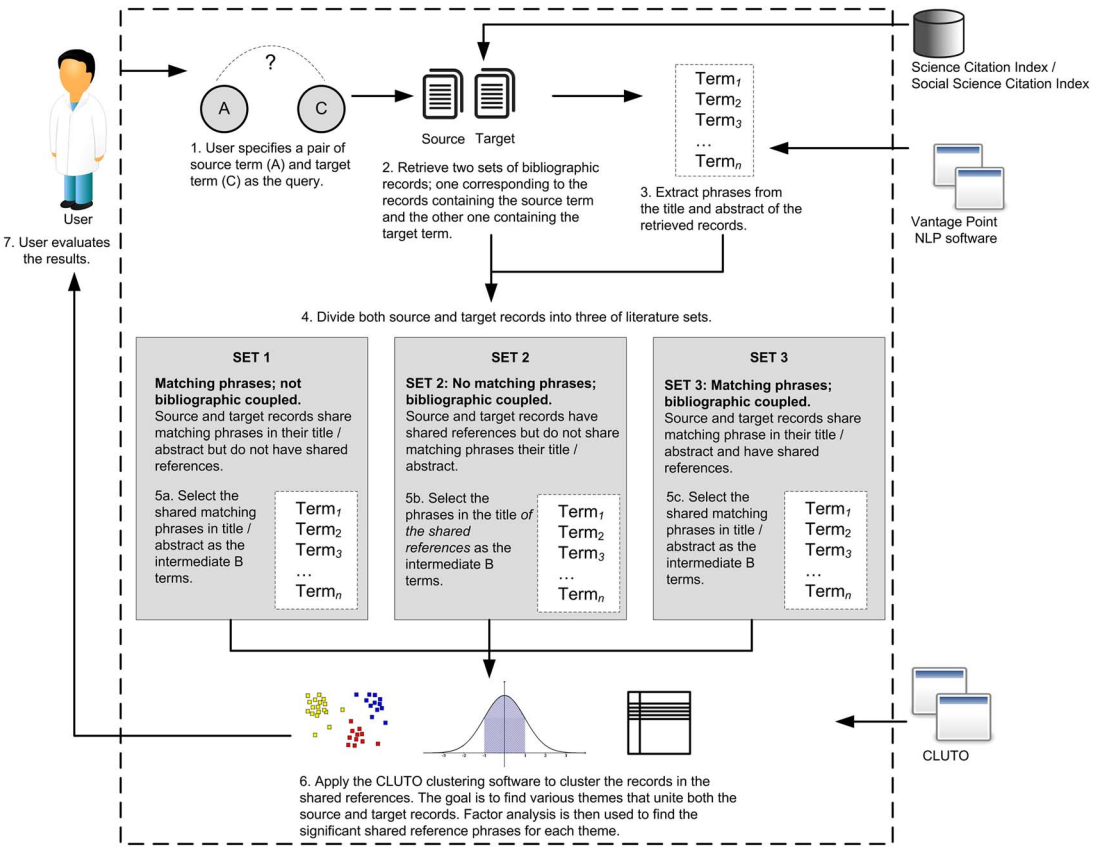

Figure 16 The steps involved in the bibliographic coupling technique (Kostoff, 2014). NLP = natural language processing

In the second evaluation, common phrases in the title or abstract of PD and CD literatures and in the title of their shared references were identified. Subsequently, the author determines whether there were novel linkages established through the phrases in the shared references, which are not found among in the title or abstract of PD and CD papers.

This technique revealed many promising concepts from the titles of the shared references between PD and CD papers (Kostoff, 2014). Specifically, it discovered three main underlying themes connecting the PD and CD literature, namely genetics, neuroimmunology, and cell death theme. It also found 1226 phrases in the titles of the shared references, of which seven new associations were considered meaningful: (1) *anthocyanins*, (2) *wogonin*, *baicalin*, *baicalein*; *Scutella rivularis extracts*, (3) *trichothecenes*, (4) *pyroptosis*, (5) *adalimumab*, (6) *cooked foods*, and (7) *ippases*.

4.2.2 Strengths and limitations

The results above strongly indicate that the shared references between two disjoint literatures could harbor many useful linking terms (Kostoff, 2014). It also suggests that a structural LBD technique that is based on bibliographic coupling, when combined with content-based analysis, could be useful in the selection of potential discovery links (Kostoff, 2012). This technique exemplifies how an emerging LBD technique begins to adopt various techniques developed in other research fields, such as computational linguistics and data mining. In this case, it uses an NLP technique to extract phrases from the titles and abstracts of records obtained from the *Science Citation Index* (SCI) and *Social Science Citation Index* (SSCI)[9], with the help of a partitional hierarchical clustering algorithm to find out the thematic structures of all references shared between PD and CD literatures.

[9] `http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery.html`

The technique's current limitation is its laborious procedure and lack of automation (Kostoff, 2014). Manual analysis needs to be performed to identify potential intermediate terms from thousands of candidates. Not only that, substantial domain expertise was needed to assess the merit of each intermediate connection between PD and CD.

4.3 Cluster similarity technique

This technique applies a text-based similarity algorithm in conjunction with a citation analysis and community detection technique (Fujita, 2012). It has been previously applied to find novel associations between literatures in *sustainability science* and *complex networks* fields (Fujita, 2012), and between *robotics* and *gerontology* fields (Ittipanuvat *et al.*, 2014).

The cluster similarity technique works by initially forming a citation network for bibliographic records downloaded from *SCI* and *SSCI* bibliographic databases, where nodes represent papers and links represent the direct citation links between the papers. Isolates or unconnected nodes were removed from the network so that only large connected components remained. The network was then partitioned into clusters of connected nodes using a modularity-based community detection algorithm (Newman, 2004).

Text cosine similarity score was then computed for all possible pairings among these clusters using the $tf-idf$ vectors of technical terms contained in the records of each cluster. Lastly, pairs of clusters which exhibited high cosine similarity scores (>0.5) were selected and the shared technical terms between these clusters were ranked based of a modified $tf-idf$ weight. The highest ranked terms were taken as the most potential intermediate terms. For instance, it was found that the term 'social network' was found to be an important term connecting the sustainability and complex network research fields (Fujita, 2012). Figure 17 describes this algorithm.

4.3.1 Performance evaluation

As many as 1630 clusters were identified from the citation network of sustainability science literature and 151 clusters from the complex network literature (Fujita, 2012). In total, 22 largest clusters were selected

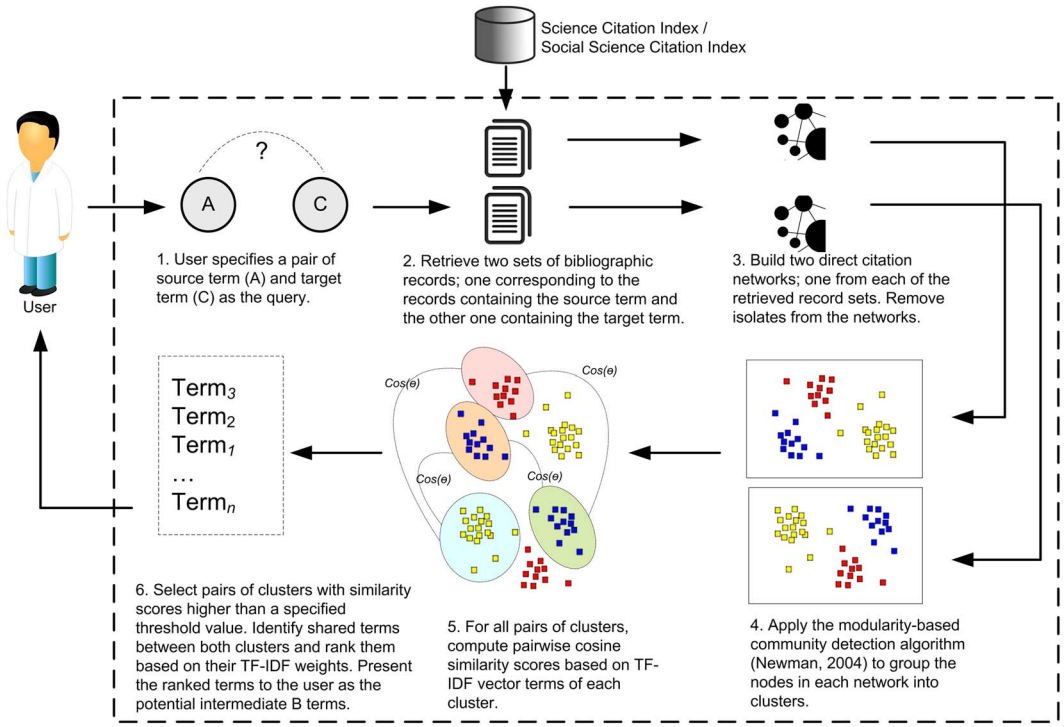

Figure 17 The steps used by the cluster similarity technique (Fujita, 2012). tf = token frequency; idf = inverse document frequency

from the former and nine largest clusters were likewise selected from the latter. Of these, three of the most textually similar pairs of clusters were chosen. Finally, the top ranking shared terms between these pairs of clusters were inspected by a domain expert to infer the possible hidden associations between the two research fields (Fujita, 2012).

Akin to Fujita, Ittipanuvat *et al*. (2014) selected pairs of textually similar clusters from the robotics and gerontology literatures. They found the cosine similarity score of $tf - idf$ vectors to be the best performing lexical statistics for identifying related clusters of papers, in comparison to other measures such as *Jaccard index*, *Dice coefficient*, and *Inclusion index* (Manning *et al*., 2008).

4.3.2 Strengths and limitations

Several cluster analytic methods have been previously proposed to discover associations between disconnected literatures. Stegmann and Grohmann (2003) and Yamamoto and Takagi (2007) used *co-word analysis* technique (Callon *et al*., 1983) to form clusters of biomedical articles in MEDLINE based on the term co-occurrence in their titles, abstracts, publication dates, and MeSH keywords. It was found that terms that linked the disconnected literatures occupied the regions of below-median centrality and density of the clusters (Stegmann & Grohmann, 2003). These methods, however, built clusters merely from term co-occurrence data and did not consider clusters formed from citation structures.

Fujita's model is different from the techniques described above. It uses citation analysis coupled with an advanced community detection algorithm instead of term co-occurrence. It also demonstrates that LBD can be successfully performed on non-biomedical research papers. Prior to this, only a few examples of non-biomedical LBD applications exist: *water purification technology* (Kostoff *et al.*, 2008), *computer science* (Gordon *et al*., 2002), *chemistry* (Valdés-Pérez, 1999), and *humanities* (Cory, 1997).

Having said so, the current technique is limited by the use of only one type of bibliographic link, that is direct citation link. Studies in scientometrics have shown that the relationships between documents can be represented by more than one types of citation links, for example, bibliographic coupling and co-citation links (Janssens *et al*., 2008; Boyack & Klavans, 2010; Waltman & Eck, 2012; Boyack *et al*., 2013). Therefore, it is possible that better discovery capabilities and more accurate cluster link prediction results can be attained through combined applications of diverse types of bibliographic links.

4.4 Entitymetrics technique

Ding *et al*. (2013) introduced *entitymetrics*, a network-based data representation that combines biological knowledge entities (e.g. diseases, drugs) with citation information. Assuming that paper *A* cites paper *B*, artificial *entity citation* links are drawn between each knowledge entity mentioned in paper *A* and each knowledge entity mentioned in paper *B*. This feature provides the model with richer information that can be used to predict latent associations between two papers. Figure 18 gives an example of the entitymetrics graph.

Besides predicting drug–disease interactions for *Metformin* (Ding *et al*., 2013), the entitymetrics method has been used to predict gene–gene interactions (Song *et al*., 2013). Its methodology consists of four major steps (Ding *et al*., 2013), as shown in Figure 19.

Biological entity extraction: At first, user selects a source term. Bibliographic records containing the source term were then downloaded from PubmedCentral[10]. From these records, biological entities were extracted from their title, abstract or full-text using a dictionary-based named entity recognition algorithm. The dictionary includes vocabularies indexed in various biomedical knowledge bases, for example, DrugBank[11], HUGO (Human Genome Organization) database[12], and Comparative Toxicogenomics Database (CTD)[13].

[10] `http://www.ncbi.nlm.nih.gov/pmc/`
[11] `http://www.drugbank.ca/`
[12] `http://www.genenames.org/`
[13] `http://ctdbase.org/`

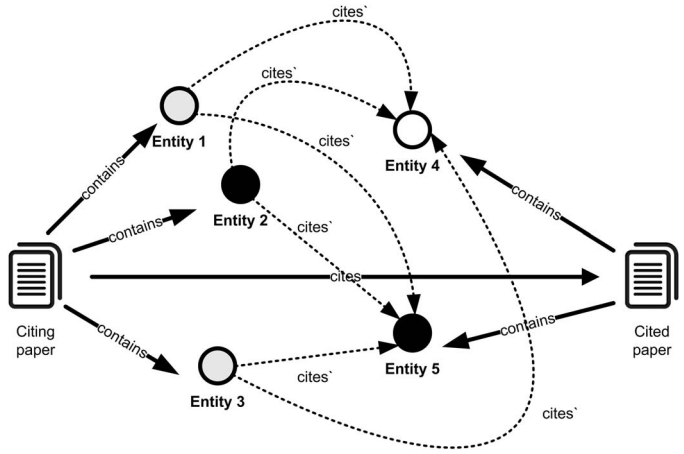

Figure 18 A partial depiction of an entitymetrics graph (Ding *et al.*, 2013)

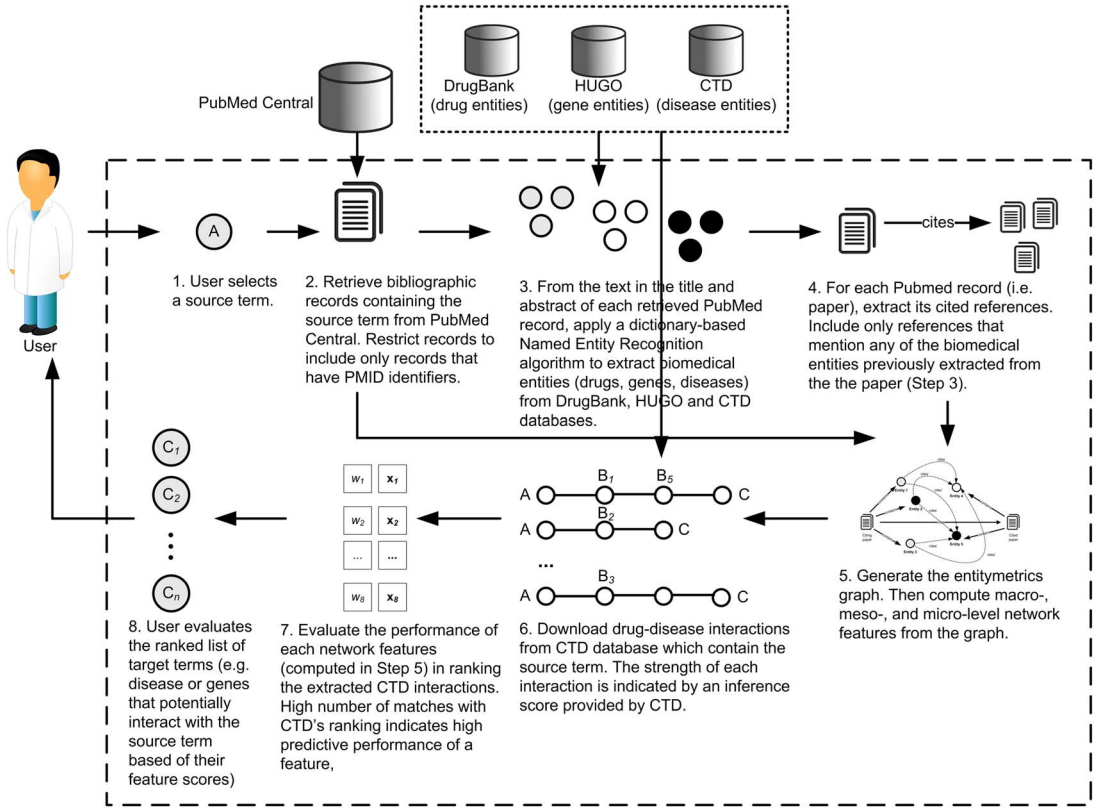

Figure 19 Methodology of the entitymetrics technique (Ding *et al.*, 2013). CTD = Comparative Toxicogenomics Database

Bio-entity citation network construction: The entitymetrics graph was then built from the previously extracted biological entities. As previously explained, if paper *A* cites paper *B* then citation links were drawn connecting each biological entity in paper *A* to each biological entity in paper *B*.

Feature construction: Next, various network-based features were calculated from the entitymetrics graph. The features included *macro*-level features (*bi-component analysis*, *K-core analysis*, *mean shortest path between node pairs*, *degree distribution*), *meso*-level features (*coefficient clustering*), and *micro*-level features (*degree centrality*, *closeness centrality*, *betweenness centrality*). All of these features represent cluster-level and node-level features of the network.

Prediction: The constructed features were finally used to predict drug–disease interactions for a drug *Metformin*. The prediction results were evaluated against the existing drug–disease interactions indexed in the CTD database.

4.4.1 Performance evaluation

The results showed that the features, especially centrality measures, can be used to predict the existing drug–disease interactions for Metformin in the CTD database with certain amount of success (Ding *et al*., 2013). Out of 697 diseases ranked in CTD for this drug, 16 matches were recovered using the in-degree centrality feature. Three of the matches were among the top 10 diseases ranked in CTD for Metformin. Other features such as the out-degree centrality features recovered 16 disease matches (lowest rank: 439th), the closeness centrality feature found 13 matches (lowest rank: 439th), and the betweenness centrality feature found 17 matches (lowest rank: 439th).

4.4.2 Strengths and limitations

This model is interesting as it represents a step toward a highly seamless integration of scientometrics techniques, knowledge-based techniques, and network analysis in a single LBD framework. Unlike Fujita's technique that combines the citation and lexical analysis in two consecutive but separate stages (Fujita, 2012), entitymetrics integrates knowledge-based entities (drugs, diseases, genes) with bibliographic entities (papers) in the same bio-entity citation network. This is advantageous because various network-based features can then be computed from the same network. Furthermore, given the large volume of research papers it was capable of analyzing, this technique suggests the promising scalability of a network-based LBD technique.

The performance evaluation results of this technique seemed to suggest that it suffers from a low recall rate with the ability to only recover less than 20 diseases for Metformin out of nearly a total of 700 diseases ranked for it in the CTD database (Ding *et al*., 2013). Due to its strong coupling with domain-specific knowledge bases, such as HUGO, CTD, and DrugBank, the effectiveness of this technique may only be limited to situations where such resources are available.

4.5 BIOMINE

The BIOMINE model views LBD as a link prediction problem. Link prediction is 'the problem of predicting the existence of a link between two entities, based on attributes of the objects and other observed links' (Getoor & Diehl, 2005). For LBD, the goal is to predict the future links between previously disjoint literatures (Chen, 2012; Sebastian, 2014). The predicted links may refer to new co-citation links that connect two previously disconnected clusters of papers in a co-citation network (Chen *et al*., 2009). Alternatively, they may represent novel word co-occurrence between previously unrelated concepts (Yetisgen-Yildiz & Pratt, 2009). To achieve these objectives, an LBD system may require automatically learning the most predictive features from large literature data sets. The learned features can then be used for predicting future links between concepts, individual papers, or clusters of research papers.

This link prediction paradigm inspired the BIOMINE technique (Eronen & Toivonen, 2012). The model constructs a biological heterogeneous weighted graph that can be used to model various biological relationships (protein interactions, gene–disease associations, gene ontology annotations). BIOMINE's main objective is to predict which biological concept nodes will be connected in the future given the present biological data.

Figure 20 illustrates the steps in the BIOMINE system. The first step weighted each edge in the graph according to the probability of it representing an actual biological relationships. This is determined based on three criteria: *relevance*, *informativeness*, and *reliability* (Equation (2)). Relevance, denoted by $q(e)$, refers to the relative importance of the relationship that is being represented by the edge. Its score ranges from 0 to ∞. The informativeness of an edge, denoted by $i(e)$, is calculated as the degrees of the nodes it connects to. Reliability, denoted by $r(e)$, measures the confidence that the relationship represented by the edge actually exists. This is determined by the reliability value obtained from the STRING protein–protein

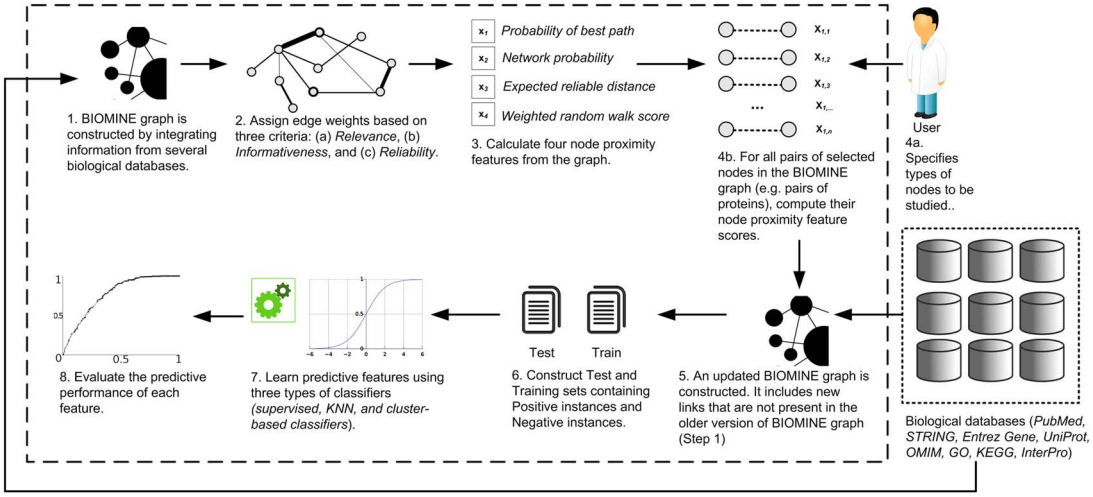

Figure 20 The algorithm of the BIOMINE system (Eronen & Toivonen, 2012)

interaction database[14]. Thus, the total weight of an edge e, $p(e)$, is the product of the scores of these three criteria:

$$p(e) = q(e) \cdot i(e) \cdot r(e) \tag{2}$$

Subsequently, four node proximity features were extracted from the BIOMINE graph to facilitate the link prediction tasks. The first feature is the *probability of best path*. Path probability is the product of all $p(e)$ of edges along a path. Consequently, assuming several paths connecting a source node to a target node, this feature selects path that has the highest probability score. The second feature is the *network probability*. Given a source node s and a target node t, this feature calculates the probability that a randomly generated subgraph of the overall graph will contain a path connecting s and t. The third feature is the *expected reliable distance*. It measures the expected shortest-path distance in all randomly generated subgraphs in which a path exist between s and t. The fourth feature is the weighted version of *standard random walk stationary distribution score*.

In the experiments, BIOMINE addressed two predictive goals: (a) predicting protein interactions that will be added to the Entrez Gene database[15], and (b) predicting pairs of genes that will affect the same disease. Training and test sets were constructed for a binary classification task and the data sets consist of positive and negative instances. The class assignment of these instances was determined by comparing an old BIOMINE graph with its updated version. Positive instances are pairs of nodes in the graph that were not linked in the older graph but which became linked in the updated BIOMINE graph. In contrast, negative instances are disconnected pairs of nodes that remain unlinked.

4.5.1 Performance evaluation

BIOMINE demonstrated a good accuracy in predicting novel disease–gene associations (Eronen & Toivonen, 2012). For task that predicts future protein interactions in the Entrez database, the random walk feature emerged as the best predictor with the *area under curve* equals to 0.82. For the disease–gene prediction task, the random walk feature also emerged as the most predictive feature.

Further, using the random walk feature as the node proximity measure, the performance of three classifiers were compared in predicting disease–gene associations: a *supervised classifier*, the *K-Nearest Neighbour (KNN)* algorithm, and a *cluster-based classifier* (Eronen & Toivonen, 2012). The results showed that the supervised classifier and the cluster-based classifier outperforming the KNN classifier.

[14] http://string-db.org/
[15] http://www.ncbi.nlm.nih.gov/gene

4.5.2 Strengths and limitations

The most salient characteristic of this technique is its utilization of a heterogeneous biological graph. The information richness of this graph allows one to generate a strongly predictive feature, such as the random walk feature. The technique was also scalable to a database with 1.1 million concepts and 8.1 million relations (Eronen & Toivonen, 2012). More importantly, this technique used both supervised and unsupervised machine learning algorithms, where the encouraging performance of the random walk feature could motivate further explorations into the use of machine learning techniques in future LBD systems.

In view of a general purpose LBD system, this technique is limited in that it is specifically tailored for discovery tasks in genomics and proteomics. It also appears to heavily depend on the availability of certain knowledge bases to operate successfully. For instance, the reliability score can only be calculated on if the reliability values are available from the STRING database. For domains where such databases do not exist, this technique may not be directly applicable.

4.6 Heterogeneous bibliographic information network (HBIN) technique

Sebastian *et al*. (2015) proposed a technique that builds predictive features from metapaths found in a HBIN. Using the information mined from a HBIN graph, this model aims at predicting future co-citation links between disparate groups of research papers. Similar to BIOMINE (Eronen & Toivonen, 2012), the authors view LBD as a link prediction problem. Figure 21 details the algorithm of this technique.

A heterogeneous information network is a directed graph consisting of multiple-typed objects and links (Sun & Han, 2012). More specifically, HBIN is a special type of heterogeneous information network that allows one to model the rich interactions between various types of bibliographic entities.

HBIN consists of four types of bibliographic entities: *paper*, *author*, *venue* (i.e. journal or conference), and *term* (i.e. those found in the titles, abstracts, subject headings, and full text of papers). The paper entity is further subdivided into *core paper*, *citing paper*, and *reference paper* entities. Core papers are papers that belong to either side of the disconnected literature. For example, in the case of Swanson's DFORS

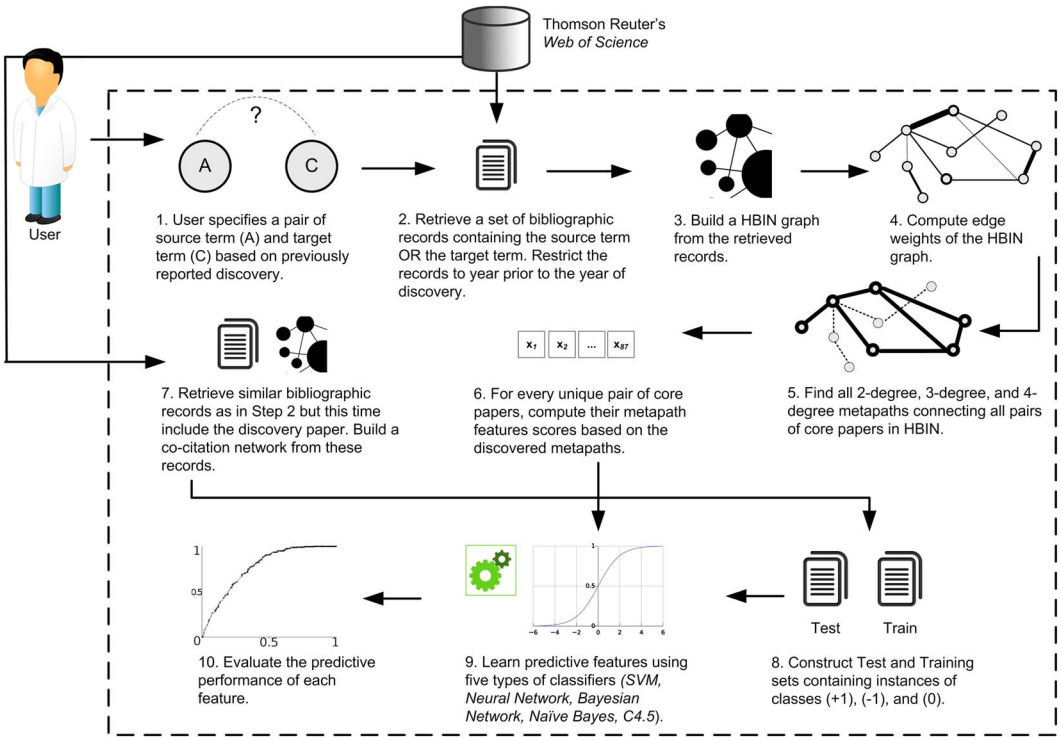

Figure 21 The algorithm used by the heterogeneous bibliographic information network (HBIN) technique (Sebastian *et al.*, 2015)

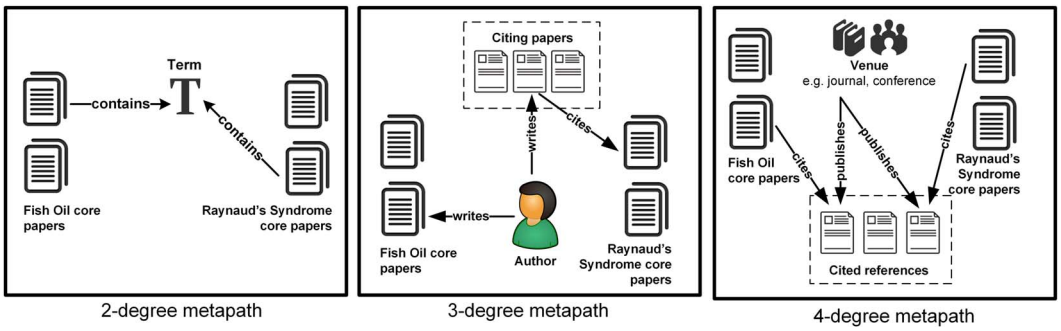

Figure 22 Examples of metapaths that can be found in the heterogeneous bibliographic information network (Sebastian *et al.*, 2015). The degree of a metapath is the number of edges that separate one core paper from the other

hypothesis, papers that belong to either the fish oil or Raynaud's syndrome literature (but *not* both) are the core papers. The information about each core paper includes its title, author(s), publication venue, its reference papers, and its citing papers. Citing papers are papers that *cite* a core paper, whereas reference papers are papers which are *cited* by a core paper. A core paper may therefore be associated with zero or more citing papers, as well as zero or more reference papers.

The entities in HBIN graphs are connected via various types of links (Sebastian *et al*., 2015). These include citation links between papers, authorship link between an author and a paper, publication link between a paper and the venue that publishes it, and semantic link between a paper and the term it contains. Composite links known as *metapaths* can then be constructed by adjoining these various types of link. The overarching idea is that although two core papers are initially disjoint (i.e. do not have any article in common, have never cited each other, and have never been co-cited) (Swanson, 1987), their hidden connections could be inferred from the latent connections via these metapaths in the HBIN graph.

Two-degree, three-degree, and four-degree metapaths could be extracted from the HBIN graph (Sebastian *et al*., 2015). Figure 22 gives examples of HBIN metapaths. Based on these path configurations, a number of different metapath-based features could then be constructed and used to predict future co-citation links between two previously disjoint core papers.

4.6.1 Performance evaluation

The model assumes the existence of a certain *discovery paper*, such as Swanson (1986a, 1988). A discovery paper provides the starting point for this model to retrospectively reconstruct a LBD scenario. For example, to reconstruct Swanson's DFORS discovery, the technique started by downloading a total of 485 core papers along with their abstracts (352 on fish oil; 133 on Raynaud's syndrome) from Thomson Reuter's *Web of Science*[16]. These core papers were identified using the same search keywords originally used by Swanson (1986a) and they were to be published prior to Swanson's discovery, that is between 1900 and 1985. A corresponding HBIN graph was then constructed using these bibliographic data, where each edge in the HBIN graph was weighted according to a specific weighting score (Sebastian *et al*., 2015). Based on these scores, 87 different metapath features were then computed from the HBIN graph.

This technique views link prediction in LBD as akin to solving a multiclass classification problem. The performance of the learned metapath features was studied using five popular machine learning classifiers: the *Sequential Minimal Optimization (SMO) variant of the Support Vector Machine algorithm*, *Neural Networks*, *Naive Bayes*, *Bayesian Network*, and *C4.5 Decision Tree* (Witten & Frank, 2005). To construct the training and test sets, the technique collected 117 370 unique core paper pairs from the 485 core papers retrieved previously. It then applied an exhaustive algorithm to search for all distinct metapaths between all unique core paper pairs. To assign a class label to each instance, the technique retrieved another set of fish oil and Raynaud's syndrome

[16] `http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/web-of-science.html`

records from the *Web of Science*, including papers published in 1986 following Swanson's discovery. A pair of core papers is labeled as class +1 if it consisted of a pair of fish oil paper and Raynaud's Syndrome paper that had no co-citation link before 1986 but which became co-cited in 1986. A pair was labeled as −1 to represent papers from the same research area (i.e. either fish oil or Raynaud's syndrome literature) that became co-cited in 1986. Lastly, a pair was labeled as 0 if they never became co-cited.

In reconstructing the DFORS discovery, the HBIN metapath features performed very well in predicting future inter-cluster co-citation links (+1) between fish oil and Raynaud's syndrome papers, with 0.851 *F*-Measure (precision: 0.845, recall: 0.857). This performance was better than the performance of other document similarity features such as bibliographic coupling similarity (0.612 *F*-Measure), LDA topic similarity (0.608 *F*-Measure), and TF–IDF similarity (0.457 *F*-Measure) (Sebastian *et al*., 2015). Further, the authors found that using features constructed from the four-degree author-sharing and term-sharing metapaths alone produced a better predictive performance (71.40 and 70.64% accuracy rate, respectively).

In a subsequent experiment to reconstruct Swanson's migraine and magnesium discovery (Swanson, 1988), the HBIN model also performed well in predicting co-citation links between migraine and magnesium literatures, with 0.80 *F*-Measure (precision: 0.80, recall: 0.80) (Sebastian *et al*., 2017). Importantly, the study found that the predictive accuracy of the model improved when it incorporated elements of topic modeling and word sense disambiguation techniques. This outcome is consistent with Preiss and Stevenson (2016), who have observed that using appropriate word sense disambiguation could positively enhance the performance of an LBD system.

4.6.2 Strengths and limitations

Unlike BIOMINE and other LBD techniques, HBIN's efficacy does not depend on the availability domain-specific knowledge-based sources. In the absence of these resources, it managed to perform considerably well in predicting which papers will form future co-citation linkages. This is probably because the HBIN allows much richer information to be exploited by the LBD algorithm compared to if a homogeneous network is used (Sun & Han, 2012). The meta-structure of the HBIN graph (i.e. the interconnections between bibliographic entities) could reveal the latent semantic relationships between papers from two separate literature.

Another distinct feature of this technique is its seamless integration of both lexical and citation information into a single unified graphical representations. This approach overcomes the limitations of a purely citation analysis method, such as bibliographic coupling (Kessler, 1963; Kostoff, 2014), as well as the limitations of purely traditional lexical analysis techniques, such as *latent Dirichlet allocation* (Blei *et al*., 2003) and $tf-idf$ (Salton & McGill, 1986). Citation analysis is known to yield high precision but low recall retrieval whereas lexical analysis tends to give high recall and low precision results (Bassecoulard & Zitt, 2004). By incorporating both citation information and lexical information in the HBIN metapath features, the more balanced precision and recall rates could be achieved.

There are several limitations of the current model. It has not completely addressed the scalability issue of the algorithm, especially its path enumeration function (Sebastian *et al.*, 2017). As such, it may be difficult to extend the current model implementation on a much larger data set. Lastly, the model did not perform author name disambiguation when constructing the HBIN graphs. The extent to which this deficiency affects the model's overall accuracy remains unknown.

4.7 Future trends in emerging literature-based discovery approaches

In short, the emerging LBD approaches have brought about new ways for addressing LBD problems. We have summarized the key aspects of these approaches in Table 1. Unlike the traditional LBD approaches, they are characteristically more data-driven, evident from the increasing adoption of machine learning algorithms to automatically learn predictive LBD features (Eronen & Toivonen, 2012; Sebastian *et al.*, 2017). The emerging approach is also more interactive and could provide better explanations for the discovered associations (Cameron *et al.*, 2015). Not only that, these LBD techniques have been designed to scale better against hundreds of thousands of documents and tackle millions of biomedical concepts and semantic relations (Eronen & Toivonen, 2012; Ding *et al.*, 2013).

Table 1 The performance summary of techniques under the emerging literature-based discovery approach

Techniques	Authors	Data sets	Performance
BIOMINE	Eronen and Toivonen (2012)	Gene–disease associations	ROC (Receiver Operating Characteristic) curve's AUC: 0.82
Entitymetrics	Ding *et al.* (2013)	Drug–disease associations	Precision: 0.30
	Song *et al.* (2013)	Drug–disease associations	Precision: 0.36
Cluster similarity	Fujita (2012)	Sustainability	Expert evaluation
	Ittipanuvat *et al.* (2014)	Robotics	Expert evaluation
Bibliographic coupling	Kostoff (2014)	Parkinson's disease	Found 7 meaningful intermediate terms out of 1226 phrases
Context-driven subgraph model	Cameron *et al.* (2013)	DFORS	For *fish oil* term, 14/2124 paths led to *Raynaud's* For *eicosapentaenoic acid*, 172/17 848 paths led to *Raynaud's*
	Cameron *et al.* (2015)	MM, somatomedin C, Alzheimer's disease, schizophrenia, cardiac hypertrophy, hypogonadism	For MM, recovered 7 out of 11 latent associations discovered by Swanson
HBIN model	Sebastian *et al.* (2015)	DFORS	*F*-Measure: 0.851
	Sebastian *et al.* (2017)	MM	*F*-Measure: 0.80

AUC = area under curve; DFORS = dietary fish oil–Raynaud's syndrome; MM = migraine–magnesium; HBIN = heterogeneous bibliographic information network.

In terms of LBD evaluation methodology, a new trend is also emerging. Even though the majority of LBD methods still rely on replicating Swanson's hypotheses for validation (Yetisgen-Yildiz & Pratt, 2009), an increasing number of non-conventional evaluation methods have been introduced, as also shown in Table 1. For instance, Ding *et al.* (2013) evaluated the effectiveness of their entitymetrics model by detecting drug–disease associations in the CTD database. Likewise, there is also a move toward using non-biomedical data sets for evaluations, as demonstrated by Fujita (2012) and Ittipanuvat *et al.* (2014).

For other future trends, we foresee a growing number of emerging LBD techniques that will be strongly driven by the utilization of advanced machine learning algorithms, network-centric techniques, and interactive visualizations. Furthermore, we also anticipate more effort will be devoted to novel ways for automatically constructing evaluation gold standards. We discuss these points below.

4.7.1 Scalable machine learning-driven literature-based discovery techniques

In the past, most LBD techniques have been developed by information scientists, digital librarians, medical researchers, and biologists (Weeber *et al.*, 2005; Bekhuis, 2006; Jensen *et al.*, 2006; Mostafa *et al.*, 2009; Hahn *et al.*, 2012). As a consequence, most LBD systems tend to be biased toward domain-oriented solutions, especially biomedical applications. Notably, they required extensive incorporation of domain background knowledge, as evident from the rampant usage of UMLS vocabularies, biological databases, and specially designed NLP tools by the majority of traditional LBD techniques. As a result, their applications are limited to certain fields only (Marsi *et al.*, 2014).

It is likely that future research trend could see more machine learning-oriented techniques capable of automatically learning useful features from a collection of literature. The learned features would be used to predict the existence of hidden connections between disjoint research areas, even in the absence of available background knowledge. In terms of scale, given recent advances in Big Data technologies, future LBD techniques may also incorporate algorithms that scale well against very large data sets. This trend would also reduce LBD systems' reliance on heuristics-based rules and domain-specific knowledge bases, making it easier to generalize them to relevant applications in various research fields.

4.7.2 Network-centric algorithms

Since the relationships among papers can naturally be modeled as interconnected nodes in complex networks (Newman, 2001), the current developments in community detection and network data mining

research may inform the future designs of LBD systems (Newman, 2003; Leskovec *et al.*, 2005, 2010). As a result of this network-centric view, there will be an increasing number of link prediction-oriented LBD algorithms, such as a recent algorithm for predicting co-occurrence associations in the MeSH co-occurrence network (Kastrin *et al.*, 2013). Next generation LBD approaches are also likely to integrate network analytic algorithms in their algorithms (Leskovec *et al.*, 2010), especially given the availability of existing open source network analytics packages such as *GraphX*[17], *NetworkX*[18], and the *Stanford Network Analysis Project*[19].

4.7.3 Better visualizations

Future LBD research may also integrate more advanced visualization methods to support LBD activities (van Mulligen *et al.*, 2002; Wilkowski *et al.*, 2011; Goodwin *et al.*, 2012; Cameron *et al.*, 2013). Smalheiser (2012) argued that the success of LBD systems should be measured based on how well they support researchers in their daily scientific endeavors. Therefore, there would be needs to create semi-automatic, highly usable, and visually attractive LBD systems that would integrate seamlessly with the users' workflows.

4.7.4 Automatic identification of evaluation gold standards

In terms of new evaluation gold standards, a possible future trajectory may see new methods capable of automatically collecting instances of scientific discoveries that have exhibited strong inter-cluster linkage properties (Chen *et al.*, 2009; Chen, 2012). In addition to minimizing the subjective element in the current gold standard selection process, this would reduce the costs associated with using domain expert evaluation. Furthermore, to address the difficulties in finding good samples of discovery papers, researchers have the option to collect evaluation data sets based on real world discoveries in fields such as medicine or physics. For example, the *Journal of American Medical Association* has identified 51 landmark medical papers (Meyer & Lundberg, 1985), whereas the *Physical Review Letters* has recognized 83 milestone physical papers[20]. Given their lasting contributions of these selected papers to their fields, future studies may focus on understanding the extent to which these papers may serve as good evaluation ground truth for LBD systems.

5 Future research areas

We have identified five immediate research problems to be addressed in future LBD research. We elaborate them in this section.

5.1 Measuring the interestingness of literature-based discovery outcomes

The first research area should consider how to accurately predict *meaningful* novel associations between the disjoint concepts. As this review shows, a common approach is to rank a list of candidate intermediate or target terms based on specific interestingness measures (Wren, 2008). User then examines which terms would yield the most interesting novel associations. Because manually examining the ranked terms is a tedious and time-consuming task, previous LBD algorithms addressed this problem either by exploring various interestingness measures that can be used to automatically filter uninteresting associations from the final LBD results (Torvik & Smalheiser, 2007), or by employing visualization techniques to help users to easily identify interesting associations (Wilkowski *et al.*, 2011).

There are considerable number of research problems to be pursued in this area. One possible research direction is to develop objective interesting measures that can be fine tuned to fit different types of knowledge discoveries (Smalheiser, 2012). Most of the existing interestingness measures have focused

[17] `https://spark.apache.org/graphx/`
[18] `https://networkx.github.io/`
[19] `http://snap.stanford.edu/`
[20] `http://prl.aps.org/50years/milestones`

only on finding frequently co-occurring terms (Wren, 2008) or semantically similar terms (Smalheiser, 2012). However, it is also possible that the less frequently co-occurring terms to harbor novel connections between disjoint literatures (Kostoff *et al.*, 2009; Petrič *et al.*, 2010), such that new interestingness measures are needed to capture this type of discovery.

5.2 Evaluating the performance of literature-based discovery systems

As suggested previously, another immediate research problem concerns the deficiency in the current evaluation methodologies. Currently, a comprehensive set of evaluation gold standards and the consistent evaluation metrics do not exist (Yetisgen-Yildiz & Pratt, 2008; Smalheiser, 2012). Rather, most evaluations typically involved replicating the historical LBD discoveries (Yetisgen-Yildiz & Pratt, 2009). Although widely used even among the most recent LBD systems (Cameron *et al.*, 2015; Novacek, 2015; Song *et al.*, 2015; Sebastian *et al.*, 2017), this evaluation approach could risk overfitting an LBD system. Another common method is to get domain experts to evaluate LBD systems' results on *ad hoc* basis (Weeber *et al.*, 2003; Srinivasan & Libbus, 2004). Alternatively, the experts may be tasked with formulating a new set of queries that might have potential LBD outcomes (Gordon *et al.*, 2002; Torvik & Smalheiser, 2007). Both approaches are costly and are prone to user bias.

To address this issue, new evaluation methods are needed. Yetisgen-Yildiz and Pratt (2009) proposed using future co-occurrence between terms that have never been co-mentioned in the literature to provide a more objective evaluation standard. The performance of LBD systems can then be evaluated based on how accurate they are in predicting the future co-occurrences between these terms from time-sliced data sets. Unfortunately, as pointed by Kostoff (2007), the term co-occurrence measure is a poor proxy for true scientific discoveries. A recent alternative evaluates domain-specific LBD systems against existing human-curated biomedical databases. For instance, Ding *et al.* (2013) judged their algorithms based on how many drug–disease interactions in the CTD can be successfully predicted. This approach is highly domain-dependent and may not be applicable to domains where such databases do not exist.

Future research should also look into building a consensus on LBD evaluation metrics. The absence of a well-accepted evaluation ground truth naturally leads to the lack of consistent evaluation metrics. The effectiveness of many past LBD systems are usually determined by their ability to recover past discoveries (such as Swanson's discoveries), without evaluating the rest of their outputs (Yetisgen-Yildiz & Pratt, 2009). Information retrieval metrics such as the precision, recall, and mean average precision have been used (Yetisgen-Yildiz & Pratt, 2006; Torvik & Smalheiser, 2007), but Kostoff *et al.* (2009) contended that such quantitative metrics cannot sufficiently account for the quality of a discovery. Instead, Kostoff *et al.* argued for a more rigorous vetting process to rule out any possible previous work related to a claimed discovery. Nevertheless, the suggested vetting procedure is likely to be a highly time-consuming task and has not seen a significant following.

5.3 Increasing the scalability of literature-based discovery systems

Wren (2004) found that the network of co-occurring terms in scientific literature, such as MEDLINE, exhibited a *small world structures* in which most nodes in a network are strongly interconnected to each other. Coupled this with the current exponential growth of scientific literatures (Larsen & Von Ins, 2010; Bornmann & Mutz, 2015), the scalability issue of the traditional LBD approaches remain an important problem to be addressed. For example, with just a single hop from the source term A to the intermediate term B, it is common to find an explosion in the number of $A-B$ associations that need to be analyzed by an LBD system (Wren, 2008; Smalheiser, 2012).

A possible strategy is to eliminate the need to evaluate the intermediate terms B by solving an open discovery problem as multiple closed discovery problems (Smalheiser, 2012). Another way is to directly measure the strength of the association between A and C without having to evaluate the strength of the $A-B$ associations, similar to the distributional LBD approach (Gordon & Dumais, 1998; Cohen *et al.*, 2014). Even in this case, the number of candidate target object C may still be very large. Thus, inventing scalable LBD algorithms is a priority research area.

5.4 Encouraging domain independence

In the past, simple statistical measures have been shown to be insufficient for modeling highly complex nature of the relationships between two disjoint concepts (Wren, 2008). For instance, Gordon *et al.* (2002) conducted experiments where lexical statistics are used to search for novel applications of genetic algorithms from World Wide Web documents, but their results showed that the technique failed to produce meaningful results, especially in the absence of substantial amount of user interventions and input.

On the other hand, knowledge-based approaches owe their effectiveness to the use of specific biomedical knowledge sources, such as MeSH vocabularies and UMLS semantic relations (Weeber *et al.*, 2001; Srinivasan, 2004; Hristovski *et al.*, 2006) or third-party NLP software, for example, *SemRep*[21] (Wilkowski *et al.*, 2011; Miller *et al.*, 2012; Cohen *et al.*, 2015). These eventually limit the applicability of these techniques to a wider range of research literatures (Symonds *et al.*, 2014).

The next important research area therefore is to design LBD algorithms whose effectiveness does not heavily rely on the availability of domain-specific knowledge sources (Marsi *et al.*, 2014). DARPA's Big Mechanism initiative (Cohen, 2015) encourages a greater usage of LBD systems for uncovering complex scientific mechanisms from diverse literatures. In addition to methods such as the HBIN model (Sebastian *et al.*, 2017), a work such as an automatic extraction of variable terms from non-biomedical literatures (Marsi &Öztürk, 2015) is a good example of the initial steps toward this direction.

5.5 Improving user acceptance

The final noteworthy research problem is to increase the acceptance and usage of LBD systems by researchers. To date, the ARROWSMITH[22] system is arguably the most popular and well-maintained LBD system. It is available as an online Web application and is relatively easy to use, although it has been previously reported to have only about 1200 unique users monthly (Li *et al.*, 2014). Certainly, having more users is desired in order to fully realize the potential and benefits of LBD systems in contemporary scientific practices (Smalheiser, 2012).

Smalheiser and Torvik (2008) observed that the problem with low user acceptance may have originated from the lack of proper understanding as to the success of LBD systems ought to be determined. Rather than defining its success by how true its outputs are, the authors argued that the merit of an LBD system should be judged instead by how well it seamlessly supports contemporary scientific practices in a way similar to how the PubMed search engine supports scientists. Even so, increasing the popularity and the rate of LBD systems' adoption in most actual scientific settings may still be hampered by the reluctance of the scientists in giving credit to an LBD system for its contributions to their scientific achievements (Smalheiser, 2012).

Smalheiser (2012) suggested two possible solutions. First, the LBD community needs to help scientists recognize the occasions during which they actually carry out LBD-like investigations in their day-to-day practices. Doing so will help them recognize the value of LBD systems as an integral part of their research activities. Second, it will be useful to increase the visibility of LBD systems to the public. For example, this can be achieved by integrating LBD functionalities to popular search engines such as PubMed[23]. Unfortunately, we have yet to witness an actual study or practical implementation of these suggested ideas.

6 Conclusion

LBD techniques have evolved from traditional approaches that primarily rely on the utilization of lexical statistics and knowledge-based techniques, to the more sophisticated emerging approaches. In this paper, we have reviewed the technical and performance evaluations of these emerging techniques. Our review has shown that, in contrast to most traditional LBD approaches, the fundamental paradigm shifts have occurred among the emerging approaches that involve the increasing adaptation of various techniques

[21] `http://semrep.nlm.nih.gov/`
[22] `http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html`
[23] `http://www.ncbi.nlm.nih.gov/pubmed`

originating from non-LBD research fields, such as graph theory, scientometrics, link prediction, and machine learning research. These techniques also employ richer forms of information representations in the form of entitymetrics, biological graphs, and HBIN. The new data structures and representational models harness both textual and non-textual features in the literatures for finding the implicit connections between disjoint sets of literature. Experimental results have shown the emerging LBD approach outperforming the traditional approaches in terms accuracy, comprehensibility, scalability, and interactivity.

Nonetheless, some research challenges remain to be addressed by future research. They include challenges associated with the scalability of LBD algorithms in dealing with very large volume of scientific papers, the need for more objective and well-accepted evaluation gold standards, and the heavy reliance on domain-specific knowledge sources. Addressing these research challenges is important to ensure that LBD systems ultimately become an invaluable resource within the contemporary scientific practices of diverse fields.

Future trends in the LBD research will see more convergence between the LBD field and other fields, especially machine learning and scientometrics. Unlike traditional artificial intelligence learning methods that are mainly constructed from rigid heuristics and formalisms, state-of-the-art machine learning techniques are capable of automatically learning hidden features from large data sets. The learned features can then be used as signal cues to find hidden connections between disjoint sets of literatures. Another trend may see new explorations into alternative discovery models other than Swanson's ABC model as previously suggested by (Smalheiser, 2012), with the discovery-by-analogy model being a good example of this (Cohen *et al.*, 2015). Lastly, with the advent of today's large-scale network analysis techniques, future LBD evaluation methodologies may consider ways to automatically search for instances of past scientific discoveries that have exhibited LBD characteristics (i.e. linking disparate clusters) within large bibliographic networks. These discovery instances may become good candidates for new LBD evaluation standards.

Acknowledgment

The first author would like to thank the School of Information Technology, Monash University Malaysia for supporting this research through the Monash Higher Degree Research Scholarship. The authors would also like to thank the two anonymous reviewers and the editor for providing valuable feedback on the initial manuscript of this paper.

References

Andronis, C., Sharma, A., Deftereos, S., Virvilis, V., Konstanti, O., Persidis, A. & Persidis, A. 2012. Mining scientific and clinical databases to identify novel uses for existing drugs. In *Drug Repositioning: Bringing New Life to Shelved Assets and Existing Drugs*, Michael J. Barrat & Donald E. Frail (eds). Wiley, 137.

Bassecoulard, E. & Zitt, M. 2004. Patents and publications. In *Handbook of Quantitative Science and Technology Research*, Henk F. Moed, Wolfgang Glänzel, & Ulrich Schmoch (eds). Springer, 665–694.

Bekhuis, T. 2006. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Libraries* **3**(1), 1.

Berry, M. W. & Castellanos, M. 2004. Survey of text mining. *Computing Reviews* **45**(9), 548.

Blei, D. M., Ng, A. Y. & Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* **3**, 993–1022.

Bornmann, L. & Mutz, R. 2015. Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology* **66**(11), 2215–2222.

Boyack, K. W. & Klavans, R. 2010. Co-citation analysis, bibliographic coupling, and direct citation: which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology* **61**(12), 2389–2404.

Boyack, K. W., Small, H. & Klavans, R. 2013. Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology* **64**(9), 1759–1767.

Brin, S. & Page, L. 2012. Reprint of: the anatomy of a large-scale hypertextual web search engine. *Computer Networks* **56**(18), 3825–3833.

Callon, M., Courtial, J.-P., Turner, W. A. & Bauin, S. 1983. From translations to problematic networks: an introduction to co-word analysis. *Social Science Information* **22**(2), 191–235.

Cameron, D., Bodenreider, O., Yalamanchili, H., Danh, T., Vallabhaneni, S., Thirunarayan, K., Sheth, A. P. & Rindflesch, T. C. 2013. A graph-based recovery and decomposition of Swanson's hypothesis using semantic predications. *Journal of Biomedical Informatics* **46**(2), 238–251.

Cameron, D. H. 2014. *A Context-Driven Subgraph Model for Literature-Based Discovery*. PhD thesis, Wright State University.

Cameron, D., Kavuluru, R., Rindflesch, T. C., Sheth, A. P., Thirunarayan, K. & Bodenreider, O. 2015. Context-driven automatic subgraph creation for literature-based discovery. *Journal of Biomedical Informatics* **54**, 141–157.

Chang, J. & Blei, D. M. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics* **4**(1), 124–150.

Chen, C. 2012. Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology* **63**(3), 431–449.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z. & Pellegrino, D. 2009. Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics* **3**(3), 191–209.

Chen, H.-H., Gou, L., Zhang, X. L. & Giles, C. L. 2013. Towards the discovery of diseases related by genes using vertex similarity measures. In *2013 IEEE International Conference on Healthcare Informatics (ICHI)*, 505–510. IEEE.

Cohen, A. M. & Hersh, W. R. 2005. A survey of current work in biomedical text mining. *Briefings in Bioinformatics* **6**(1), 57–71.

Cohen, P. R. 2015. Darpa's big mechanism program. *Physical Biology* **12**(4), 045008.

Cohen, T., Schvaneveldt, R. & Widdows, D. 2010. Reflective random indexing and indirect inference: a scalable method for discovery of implicit connections. *Journal of Biomedical Informatics* **43**(2), 240–256.

Cohen, T., Widdows, D. & Rindflesch, T. 2015. Expansion-by-analogy: a vector symbolic approach to semantic search. In *Quantum Interaction: 8th International Conference, QI 2014, Filzbach, Switzerland, June 30–July 3*, Atmanspacher, H., Bergomi, C., Filk, T. & Kitto, K. (eds). Springer International Publishing, 54–66.

Cohen, T., Widdows, D., Schvaneveldt, R. W., Davies, P. & Rindflesch, T. C. 2012. Discovering discovery patterns with predication-based semantic indexing. *Journal of Biomedical Informatics* **45**(6), 1049–1065.

Cohen, T., Widdows, D., Stephan, C., Zinner, R., Kim, J., Rindflesch, T. & Davies, P. 2014. Predicting high-throughput screening results with scalable literature-based discovery methods. *CPT: Pharmacometrics & Systems Pharmacology* **3**(10), 1–9.

Cory, K. A. 1997. Discovering hidden analogies in an online humanities database. *Computers and the Humanities* **31**(1), 1–12.

Davies, R. 1989. The creation of new knowledge by information retrieval and classification. *Journal of Documentation* **45**(4), 273–301.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* **41**(6), 391.

DiGiacomo, R. A., Kremer, J. M. & Shah, D. M. 1989. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. *The American Journal of Medicine* **86**(2), 158–164.

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L. & Chambers, T. 2013. Entitymetrics: measuring the impact of entities. *PloS One* **8**(8), e71416.

Eronen, L. & Toivonen, H. 2012. Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics* **13**(1), 1.

Feller, I. & Stern, P. C. 2007. *A Strategy for Assessing Science: Behavioral and Social Research on Aging*. National Academies Press.

Freeman, L. C. 1978. Centrality in social networks conceptual clarification. *Social Networks* **1**(3), 215–239.

Frijters, R., Van Vugt, M., Smeets, R., Van Schaik, R., De Vlieg, J. & Alkema, W. 2010. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Computational Biology* **6**(9), e1000943.

Fujita, K. 2012. Finding linkage between sustainability science and technologies based on citation network analysis. In *2012 Fifth IEEE International Conference on Service-Oriented Computing and Applications (SOCA)*, 1–6. IEEE.

Ganiz, M., Pottenger, W. M. & Janneck, C. D. 2005. *Recent Advances in Literature Based Discovery*. Technical report, Lehigh University.

Getoor, L. & Diehl, C. P. 2005. Link mining: a survey. *ACM SIGKDD Explorations Newsletter* **7**(2), 3–12.

Goodwin, J. C., Cohen, T. & Rindflesch, T. 2012. Discovery by scent: discovery browsing system based on the information foraging theory. In *Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, 232–239. IEEE.

Gordon, M. D. & Dumais, S. 1998. Using latent semantic indexing for literature based discovery. *Journal of the American Society for Information Science* **49**(8), 674–685.

Gordon, M. D. & Lindsay, R. K. 1996. 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. *Journal of the American Society for Information Science* **47**(2), 116–128.

Gordon, M., Lindsay, R. K. & Fan, W. 2002. Literature-based discovery on the world wide web. *ACM Transactions on Internet Technology* **2**(4), 261–275.

Hahn, U., Cohen, K. B., Garten, Y. & Shah, N. H. 2012. Mining the pharmacogenomics literature: a survey of the state of the art. *Briefings in Bioinformatics* **13**(4), 460–494.

Hristovski, D., Džeroski, S., Peterlin, B. & Rožić, A. 2000. Supporting discovery in medicine by association rule mining of bibliographic databases. In *Principles of Data Mining and Knowledge Discovery: 4th European Conference, PKDD 2000 Lyon, France, September 13–16, 2000 Proceedings*, Zighed, D. A., Komorowski, J, Żytkow, J. (eds). Springer Berlin Heidelberg, 149–159.

Hristovski, D., Friedman, C., Rindflesch, T. C. & Peterlin, B. 2006. Exploiting semantic relations for literature-based discovery. In *Proceedings of the 2006 AMIA Symposium*, 349–353.

Hu, X., Yoo, I., Song, M., Zhang, Y. & Song, I.-Y. 2005. Mining undiscovered public knowledge from complementary and non-interactive biomedical literature through semantic pruning. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management, CIKM '05*, 249–250. ACM.

Ittipanuvat, V., Fujita, K., Sakata, I. & Kajikawa, Y. 2014. Finding linkage between technology and social issue: a literature based discovery approach. *Journal of Engineering and Technology Management* **32**, 160–184.

Janssens, F., Glänzel, W. & De Moor, B. 2008. A hybrid mapping of information science. *Scientometrics* **75**(3), 607–631.

Jensen, L. J., Saric, J. & Bork, P. 2006. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics* **7**(2), 119–129.

Juršič, M., Sluban, B., Cestnik, B., Grčar, M. & Lavrač, N. 2012. Bridging concept identification for constructing information networks from text documents. In *Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools, and Applications*, M. R. Berthold (ed.). Springer Berlin Heidelberg, 66–90.

Kastrin, A., Rindflesch, T. C. & Hristovski, D. 2013. Link prediction in a mesh co-occurrence network: preliminary results. *Studies in Health Technology and Informatics* **205**, 579–583.

Kessler, M. M. 1963. Bibliographic coupling between scientific papers. *American Documentation* **14**(1), 10–25.

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM* **46**(5), 604–632.

Kostoff, R. N. 2007. Validating discovery in literature-based discovery. *Journal of Biomedical Informatics* **40**(4), 448–450.

Kostoff, R. N. 2008. Literature-related discovery (LRD): potential treatments for cataracts. *Technological Forecasting and Social Change* **75**(2), 215–225.

Kostoff, R. N. 2012. Literature-related discovery and innovation update. *Technological Forecasting and Social Change* **79**(4), 789–800.

Kostoff, R. N. 2014. Literature-related discovery: common factors for Parkinson's disease and Crohn's disease. *Scientometrics* **100**(3), 623–657.

Kostoff, R. N., Block, J. A., Solka, J. L., Briggs, M. B., Rushenberg, R. L., Stump, J. A., Johnson, D., Lyons, T. J. & Wyatt, J. R. 2009. Literature-related discovery. *Annual Review of Information Science and Technology* **43**(1), 1–71.

Kostoff, R. N. & Briggs, M. B. 2008. Literature-related discovery (LRD): potential treatments for Parkinson's disease. *Technological Forecasting and Social Change* **75**(2), 226–238.

Kostoff, R. N., Briggs, M. B. & Lyons, T. J. 2008. Literature-related discovery (LRD): potential treatments for multiple sclerosis. *Technological Forecasting and Social Change* **75**(2), 239–255.

Kostoff, R. N., Solka, J. L., Rushenberg, R. L. & Wyatt, J. A. 2008. Literature-related discovery (LRD): water purification. *Technological Forecasting and Social Change* **75**(2), 256–275.

Kraines, S. B., Guo, W., Hoshiyama, D., Makino, T., Mizutani, H., Okuda, Y., Shidahara, Y. & Takagi, T. 2010. Literature-based knowledge discovery from relationship associations based on a DL ontology created from mesh. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*, 87–106. Springer.

Larsen, P. O. & Von Ins, M. 2010. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics* **84**(3), 575–603.

Leskovec, J., Kleinberg, J. & Faloutsos, C. 2005. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 177–187. ACM.

Leskovec, J., Lang, K. J. & Mahoney, M. 2010. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th International Conference on World Wide Web*, 631–640. ACM.

Li, C., Liakata, M. & Rebholz-Schuhmann, D. 2014. Biological network extraction from scientific literature: state of the art and challenges. *Briefings in Bioinformatics* **15**(5), 856–877.

Li, J., Zhu, X. & Chen, J. Y. 2010. Discovering breast cancer drug candidates from biomedical literature. *International Journal of Data Mining and Bioinformatics* **4**(3), 241–255.

Lindsay, R. K. & Gordon, M. D. 1999. Literature-based discovery by lexical statistics. *Journal of the Association for Information Science and Technology* **50**(7), 574.

Lytras, M., Sicilia, M.-A., Davies, J., Kashyap, V. & Hu, X. 2005. Mining novel connections from large online digital library using biomedical ontologies. *Library Management* **26**(4/5), 261–270.

Manning, C. D., Raghavan, P. & Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Marsi, E. & Öztürk, P. 2015. Extraction and generalisation of variables from scientific publications. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*.

Marsi, E., Oztürk, P., Aamot, E., Sizov, G. & Ardelan, M. V. 2014. Towards text mining in climate science: extraction of quantitative variables and their relations. In *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing, Reykjavik, Iceland*.

Meyer, H. S. & Lundberg, G. D. 1985. *Fifty-One Landmark Articles in Medicine: The JAMA Centennial Series*. Chicago Review Press.

Miller, C. M., Rindflesch, T. C., Fiszman, M., Hristovski, D., Shin, D., Rosemblat, G., Zhang, H. & Strohl, K. P. 2012. A closed literature-based discovery technique finds a mechanistic link between hypogonadism and diminished sleep quality in aging men. *Sleep* **35**(2), 279–285.

Mostafa, J., Seki, K. & Ke, W. 2009. Beyond information retrieval: literature mining for biomedical knowledge discovery. In J. Y. Chen & S. Lonardi (eds). *Biological Data Mining*. CRC Press, 449–485.

Nakamura, H., Ii, S., Chida, H., Friedl, K., Suzuki, S., Mori, J. & Kajikawa, Y. 2014. Shedding light on a neglected area: a new approach to knowledge creation. *Sustainability Science* **9**(2), 193–204.

Narayanasamy, V., Mukhopadhyay, S., Palakal, M. & Potter, D. A. 2004. Transminer: Mining transitive associations among biological objects from text. *Journal of Biomedical Science* **11**(6), 864–873.

Newman, M. E. 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* **98**(2), 404–409.

Newman, M. E. 2003. The structure and function of complex networks. *SIAM Review* **45**(2), 167–256.

Newman, M. E. 2004. Fast algorithm for detecting community structure in networks. *Physical Review E* **69**(6), 066133.

Novacek, V. 2015. Formalising hypothesis virtues in knowledge graphs: a general theoretical framework and its validation in literature-based discovery experiments. *arXiv preprint arXiv:1503.09137*.

Perez-Iratxeta, C., Bork, P. & Andrade, M. A. 2002. Association of genes to genetically inherited diseases using data mining. *Nature Genetics* **31**(3), 316–319.

Perez-Iratxeta, C., Wjst, M., Bork, P. & Andrade, M. A. 2005. G2d: a tool for mining genes associated with disease. *BMC Genetics* **6**(1), 1.

Petrič, I., Cestnik, B., Lavrač, N. & Urbančič, T. 2010. Outlier detection in cross-context link discovery for creative literature mining **55**(1). *The Computer Journal*, 47–61.

Piatetsky-Shapiro, G., Djeraba, C., Getoor, L., Grossman, R., Feldman, R. & Zaki, M. 2006. What are the grand challenges for data mining?: Kdd-2006 panel report. *ACM SIGKDD Explorations Newsletter* **8**(2), 70–77.

Pratt, W. & Yetisgen-Yildiz, M. 2003. Litlinker: capturing connections across the biomedical literature. In *Proceedings of the 2nd International Conference on Knowledge Capture, K-CAP '03*, 105–112. ACM.

Preiss, J. & Stevenson, R. 2016. The effect of word sense disambiguation accuracy on literature based discovery. *BMC Medical Informatics and Decision Making* **16**(Suppl 1), 57.

Preiss, J., Stevenson, M. & Gaizauskas, R. 2015. Exploring relation types for literature-based discovery, *Journal of the American Medical Informatics Association* **22**(5), 987–992.

Salton, G. & McGill, M. J. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Sebastian, Y. 2014. Cluster links prediction for literature based discovery using latent structure and semantic features. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1275–1275. ACM.

Sebastian, Y., Siew, E.-G. & Orimaye, S. O. 2015. Predicting future links between disjoint research areas using heterogeneous bibliographic information network. In *Advances in Knowledge Discovery and Data Mining: 19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, May 19-22*, T. Cao, E.-P. Lim, Z.-H. Zhou, T.-B. Ho, D. Cheung H. Motoda (eds). Springer International Publishing, 610–621.

Sebastian, Y., Siew, E.-G. & Orimaye, S. O. 2017. Learning the heterogeneous bibliographic information network for literature-based discovery. *Knowledge-Based Systems* **115**, 66–79.

Seki, K. 2015. Hypothesis discovery exploiting closed chains of relation. In A. Hameurlain, J. Küng & R. Wagner (eds). *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXII*. Springer Berlin Heidelberg, 145–164.

Shang, N., Xu, H., Rindflesch, T. C. & Cohen, T. 2014. Identifying plausible adverse drug reactions using knowledge extracted from the literature. *Journal of Biomedical Informatics* **52**, 293–310.

Smalheiser, N. R. 2012. Literature-based discovery: beyond the ABCs. *Journal of the American Society for Information Science and Technology* **63**(2), 218–224.

Smalheiser, N. R. & Swanson, D. R. 1996a. Indomethacin and Alzheimer's disease. *Neurology* **46**(2), 583–583.

Smalheiser, N. R. & Swanson, D. R. 1996b. Linking estrogen to Alzheimer's disease an informatics approach. *Neurology* **47**(3), 809–810.

Smalheiser, N. R. & Torvik, V. I. 2008. The place of literature-based discovery in contemporary scientific practice. In P. Bruza & M. Weeber (eds). *Literature-Based Discovery*. Springer Berlin Heidelberg, 13–22.

Small, H. 2010. Maps of science as interdisciplinary discourse: co-citation contexts and the role of analogy. *Scientometrics* **83**(3), 835–849.

Sneed, W. A. 2003. *Knowledge Synthesis in the Biomedical Literature: Nordihydroguaiaretic Acid and Breast Cancer*. PhD thesis, University of North Texas.

Song, M., Han, N.-G., Kim, Y.-H., Ding, Y. & Chambers, T. 2013. Discovering implicit entity relation with the gene-citation-gene network. *PloS One* **8**(12), e84639.

Song, M., Heo, G. E. & Ding, Y. 2015. SemPathFinder: semantic path analysis for discovering publicly unknown knowledge. *Journal of Informetrics* **9**(4), 686–703.

Srinivasan, P. 2004. Text mining: generating hypotheses from medline. *Journal of the American Society for Information Science and Technology* **55**(5), 396–413.

Srinivasan, P. & Libbus, B. 2004. Mining medline for implicit links between dietary substances and diseases. *Bioinformatics* **20**(Suppl 1), i290–i296.

Srinivasan, P., Libbus, B. & Sehgal, A. K. 2004. Mining medline: postulating a beneficial role for curcumin longa in retinal diseases. In *Workshop BioLINK, Linking Biological Literature, Ontologies and Databases at HLT NAACL*, 33–40.

Stegmann, J. & Grohmann, G. 2003. Hypothesis generation guided by co-word clustering. *Scientometrics* **56**(1), 111–135.

Sun, Y. & Han, J. 2012. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery* **3**(2), 1–159.

Swanson, D. 2008. Literature-based discovery? The very idea. In *Literature-Based Discovery*, Peter Bruza & Marc Weeber (eds.). Springer, 3–11.

Swanson, D. R. 1979. Libraries and the growth of knowledge. *The Library Quarterly* **49**(1), 3–25.

Swanson, D. R. 1986a. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine* **30**(1), 7–18.

Swanson, D. R. 1986b. Undiscovered public knowledge. *The Library Quarterly* **56**(2), 103–118.

Swanson, D. R. 1987. Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science* **38**(4), 228.

Swanson, D. R. 1988. Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine* **31**(4), 526–557.

Swanson, D. R. 1990. The absence of co-citation as a clue to undiscovered causal connections. *Scholarly Communication and Bibliometrics*, 129–137.

Swanson, D. R. 1993. Intervening in the life cycles of scientific knowledge. *Library Trends* **41**(4), 606–631.

Swanson, D. R. & Smalheiser, N. R. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence* **91**(2), 183–203.

Symonds, M., Bruza, P. & Sitbon, L. 2014. The efficiency of corpus-based distributional models for literature-based discovery on large data sets. In *Proceedings of the Second Australasian Web Conference – Volume 155, AWC '14*, 49–57.

Tarjan, R. 1972. Depth-first search and linear graph algorithms. *SIAM Journal on Computing* **1**(2), 146–160.

Torvik, V. I. & Smalheiser, N. R. 2007. A quantitative model for linking two disparate sets of articles in medline. *Bioinformatics* **23**(13), 1658–1665.

Uzzi, B., Mukherjee, S., Stringer, M. & Jones, B. 2013. Atypical combinations and scientific impact. *Science* **342**(6157), 468–472.

Valdés-Pérez, R. E. 1999. Principles of human-computer collaboration for knowledge discovery in science. *Artificial Intelligence* **107**(2), 335–346.

van Haagen, H.H., AC't Hoen, P., Bovo, A.B., de Morrée, A., van Mulligen, E.M., Chichester, C., Kors, J.A., den Dunnen, J.T., van Ommen, G.J.B., van der Maarel, S.M. & Kern, V.M. 2009. Novel protein-protein interactions inferred from literature context. *PLoS One* **4**(11), e7894.

van Haagen, H. H., 't Hoen, P. A., de Morree, A., van Roon-Mom, W., Peters, D. J., Roos, M., Mons, B., van Ommen, G.-J. & Schuemie, M. J. 2011. In silico discovery and experimental validation of new protein–protein interactions. *Proteomics* **11**(5), 843–853.

van Mulligen, E. M., van Der Eijk, C., Kors, J. A., Schijvenaars, B. J. & Mons, B. 2002. Research for research: tools for knowledge discovery and visualization. In *Proceedings of the 2002 AMIA Symposium*, 835. American Medical Informatics Association.

Waltman, L. & Eck, N. J. 2012. A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology* **63**(12), 2378–2392.

Weeber, M., Klein, H., de Jong-van den Berg, L. & Vos, R. 2001. Using concepts in literature-based discovery: simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. *Journal of the American Society for Information Science and Technology* **52**(7), 548–557.

Weeber, M., Kors, J. A. & Mons, B. 2005. Online tools to support literature-based discovery in the life sciences. *Briefings in Bioinformatics* **6**(3), 277–286.

Weeber, M., Vos, R., Klein, H., Aronson, A. R. & Molema, G. 2003. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *Journal of the American Medical Informatics Association* **10**(3), 252–259.

Wei, C.-P., Chen, K.-A. & Chen, L.-C. 2014. Mining biomedical literature and ontologies for drug repositioning discovery. In *Advances in Knowledge Discovery and Data Mining: 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13-16*, V. S. Tseng, T. B. Ho, Z.-H. Zhou, A. L. P. Chen & H.-Y. Kao (eds). Springer International Publishing, 373–384.

White, H. D. & Griffith, B. C. 1981. Author cocitation: a literature measure of intellectual structure. *Journal of the American Society for Information Science* **32**(3), 163–171.

Wilkowski, B., Fiszman, M., Miller, C. M., Hristovski, D., Arabandi, S., Rosemblat, G. & Rindflesch, T. C. 2011. Graph-based methods for discovery browsing with semantic predications. In *Proceedings of the 2011 AMIA Symposium, 2011*, 1514. American Medical Informatics Association.

Witten, I. H. & Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Wren, J. D. 2004. Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics* **5**(1), 1.

Wren, J. D. 2008. The 'open discovery' challenge. In *Literature-Based Discovery*, P. Bruza & M. Weeber (eds). Springer Berlin Heidelberg, 39–55.

Wren, J. D., Bekeredjian, R., Stewart, J. A., Shohet, R. V. & Garner, H. R. 2004. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics* **20**(3), 389–398.

Yamamoto, Y. & Takagi, T. 2007. Biomedical knowledge navigation by literature clustering. *Journal of Biomedical Informatics* **40**(2), 114–130.

Yetisgen-Yildiz, M. 2006. Litlinker: a system for searching potential discoveries in biomedical literature. In *Proceedings of 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR'06) Doctoral Consortium, Seattle, WA*.

Yetisgen-Yildiz, M. & Pratt, W. 2006. Using statistical and knowledge-based approaches for literature-based discovery. *Journal of Biomedical Informatics* **39**(6), 600–611.

Yetisgen-Yildiz, M. & Pratt, W. 2008. Evaluation of literature-based discovery systems. In *Literature-Based Discovery*, P. Bruza & M. Weeber (eds). Springer Berlin Heidelberg, 101–113.

Yetisgen-Yildiz, M. & Pratt, W. 2009. A new evaluation methodology for literature-based discovery systems. *Journal of Biomedical Informatics* **42**(4), 633–643.

Youn, H., Strumsky, D., Bettencourt, L. M. & Lobo, J. 2015. Invention as a combinatorial process: evidence from US patents. *Journal of The Royal Society Interface* **12**(106), 20150272.