

Literature Based Discovery: Models, methods, and trends



Sam Henry*, Bridget T. McInnes

Department of Computer Science, Virginia Commonwealth University, 401 S. Main St., Rm E4222, Richmond, VA 23284, USA

ARTICLE INFO

Article history:

Received 3 March 2017

Revised 21 July 2017

Accepted 20 August 2017

Available online 31 August 2017

Keyword:

Literature-Based-Discovery

ABSTRACT

Objectives: This paper provides an introduction and overview of literature based discovery (LBD) in the biomedical domain. It introduces the reader to modern and historical LBD models, key system components, evaluation methodologies, and current trends. After completion, the reader will be familiar with the challenges and methodologies of LBD. The reader will be capable of distinguishing between recent LBD systems and publications, and be capable of designing an LBD system for a specific application.

Target audience: From biomedical researchers curious about LBD, to someone looking to design an LBD system, to an LBD expert trying to catch up on trends in the field. The reader need not be familiar with LBD, but knowledge of biomedical text processing tools is helpful.

Scope: This paper describes a unifying framework for LBD systems. Within this framework, different models and methods are presented to both distinguish and show overlap between systems. Topics include term and document representation, system components, and an overview of models including co-occurrence models, semantic models, and distributional models. Other topics include uninformative term filtering, term ranking, results display, system evaluation, an overview of the application areas of drug development, drug repurposing, and adverse drug event prediction, and challenges and future directions. A timeline showing contributions to LBD, and a table summarizing the works of several authors is provided. Topics are presented from a high level perspective. References are given if more detailed analysis is required.

© 2017 Elsevier Inc. All rights reserved.

Contents

1. Introduction	21
2. Models	21
2.0.1. Term representation	21
2.0.2. Document representation	22
2.0.3. System components	22
2.1. Co-occurrence models	23
2.2. Semantic models	23
2.3. Distributional models	23
2.3.1. Vector construction	23
2.3.2. Knowledge generation	24
2.4. User interaction models	24
2.5. Other models	24
3. Uninformative term removal	24
3.1. Stop word removal	24
3.2. Hierarchical filters	25
3.3. Semantic type filters	25
3.4. Relation type filters	25
4. Term ranking	25
5. Results display	25

* Corresponding author.

E-mail address: henryst@vcu.edu (S. Henry).

6.	Evaluation	26
6.1.	Discovery replication evaluation	26
6.2.	New discovery proposal and empirical evaluation	26
6.3.	Time slicing	26
6.3.1.	Gold standard generation	27
6.3.2.	Time slicing quantification	27
6.3.3.	User interaction studies	27
6.4.	Other evaluation metrics	27
7.	Application areas	27
7.1.	Drug discovery	27
7.2.	Drug repurposing	28
7.3.	Adverse drug event prediction	28
8.	Challenges and future directions	28
8.1.	Lack of adoption	28
8.2.	Methodological gaps	29
8.2.1.	Implicit term ranking	29
8.2.2.	Grouping output terms	29
8.2.3.	Query expansion	30
8.2.4.	Word sense disambiguation	30
9.	Conclusion	30
	Conflicts of interest	30
	References	30

1. Introduction

Literature-Based-Discovery (LBD) seeks to discover new knowledge from existing literature in an automated or semi-automated way. Scientific literature is growing at an exponential rate [1] causing researchers to become increasingly specialized, and making it difficult for researchers to stay current in even their narrow discipline. There is too much information for anyone to read, much less understand. This overwhelming volume of publications has led to specialized, non-interacting literatures, creating islands of knowledge in which discoveries in one area are not known outside of it [2]. LBD seeks to build bridges between these islands, increasing interdisciplinary information sharing. As the scientific literature grows, LBD is becoming an increasingly necessary tool for facilitating research.

LBD has led to countless discovery proposals ranging from treatments for cataracts [3], multiple sclerosis [4], and Parkinson's Disease [5], to understanding and discovering new health benefits of curcumin [6], and potential treatments for cancer [7]. Perhaps the most promising application areas are drug development [8–10], and repurposing [7,11,12,9,13–16], and adverse drug event (ADE) prediction [11,17–20]. Application areas outside the biomedical domain include: development of efficient water purification systems [21], accelerating the development of developing countries [22], categorizing potential bio-warfare agents [23], studying climate change [24], and identifying promising research collaborations [25].

This paper gives an overview of current LBD techniques with a focus on the biomedical domain. It begins with a description of a general model, the theoretical framework of LBD systems. Next, different methodologies are presented, followed by components common to most systems, and evaluation methodologies. Lastly a discussion of challenges and trends, and future directions is presented, and an overview of three application areas of LBD is provided.

2. Models

Nearly all LBD systems are based on or derived from Swanson's ABC co-occurrence model [26]. In this model, explicit knowledge is found in text to generate “A implies B” and “B implies C” relation-

ships. Implicit knowledge is discovered by drawing a “therefore A implies C” conclusion. There are two main ways to perform LBD, open discovery and closed discovery [27]. In open discovery, the user inputs a start term, and the system outputs a list of target terms. In closed discovery, the user inputs both a start term and a target term, and the system outputs a set of linking terms. Open discovery is used to generate new discoveries, where as closed discovery is primarily used to explain correlations or observations. Fig. 1 shows how these two methodologies differ.

2.0.1. Term representation

Using the ABC co-occurrence model as a theoretical framework, several core questions arise:

1. How do I represent a term?
2. What constitutes a relationship?
3. How do I find linking and target terms?

The answers to these questions distinguish LBD systems into three high level categories:

1. Co-occurrence models – represent terms with words, or United Medical Language System (UMLS) concepts.¹ A co-occurrence in text constitutes a relationship. Linking terms are found iteratively through co-occurrences.
2. Semantic models – represent terms with words, or UMLS concepts. Semantic parsers extract relationships from text. Linking terms are found iteratively through semantic relationships.
3. Distributional models – represent terms as context vectors. Explicit relationships (“A implies B”) are found as co-occurrences or as semantic relationships when constructing the context vectors. Implicit relationships (“A implies C”) are found in vector space via vector operations and nearest neighbor search.

These models are discussed in more detail in the next few subsections, but first an overview of document representations, and

¹ A UMLS concept represents a single meaning to which synonymous terms map.

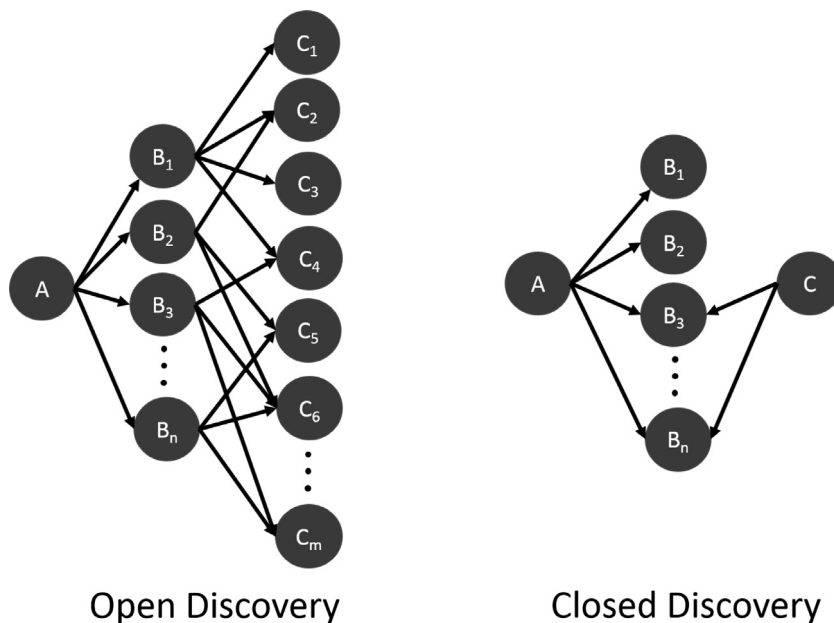


Fig. 1. Open and closed literature based discovery. Open discovery generates linking (B_i) and target (C_i) terms from just a user-input starting term (A). Closed discovery generates linking terms (B_i) from a user-input start term (A) and target term (C). The intersection between the two sets of linking terms is returned.

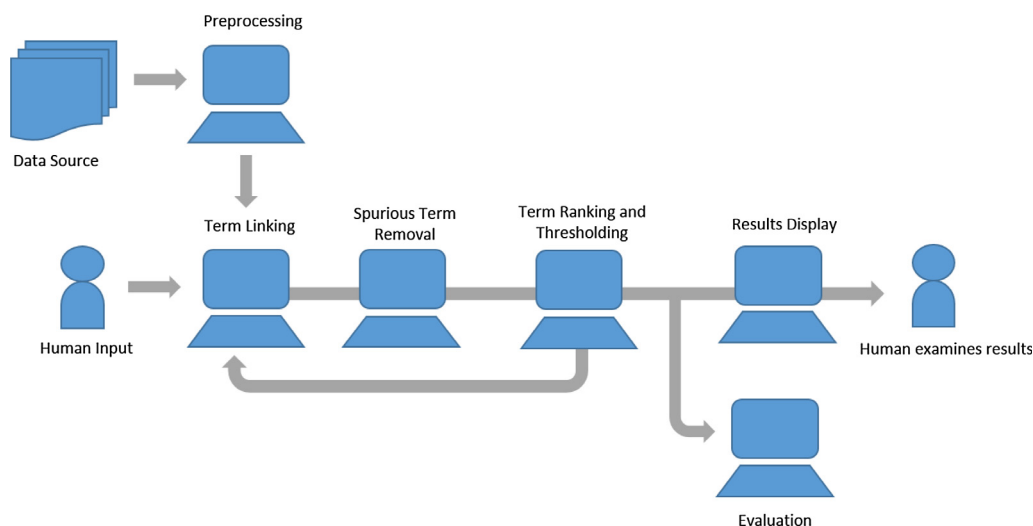


Fig. 2. A generic framework for LBD. Most systems follow a workflow similar to this one. A data source is preprocessed, or parsed to extract features of interest (be it CUIs, predication, or word vectors). A human inputs start terms and linking terms are found. Next terms are filtered from the list of linking terms, ranked and thresholded. This process is repeated, using the set of linking terms to produce a set of target terms. The target terms are then filtered, thresholded, displayed, and evaluated.

system components are presented. These are useful for distinguishing systems within each high level category.

2.0.2. Document representation

Systems within models differ in many ways, but decisions on how a document is represented, and how a co-occurrence is defined are key design decisions. Document representations may be an article title [26], MeSH descriptors [28], an abstract [29], a combination of those three [30], or even the full text of an article [31]. Co-occurrences have been represented as bigrams [32] (i.e. a co-occurrence is judged as two words appearing together as a bigram), or as co-occurrences within a window [33], sentence [34], or document [30]. Both of these decisions come down to how to most compactly represent important relationships within

a document. As documents are represented with less information, and as the distance between co-occurring term pairs narrows the number of relationships found decreases, but the chance of missing interesting relationships increases. This trade-off between precision and recall is seen again and again in LBD literature, and in decisions about system components.

2.0.3. System components

With the core of an LBD system defined, several design decisions are common to all LBD systems, including:

1. How do I eliminate uninteresting terms?
2. How do I explain and/or display the results?
3. How do I evaluate my system?

The answers to these questions are generally system independent, and techniques can be used across systems. Fig. 2 shows the workflow of a typical LBD system. From a theoretical perspective, each component in this workflow is independent of one another.

2.1. Co-occurrence models

Co-occurrence models directly use co-occurrences in text as relationships between terms. Benefits of co-occurrence models include their simplicity, and recall rates. Directly using co-occurrences ensures all possible relations in text will be captured. Co-occurrence models directly follow the ABC co-occurrence model by finding linking terms with the starting term via co-occurrences in text, and repeating this process by finding co-occurrences with each linking term to form the set of target terms. Terms may be represented as n-grams or UMLS concepts.

The first LBD systems used n-gram co-occurrences to generate linking and target terms. **N-gram co-occurrence** models do not necessarily rely on external knowledge sources, and can easily be adapted to any corpus or domain. Their theoretical background is in information retrieval. Primary authors of the method include Swanson and Smalheiser [26], and Gordon and Lindsay [32]. N-gram co-occurrence models have been largely replaced by other models.

Concept co-occurrence models use concepts rather than n-grams to represent terms. Using UMLS concepts provides normalization, stop word removal, and identification of multi-word terms. Concept co-occurrence methods typically use MeSH descriptors, or MetaMapped MEDLINE titles and/or, abstracts as document representations. Typical authors of concept co-occurrence models include Weeber et al. [27], Srinivasan et al. [28], and Yetisgen-Yildiz and Pratt [35].

Association rules [36,37] are another co-occurrence model that incorporate statistical measures to determine the likelihood of a relationship existing between terms. The theoretical background is in data mining. Once term co-occurrences are found, two statistics are computed for each linking term, *confidence* and *support*. *Confidence* estimates the percentage of articles containing the linking term that also contain the starting term, and *support* estimates the count of articles containing both the starting and linking term. Using these measures the strength of a relationship is estimated, and a threshold is applied to remove low likelihood relationships.

2.2. Semantic models

Semantic models incorporate semantic parsers to determine what constitutes a relationship. Co-occurrences, even those with high frequency can only be interpreted as an association. Relationships can be established by using the semantics of a sentence. SemRep [38] is the most popular semantic parser [39,40,34,14], but ReVerb [41] and Stanford Parser [42] have also been used [34].

Semantic parsers increase the precision of linking at the expense of recall. Some relationships may be missed during the semantic parsing process, but the relations that are extracted are more accurate and have a labeled type (e.g. *TREATS*, *COEXISTS_WITH*, *PROCESS_OF*, or *NEG_CAUSES*). This allows uninteresting relation types to be removed [39], and negative relations to be ignored [43,14].

SemMedDB² (a database of semantic predications from SemRep) is often used as data source [39,40,34,14]. Relationships can be extracted from SemMedDB in the form of a **discovery pattern** [39], which is one or more UMLS semantic type - SemRep relation

type - UMLS semantic type triplet. For example, the discovery pattern *(may_disrupt)* [7] is defined by:

*Substance X (inhibits) Substance Y and
Substance Y (causes) Pathology Z therefore
Substance X (may_disrupt) Pathology Z*

where *Substance* and *Pathology* are UMLS semantic types, *(inhibits)* matches the predication types *INHIBITS*, *(causes)* matches predication types of *CAUSES*, *PREDISPOSES*, or *ASSOCIATED WITH*, and *(may_disrupt)* matches predication types *TREATS* or *PREVENTS*. Ahlers et al. [7] use this discovery pattern to find links between anti-psychotic drugs and cancer in a closed discovery manner by defining X to be Anti-Psychotic Agents, Z to be Cancer, and retrieving all matching Y terms. Implementing discovery patterns with semantic parsers other than SemRep may be difficult. For instance, ReVerb and Stanford Parser do not map to a controlled terminology (as SemRep does to the UMLS), and they produce a greater diversity of output compared to SemRep.

Recently SemMedDB was converted to a Neo4j graph database [44] making relation extraction using discovery patterns easier, and enabling the intuitive fusing of information sources. Hristovski et al. [8] generate “upregulates” and “downregulates” relations between genes and diseases using DNA microarray analysis. The genes are mapped to identifiers in the literature, and viewed as nodes in a graph. The “upregulates” and “downregulates” relations create edges in that graph. Combining the relations found through microarray data and SemMedDB, *(inhibits)* and *(stimulates)* relations between the drugs and genes are found in SemMedDB to create “Inhibit the Upregulated” and “Stimulate the Downregulated” relation pairs. The result is a *Maybe_Treats* discovery pattern used to discover new drugs.

2.3. Distributional models

Co-occurrence models and semantic models are similar. They differ primarily on how a relationship is defined. Distributional models are more distinct. They use co-occurrence information to construct vector representations of terms. The explicit “A implies B” relationships are found during vector construction, when co-occurrences or relationships between all terms in a corpus are recorded. Similar vectors contain similar co-occurrence patterns, so “A implies C” relationships are found via nearest neighbor search (NNS) or vector operations in vector space (sometimes called *semantic space*). This is a computational convenience, as only vector comparisons are made, and the “B implies C” relationships do not need to be explicitly found. These vector representations have theoretical backgrounds in cognitive representations of terms, and approximate the idea of **conceptual spaces** in Gärdenfors’ model of human cognition [45]. These distributional models attempt to approximate how humans conceptualize and assemble knowledge.

2.3.1. Vector construction

Vector construction may be done in several ways: Latent Semantic Indexing (LSI) [30], Associative Concept Space (ACS) [29], Hyperspace Analogue to Language (HAL) [33], Tensor Encoding (TE) [46], Reflective Random Indexing (RRI) [47], or Predication-Based Semantic Indexing (PSI) [48]. These techniques either record co-occurrence information in a term-document matrix, within a sliding window, or within semantic predications.

A **term-document matrix** considers a co-occurrence as any two terms co-occurring in the document as a whole. The matrix has dimensionality of vocabulary size by number of documents. Terms present in each document are recorded in the matrix. Dimensionality reduction is then performed, most often via singular value

² footnote: <https://skr3.nlm.nih.gov/SemMedDB/>.

decomposition (SVD) based techniques [30,49]. SVD captures as much of the variation of the data as possible in the number of dimensions specified [45]. Since data often exhibits regularities SVD is effective for dimensionality reduction.

Sliding windows may also be used to collect co-occurrence information. A matrix of dimensions $|V| \times |V|$, where $|V|$ is the vocabulary size is constructed by tallying co-occurrences within a pre-defined window size (generally eight [45] or ten [50] words on each side of a focus term in the center of the window). As the sliding window moves across text in one word increments, co-occurrences are tallied in the matrix with a weight of one [45]. This has the effect of making co-occurrence counts between words proportional to the distances between them. More tightly coupled words will have higher counts, and words used further apart will have lower counts because they co-occur in fewer windows. Each row of the matrix forms a word vector. The rows may be normalized to form the final vector representations [45]. Sliding windows may also be used to encode reduced dimensionality word vectors directly, without the need to calculate an explicit term-by-term or term-by-document matrix, as is the case with RRI techniques [51].

Similar to RRI techniques is PSI, which encodes object-relation-object triplets in a vector space. This gives the ability to model both term (e.g. “Prozac” or “depression”) and relationship (e.g. “treats”) meanings in a vector space. PSI uses SemRep predications to learn the encodings, and vector operations are performed to find predication types that most strongly relate the two objects, or to generate knowledge in the form of analogies such as “Prozac is to depression as what is to schizophrenia” [48].

2.3.2. Knowledge generation

Implicit knowledge can be gained within vector space. This is typically done using a nearest neighbor search (NNS) that uses cosine, euclidean distance, or information flow [45] as a distance measure. Cole et al. [50] found that cosine outperforms other metrics in the presence of noisy B-terms, or when using both titles and abstracts as corpora. NNSs are typically performed around the starting (A) term for open discovery, and between the sum of A and C terms to discover B terms for closed discovery. Although NNSs in vector space can generate knowledge, it often produces noisy results, and several modifications have been proposed to increase accuracy. Vector priming [50] artificially boosts the weights of co-occurring terms in the start term vector. **Discovery by analogy** [52] computes both SemRep relation type vectors and UMLS concept vectors. These are combined using vector operations to form a product vector, around which NNS is performed.

2.4. User interaction models

Systems that focus on user interaction place the user as a central part of the discovery process. Discovery begins with a flash of insight, followed by an effort to realize and understand that insight [47]. User interaction focused systems are designed to promote abductive reasoning, and provide tools for deductive and inductive reasoning once a hypothesis has been generated. Their focus is on user interaction, and displaying information in a manner that facilitates greater understanding. User interaction systems are based on theories of how humans assemble new information and create new connections [53]. These systems are an aid to human creativity, rather than a fully automated hypothesis generation machine. “**Abductive reasoning**, as defined by the philosopher and logician, C. S. Peirce (1839–1914) is concerned with the generation of new explanatory hypotheses given a set of observations.” [54] Inductive and deductive reasoning can then be applied to confirm or disprove these hypotheses. Although theories of abductive reasoning have been applied to other models of LBD

[45], it is an important theory for user interaction systems. In these systems, abductive reasoning is accomplished through the theoretical framework of distributed cognition [47] in which a machine is viewed as complementary to the human mind. Users interact with the system to produce reasoning that is greater than the sum of its parts. The goal is not to automatically produce new discoveries, but rather to provide a “dynamic and interactive experience that allows scientists to both explore and validate conceptual connections” [47]. Exemplary of these systems is Epiphanet [47] which uses distributional term representations and facilitates exploration of connections between associated concepts.

Theories of discovery browsing [55–57] may also guide the design of systems. **Discovery browsing** is based on **Information Foraging Theory**, and was first proposed for LBD by Wilkowski et al. [53], and later implemented by Goodwin et al. [55] and Workman et al. [57]. In “discovery browsing” information is displayed to the user, and the user selects topics they find interesting or surprising. The Spark system [57] uses SemRep predications with a highly interactive graphical user interface to spark the creativity of the user.

2.5. Other models

The majority of LBD systems fall into one of the above mentioned model paradigms, but there are other unique systems. Some systems incorporate a **rarity** principle [31,58,57], and focus on finding infrequently co-occurring terms rather than frequently co-occurring ones. RaJoLink [31] epitomizes the idea. RaJoLink operates by first finding rare terms in the starting term literature. Several of the rare terms are selected, and common terms within the selected rare term’s literatures are found which forms a set of target terms. Linking terms are found in the last step in a traditional closed discovery manner. Later work [59] focused on also finding document outliers.

Bibliometric based systems use citation information to find linking and target literatures. Terms are then extracted from the cited literatures. Kostoff et al. [58] uses the Science Citation Index (SCI), a database of reference information to find articles related to starting, linking, and target terms.

3. Uninformative term removal

Uninformative terms are terms that provide no new or interesting information to the user. Uninformative terms may be existing, uninteresting, obvious, or spurious. Terms that are overly general or broad, such as *disease*, *drug*, *test*, or *therapeutic* don’t provide any useful information, and are correlated with most terms [60]. Eliminating them is vital to providing a concise set of information that is interpretable by a human user. Uninformative terms can be removed via stop word removal, semantic type or relation type filtering, or thresholding. Techniques are discussed in this section.

3.1. Stop word removal

Stop words may be general English words (e.g. “the”, “and”) or words uninteresting for the biomedical domain (e.g. “doctor”, “patient”). Swanson and Smalheiser [26] manually created a stop word list of 9500 + terms. It contains uninteresting and general English words. Manual lists are difficult to create and automated methods are preferred.

Stop word lists may be automatically generated using single term occurrence count thresholds, but determining an appropriate threshold is difficult and corpus dependent. Cohen et al. [61] eliminate terms that occur more than 100,000 times. Pratt and Yetisgen-Yildiz [60] eliminate terms that occur in more than

Table 1

The measures shown in this table have been used to rank and/or threshold terms of LBD systems. The measures are divided into groups based on how the measure is calculated. Term co-occurrence based measures are based on co-occurrence rates of two terms. Measures of independence are co-occurrence measures that test for statistical independence of terms. Implicit term based measures are ranking measures designed specifically to rank implicit knowledge generated by an LBD system. Predication based measures are specific to semantic predication based techniques, and vector space nearest neighbor search methods are specific to distributional methods.

<i>Term co-occurrence</i>	
Gordon and Lindsay [32]	Relative frequency
Hristovski et al. [69]	Confidence ^a
Hristovski et al. [36]	Support
Swanson et al. [70]	Literature cohesiveness (COH)
Cole and Bruza [50]	Odds-ratio
Stegmann and Grohmann [71]	Equivalence index
<i>Measures of independence</i>	
Yetisgen-Yildiz and Pratt [35,68]	Z-score
Wren et al. [67]	Mutual Information Measure (MIM)
Cole and Bruza [50]	Log Likelihood (ll)
<i>Semantic predication</i>	
Hristovski et al. [72]	Predication frequency
Wilkowski et al. [53]	Degree centrality
Cameron et al. [73]	Intra-cluster predication similarity
<i>Nearest neighbor search</i>	
Gordon and Dumais [30]	Cosine distance
Bruza et al. [33]	Euclidean distance
Bruza et al. [33]	Information flow
<i>Implicit term</i>	
Hristovski et al. [69]	$X \rightarrow Z$ Support
Wren et al. [67]	Average Mutual Information Measure (AMIM)
Wren et al. [67]	Minimum Mutual Information Measure (MMIM)
Wren et al. [67]	Average Minimum Weight (AMW)
Swanson and Smalheiser [26]	Linking Term Count (LTC)
Yetisgen-Yildiz and Pratt [68]	Linking Term Count with Average Minimum Weight (LTC-AMW)
Rastegar et al. [14]	Predicate independence/interdependence

^a Confidence and relative frequency are equivalent.

10,000 documents, and Preiss et al. [34] eliminate concepts that occur in greater than 150,000 abstracts. Gordon and Lindsay use both Term Frequency-Inverse Global Record Frequency [32] and Term Frequency-Inverse Document Frequency (TF-IDF) [62] to apply thresholds for stop word generation.

MetaMap automatically eliminates general English words, and removing general English words for distributional methods is likely unnecessary [45]. As such, automatic stop word generation techniques focus on biomedical specific words. Stop word lists may be automatically generated by observing that spurious terms will likely generate many linking terms. By repeatedly and randomly finding hidden knowledge, highly connected terms can be identified and removed [43].

3.2. Hierarchical filters

Broad terms may be eliminated using the UMLS hierarchy. Concepts on the first, second, and third level of the hierarchy [60,63] may be removed, but since vocabularies have different hierarchical structures, this technique alone is not sufficient.

The UMLS hierarchy has also been used to remove terms that are too similar to either the start or linking terms. Using UMLS concepts instead of n-grams maps all synonymous terms to the same concept, but the concept distinctions are often too fine grained. For example *migraine* and *common migraine* [64] are very similar, but correspond to different UMLS concepts. Pratt and Yetisgen-Yildiz [60] eliminate terms that are parents and children of the starting

concept, and later expand this to include grandparents and siblings [35]. Similarly, the UMLS contains a list of synonymous concepts which can also be used to find similar terms [43,34].

3.3. Semantic type filters

The UMLS classifies all concepts into one or more of 134 semantic types. UMLS semantic types range from *Reptile* to *Vitamin* to *Disease or Syndrome*. Each type is grouped into one more more of 15 semantic groups [65], such as *Disorder*, *Organizations*, or *Anatomy*. By restricting linking and target terms to specific semantic types or groups, uninformative terms can be eliminated. Semantic type filtering has become standard for most systems [27,66,28,39,63,35,53]. Selecting appropriate semantic types is challenging. A system that is too restrictive may eliminate important linking and target terms, and a system that is not restrictive enough will produce too many uninformative terms. Selecting the appropriate semantic types requires an understanding of both the UMLS and medical terminology. Since semantic groups are more broad than semantic types, selecting appropriate groups [35] may be easier. Selecting the desired relationships between terms may be even easier. Hu et al. [63] exploit the UMLS semantic network to automatically derive appropriate semantic types using user input of the desired relationship types between start, linking, and target terms.

3.4. Relation type filters

SemRep assigns one of 58 predefined relationship types to predication extracted from text. Relation type filtering can be used to eliminate uninteresting (e.g. PREVENTS when you want to find STIMULATES or AUGMENTS type relationships) or negative (e.g. NEG_CAUSES) relation types. Discovery patterns [39] (introduced in Section 2.2) combine semantic type and relation filters.

4. Term ranking

Term ranking can be used for ordering and displaying linking and target terms, and for removing uninformative terms by applying a threshold. The “small world” problem [67] states that a start term will most likely co-occur with a highly connected linking term. Since a highly connected linking term co-occurs with many terms, the “B implies C” linking step will cause the set of target terms to approach the vocabulary size. Put simply, linking always generates too many target terms, and thresholds may be applied to eliminate many of them.

Statistical thresholds are less affected by corpus size than term occurrence thresholds, and many statistical ranking measures have been applied to LBD. A list of ranking methods are presented in Table 1. The references provided give more detailed information. Yetisgen-Yildiz and Pratt [68] evaluate several of these ranking measures, and find that among these methods, using Linking Term Count with Average Minimum Weight as a tie breaker (LTC-AMW) is the best performing term ranking method.

5. Results display

Displaying results and explaining the generated discoveries is important. Ranked lists of terms are the most common system output, but this does not provide sufficient evidence explaining the discovery. Swanson and Smalheiser [26] display terms and the article titles where the terms co-occurred. This allows the user to investigate further, and draw their own conclusions. Wren [74] ranks and displays both implicit and explicit knowledge so that the user can quickly see implicit knowledge that is ranked as highly as explicit knowledge, a good indication of a discovery's

Table 2
Discoveries replicated and dates used.

Author	Discovery replicated	Dates used
Gordon and Lindsay [32]	Raynaud's disease & fish oil	1983–1985
Hu et al. [63]	Raynaud's disease & fish oil	1980–1985
	Migraine & magnesium	1980–1984
	Raynaud's disease & fish oil	1960–1986
	Migraine & magnesium	1960–1986
	Raynaud's disease & fish oil	1960–1985
Weeber et al. [27]	Somatomedin C & arginine	1960–1989
	Migraine & magnesium	1980–1984
	Magnesium deficiency & neurologic disease	1966–1994 ^a
	Alzheimer's disease & indomethacin	1966–1996
	Alzheimer's disease & estrogen	1974–June 1995 ^a
Preiss et al. [34]	Schizophrenia & calcium-independent phospholipase A2	1960–1997

^a Note: exact dates were not listed, but dates from the provided reference are stated.

validity. Cameron et al. [73] create graphs showing complex interactions between terms of interest. This provides a rich explanatory layer. Van der Eijk [29] constructs a vector space in which term location represents similarity, and co-occurrences are shown as edges. The display of results is critical to the adoption of LBD in real-world laboratory environments, and has become a popular research area [75,76].

6. Evaluation

Evaluation of LBD systems is challenging. This is due to the difficulty of acquiring a gold standard dataset. What constitutes a discovery? How can one predict all future discoveries? Even if those questions could be answered, the datasets are necessarily very large, and human evaluation of all possibilities is likely impossible. There are however, four evaluation methodologies that have become standard:

1. Discovery replication – replicating previous discoveries, particularly Swanson's initial discoveries.
2. New discovery proposal and empirical evaluation – using an LBD system to propose new discoveries.
3. Time slicing – dividing the dataset into pre-discovery and post-discovery segments. The pre-discovery segment is used to generate knowledge, while the post-discovery segment is used to evaluate the goodness of the generated knowledge.
4. User interaction studies – evaluating how well a system informs and engages users, and its usefulness in a real-world environment.

6.1. Discovery replication evaluation

Discovery replication consists of replicating a discovery made by previous systems. It is a very constrained task, and the best parameters for one discovery may not generalize well to another. Other evaluation techniques should be used in combination with discovery replication. Regardless, discovery replication is a proof-of-concept of a system [32,26,77,78,71,60,79,28,33,39,63,40], and it is the only evaluation technique used in many older publications. For discovery replication, literature published before the to-be-replicated discovery is used. For instance to replicate Swanson's *Raynaud's Disease-fish oil* discovery [80], only data prior to 1986 (the publication year of Swanson's paper) may be used. Discoveries are generated using the pre-discovery literature, and if the term of interest is returned as a target term, the discovery is deemed

successfully replicated. The *Raynaud's Disease-Fish Oil* discovery is the most commonly replicated, but authors have replicated as many as fourteen discoveries [40]. Table 2 gives a few examples of discoveries that authors have replicated, and the MEDLINE date ranges used. The presence of the desired term in a list of target terms doesn't indicate the likelihood of the term being noticed by a would-be researcher, or allow for quantitative comparisons between systems or system components. Reporting the rank of the terms of interest (e.g. *fish oil* is the 10th term in the list of target terms) is a more quantitative approach. The higher the rank, the better the system. These techniques may also be used to evaluate closed discovery systems. The ranks of the linking terms of interest are reported rather than the target terms.

6.2. New discovery proposal and empirical evaluation

A limitation to discovery replication is that it does not evaluate the ability of the system to actually make new discoveries. New discovery proposal [26,71,79,35,81,48,29,8,82,83,76,9] shows a system is capable of generating practical new knowledge. Discovery replication and new discovery proposal evaluation are often used together to prove a system's performance. Discovery proposal without expert vetting or empirical evaluation is no longer sufficient. A major critique of LBD has been the failure of proposed discoveries to withstand expert assessment, and the lack of adoption of LBD systems in their intended application domains [84,58,85]. Expert assessment and empirical evaluation attract the attention of biological and biomedical scientists, and will likely alleviate these criticisms.

Expert vetting may consist of evaluation by an expert or publication in the application domain. This allows for obvious, uninteresting, or incorrect hypotheses to be eliminated. Promising hypotheses, though should be empirically evaluated via laboratory testing. Examples of empirical evaluation include:

- DiGuacomo et al. [86] test Swanson's Raynaud's disease – Fish Oil hypothesis in a clinical trial.
- Fritjers et al. [87] confirm in vitro, their predicted associations between compounds and cell proliferation.
- Cohen et al. [12] confirm their predicted therapies for prostate cancer in vitro with cell cultures
- Wren et al. [79] perform in vivo testing of their predictions of compounds affecting the development of cardiac hypertrophy with rodent models.
- Lekka et al. [88] perform in vivo experimentation to support their treatment for Multiple Sclerosis.
- Hu et al. [10] use microarray and proteomic data to confirm hypothesized associations between specific genes and breast cancer.
- Hristovski et al. [8] use microarray data to support their hypotheses on Parkinson's disease.

New discovery proposal and empirical evaluation is a critical step, both for proving a system's viability, but perhaps more importantly for supporting and promoting LBD's viability to outside disciplines. LBD is not just a theoretical tool for information scientists, it is useful in laboratory environments, and using it to produce proven new hypotheses is vital to its adoption.

6.3. Time slicing

Another limitation to discovery replication is overfitting. It only shows a system can produce a single discovery. Time slicing attempts to alleviate this limitation by showing a system can generalize and make many new discoveries. Time slicing evaluation techniques use a cutoff date to divide the data set into pre- and

post-cutoff segments. The pre-cutoff segment is used as a training set to generate discoveries, and the post-cutoff segment is used as a test set to evaluate the generated discoveries. This leaves two questions:

1. How do I generate a gold standard?
2. How do I quantify the results?

6.3.1. Gold standard generation

A gold standard dataset is ideally a list of all new real world knowledge discovered after the cutoff date, and all potential knowledge that will be discovered in the future, something impossible to attain. Instead the gold standard is estimated by finding relationships present in the test set and absent from the training set. These relationships represent new discoveries, but the question of what constitutes a relationship, is nearly identical to that same question when designing a system. Co-occurrences within a sentence, or document may be used [66,35,68] to represent a relationship. This however, creates a noisy gold standard in which many discoveries will be falsely reported. Semantic parsers increase precision at the expense of recall [34], and using several semantic parsers further expands this trade-off. Preiss et al. [34] uses SemRep [38], ReVerb [41], and Stanford Parser [42] to generate three relationship sets. The presence of a relationship in one, any two, or all three sets can indicate with increasing confidence that the discovery is legitimate. Expert opinion can be used to generate a list of gold standard terms [89]. This will have the lowest recall rates of all the techniques, and highest precision.

6.3.2. Time slicing quantification

Time slicing evaluation is quantified using precision and recall. Early techniques using precision and recall were calculated for a single start term [66]. Relationships were extracted for a single term in the post-cutoff segment to form the gold standard, and potential discoveries were generated for that term on the pre-cutoff segment. The potential discoveries were compared to the gold standard to calculate precision and recall. This idea was expanded to use several start terms rather than just one [35], and was proposed as a formalized evaluation framework by Yetisgen-Yildiz and Pratt [68]. Time slicing methods borrow from information retrieval metrics, and include:

- **Precision and recall graphs over time** [35] which measure precision and recall rates based on the amount of data available to a system.
- **Average interpolated precision curve** [68] calculates precision and recall rates at evenly spaced intervals for several terms, and takes the average. Yetisgen-Yildiz and Pratt use 100 random starting terms [68].
- **Precision at k** [68] which calculates precision using only the top k ranked target terms. This may be averaged over several starting terms.
- **Mean average precision (MAP)** [68] calculates the average precision (the average precision at the point of retrieval of each relevant result) for multiple starting terms and takes the mean. This awards systems that rank gold standard terms highly, and provides a single number that can be compared across systems.
- **F-measure** [34] which is the harmonic mean of precision and recall. This produces a single number to quantify performance and make system comparisons easier. F-measure may be calculated for one term up to all terms in the vocabulary.

6.3.3. User interaction studies

User interaction studies monitor how users interact with an LBD system. The goals may be to improve the user interface,

improve how information is displayed, or learn how the LBD system is being used. User interaction studies are a particularly important for user interaction based systems [47,57], but are valuable for any system. User interaction is a critical role in LBD, and has traditionally been a neglected area. Yetisgen-Yildiz and Pratt [89] state “The success of an LBD system in facilitating new discoveries depends on its interface’s ability to inform and engage its users as they attempt to interpret and evaluate the proposed connections.” User interaction/usability studies can help guide the development of an effective user interface [89]. The studies focus on how the user interacts with a system, and their ability to use the system to actually make discoveries. It can lead to the redesign and refinement of how information is displayed, and give insights into how a system is actually being used [75].

Smalheiser et al. [76] provide an excellent long term study of how users interact with the ArrowSmith LBD tool. Their study reveals that users tend to use the system for concrete tasks such as obtaining information for discussion sections of papers, or “assessing whether unexpected, anomalous findings in the laboratory warranted a follow-up”. Users also began to use the system in new, unexpected ways such as constructing a list of terms that are common to two literatures, and browsing articles “in light of another context (e.g. a specific disease)”, such as the mitochondrial complex as it relates to Parkinson’s Disease. User interaction studies are critical to bringing about the adoption of LBD systems into laboratory environments. It is the responsibility of the developer to create a useful, easy-to-use tool for researchers.

6.4. Other evaluation metrics

Other evaluation techniques have been proposed [79,61,29,14], but are often specific to a particular system or methodology. Such techniques are usually accompanied by discovery replication or new discovery proposals to validate their techniques in a more conventional way. Ahmed and Alhashmi [90] propose new ideas for evaluation metrics, but do not provide sufficient details for implementation. Briefly they are: **goodness of path** which evaluates systems based on the quality of the path that links the two concepts; **early discovery** which evaluates a system based on the length of time between of the newly found hypothesis generation and discovery; and **noise discrimination** which evaluates how well a system can recover a target discovery with different levels of artificial noise added.

7. Application areas

Successfully applying LBD to new drug development, drug repurposing, and adverse drug event prediction can save millions of dollars, and save lives by bringing new drugs to the market faster, and preventing fatal adverse drug events. This section gives an overview of these application areas, and discusses some of the challenges of LBD today.

7.1. Drug discovery

Much of the work with LBD and drug discovery has focused on incorporating genetic microarray information into LBD systems [8–10]. Incorporating microarray data is promising because it adds empirical evidence to support generated hypotheses. Hu et al. [10] correlate microarray analysis of genes and diseases with the strength of those relationships in literature. Hristovski et al. [8] incorporate microarray correlations into discovery patterns, and Zhang et al. [9] identify potential prostate cancer drugs using a combination of SemRep predications and microarray data. These systems primarily use genes as linking terms, however using pro-

teins as linking terms also makes sense, “because proteins are the agents behind most physiological processes” [16]. Both proteins and genes affect disease development and progression, and can be targeted by drugs and chemicals.

7.2. Drug repurposing

Drug Repurposing is the process of finding new applications for existing drugs. Drug repurposing is on the rise, accounting for “approximately 30% of the new US Food and Drug Administration approved drugs and vaccines in recent years” [14]. Classic examples of drug repurposing include Viagra, which was developed as a treatment for angina, and was repurposed to treat erectile dysfunction; Rogaine, originally developed for high blood pressure, found success as a baldness treatment [74]; Topiramate, an anti-epileptic drug was developed to treat obesity, and Prozac, an anti-depressant was developed to treat premenstrual dysphoria [11]. Although LBD did not play a role in these repurposings, LBD is increasingly being used towards that goal [7,11,12,9,13–16].

LBD is useful for drug repurposing because it yields a better understanding of the biological effects of a drug, and may be used to evaluate a drugs benefit/risk profile. This allows one to arrive at novel discoveries [11]. As of 2011, drugs developed using LBD are in the preclinical stage [11].

A new drug costs between 500 million and 2 billion dollars to develop, and can take between 10 and 15 years [14] to come to market. The success rate is less than 10% [14]. The number of new drugs approved by the FDA is declining [11], but currently there are about 4,000 drugs approved for human use, and about 5,000 more drugs registered for investigational use [12]. Many of the investigational drugs have been extensively studied and satisfy basic regulatory requirements. By applying LBD to drug repurposing drug development costs may be reduced by up to 50%, and bring drugs to market much more quickly [11].

7.3. Adverse drug event prediction

LBD provides a better understanding of drug mechanisms and side effects, and in a similar way that this knowledge can be

applied for drug repurposing, it can also be applied to adverse drug event (ADE) prediction [11,17–20]. Adverse events can be caused by normal use, misuse, or sudden discontinuation of medications. ADEs often lead to hospitalization, and account for an estimated 12% of all emergency room visits [17]. Furthermore, the number of serious or life-threatening ADEs is increasing [11]. ADEs pose significant health and financial problems worldwide [19].

Since LBD can explain drug mechanisms and side effects it makes ADEs more easily predicted and avoided. A recent study by the Food and Drug Administration [91], found that ADE prediction systems were able to predict many life-threatening cardiac-related ADEs, and anticipated that development of similar technologies are in line with their initiatives and will be helpful tools in the future. Unforeseen ADEs may occur after drugs are released to the market, and LBD allows for early detection of these ADEs through automated analysis of literature and clinical notes. By quickly identifying ADEs both safety and quality of patient health care increase [17].

8. Challenges and future directions

8.1. Lack of adoption

LBD has been around for over 30 years, but has not been widely adopted outside of the information retrieval and text processing community. A lot of criticism has focused on this lack of adoption into laboratory and research environments. LBD’s lack of adoption can be attributed to two primary concerns: lack of empirical evaluation, and a disconnect between users and developers. Both of these topics have received considerable attention in recent years.

Concerns of lack of empirical evaluation are also discussed in the Evaluation section (Section 6) since empirical evaluation has become common for LBD systems. LBD publications are increasingly focused on applications, making discoveries, and **self-validation** of those discoveries. The biomedical domain is the primary application area. Bekhuis et al. [92] recommend that developers work on substantive problems for specific translational purposes. By self-validating proposed discoveries and creating dis-

Table 3
Comparison between LBD systems. Each row represents a different system, and each column shows how the systems differ. Document representations of abstracts, titles, and MeSH descriptors come from MEDLINE. Term representations of CUIs indicate the system uses MeSH or UMLS concepts. The Filter columns show which types of filters are used for a system. An X indicates the system uses the filter type in that column. The columns correspond to “S” for semantic type filter, “R” for relationship type filter, and “H” for hierarchical filter. An X indicates the system incorporates that filter type in its implementation as described in the referenced paper. The evaluation columns show how a system was evaluated, an X indicates the system was evaluated using that technique in the referenced paper. The columns correspond to “R” for discovery replication, “P” for discovery proposal, and “T” for time-slicing. Many of the systems and author’s works have evolved over time, and this table is meant to be indicative only of the references provided.

Author	Model	Document representation	Term representation	Filters			Evaluation		
				S	R	H	R	P	T
Kostoff et al. [58]	bibliometric	abstracts,titles,MesH	n-grams/MeSH	X			X	X	
Gordon and Lindsay [32]	co-occurrence	abstracts,titles	n-grams				X		
Weeber et al. [27]	co-occurrence	abstracts,titles	CUIs	X			X		
Wren et al. [79]	co-occurrence	abstracts,titles	CUIs					X	
Hu et al. [63]	co-occurrence	MeSH	CUIs	X	X	X	X		
Hristovski et al. [36]	co-occurrence	MeSH	CUIs	X					X
Srinivasan [28]	co-occurrence	MeSH	CUIs	X			X		
Yetisgen-Yildiz [64]	co-occurrence	MeSH	CUIs	X		X		X	X
Stegmann and Grohmann [71]	co-occurrence	MeSH	CUIs				X	X	
Pratt and Yetisgen-Yildiz [60]	co-occurrence	titles	CUIs	X		X	X		
Swanson and Smalheiser [26]	co-occurrence	titles	unigrams				X	X	
Preiss [43]	semantic	abstracts,titles	CUIs	X		X	X		
van der Eijk et al. [29]	distributional	abstracts	CUIs					X	
Gordon and Dumais [30]	distributional	abstracts,titles,MesH	n-grams				X		
Bruza et al. [33]	distributional	titles	unigrams				X		
Cohen et al. [61]	distributional	SemMedDB	CUIs	X	X			X	
Wilkowski et al. [53]	interactive	SemMedDB	CUIs	X	X			X	
Workman et al. [56]	interactive	SemMedDB	CUIs	X	X			X	
Petric et al. [31]	rarity	PMC Full Text	n-grams	X				X	
Hristovski et al. [39]	semantic	SemMedDB	CUIs	X	X		X	X	
Cameron et al. [73]	semantic	SemMedDB,MeSH	CUIs				X		

coveries with interdisciplinary teams, LBD will likely gain credibility and adoption in the intended application domains.

The **disconnect between developers and users** is being addressed by many systems. This can be seen with the increased attention given to user interaction studies, the increasing popularity of user-interaction focused systems, and increasing attention given to the display of results and interpretability. Skeels et al. [75] state that the “interface must facilitate comprehension, investigation, and evaluation of the connections proposed”. User interaction studies have also revealed that users often use systems differently than originally intended. Systems should be designed to support those uses, as well as traditional ones.

Validation and effective interfaces are important, but Bekhuis et al. [92] attribute some of the lack of adoption to differences in thinking between biomedical scientists and data scientists. “Biology has a solid foundation on experimental, empirical science. The notion that experiments can be conducted on data alone, even when the data was collected by other researchers, is a difficult paradigm shift for many scientists.” [16]. Moving beyond Swanson’s ABC model to **develop new paradigms** that more closely resemble traditional experimentation methods may bring about both acceptance by outside scientists, and ease of use, since the methods are instantly familiar to those scientists. Recently Baker et al. [93] developed a system that mimics traditional drug repurposing procedures. The system extracts side effect information from Medline and uses machine learning techniques to predict the molecular activity of chemicals. Specifically, it predicts whether a molecule will bind to receptors of interest based on its side effects. This parallels how traditional drug repurposing is done, leading to a better understanding and trust of the process. This tool is instantly recognizable, leading to more trust, familiarity, and ease of use by its users.

8.2. Methodological gaps

8.2.1. Implicit term ranking

Removal of uninformative linking and target terms has been a major area of research for LBD. This is due both to computational difficulties and the overwhelming volume of data LBD systems generate. Less focus however has been on the development of ranking measures specifically for implicit (A to C) knowledge generated by LBD systems. Ranking measure development has focused on explicit (A to B) measures. The majority of implicit ranking measures proposed (highlighted in Table 1) are adaptations of explicit measures, or rely purely on frequency (e.g. linking term count). Wren et al. [67] state that “it is unclear how these can be adapted to implied relations of interest”, and although they develop ranking measures specifically for implicit knowledge, these again are modifications of explicit measures (e.g. mutual information measure). Development of ranking measures specifically for implicit knowledge can reduce the impact of uninformative knowledge generation, since more effective thresholds can be applied, and the most interesting terms will rank higher and therefore be easily recognized by the user.

8.2.2. Grouping output terms

The idea of systematically grouping similar output terms of an LBD system was proposed by Weeber et al. [27] when they outlined their idea of *functional groups*. Weeber et al. analyze the linking terms their system generates when replicating Swanson’s Raynaud’s-Fish Oil discovery. They find three primary groupings of interest: blood viscosity, platelet aggregation, and vascular reactivity. Each of these groupings contain multiple, closely related terms, and when each functional group is separately analyzed, the output becomes more interpretable and meaningful. Other authors have performed similar grouping schemes. Baker [16]

assigns a high level classification to terms by exploiting the MeSH hierarchy. She provides a broad categorization by assigning the descriptor at the third level of the ancestor tree relating to each term. Although effective for her application, using the MeSH hierarchy alone will be problematic when using multiple taxonomies

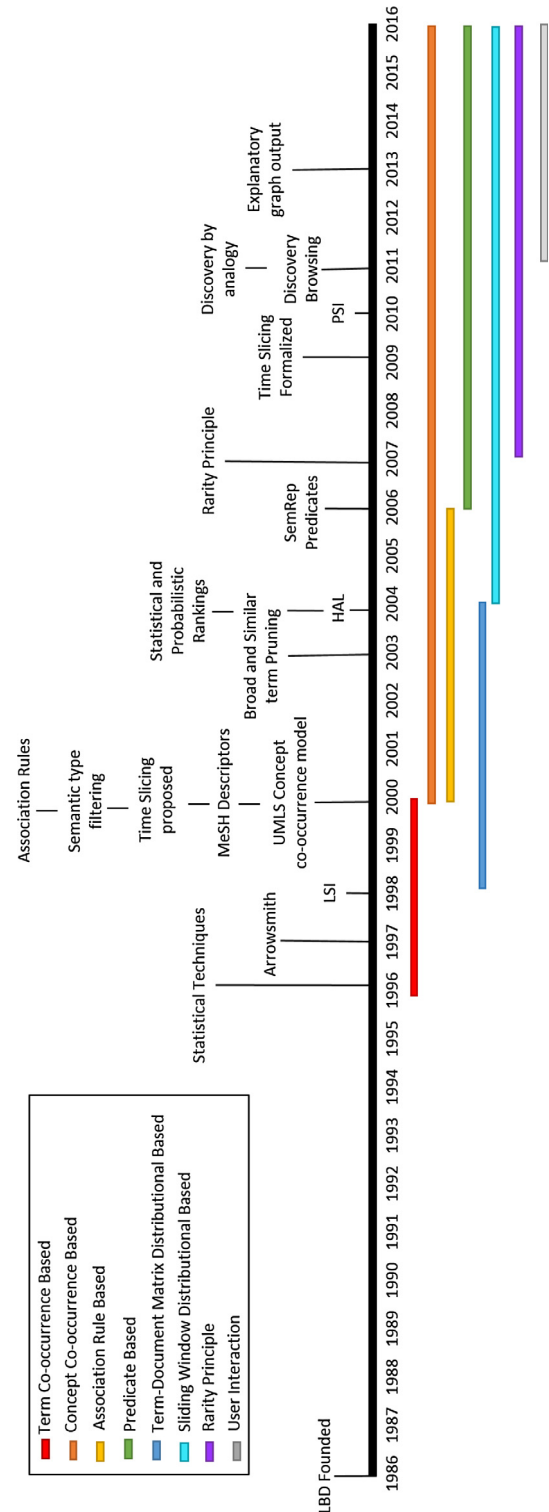


Fig. 3. Timeline of the development of LBD. Events are shown above the timeline at the date of first publication. Colored bars represent publications for different methodologies as they start and end over time. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of the UMLS (e.g. MeSH and SNOMED CT), as their structures differ significantly. Cameron et al. [73] uses graph-based similarity measures, and hierarchical agglomerative clustering to group similar “contexts” of SemRep predications. The system creates an easily interpretable graphical output that succinctly explains the interaction between terms of interest. Their system output is impressive, and is inspirational for the development of more advanced, generalizable, and efficient output grouping methods.

8.2.3. Query expansion

Query expansion is a critical component of most information retrieval systems, however it has received little attention for LBD. This is likely due to the information explosion already problematic for the field. It is likely, however, that without query expansion many terms and documents of interest are being excluded from the linking step. A user's query is an imprecise description of their information need, and query expansion augments the query to be a more precise representation of that information need [94]. Some examples of query expansion for LBD include: Kostoff et al. [58], who perform a process of “core literature expansion”. They stress the importance of this step. Manual query expansion has been performed by Wilkowski et al. [53] who manually expand the start term, *serotonin* to 183 concepts. They state that, in the future, ontology resources could be used to automate this expansion. Cameron et al. [73] manually augment starting and target terms in his closed discovery process. Automated query expansion techniques for LBD deserve more research, and it is likely that similar methods can be used for query expansion, uninformative term elimination of similar terms, and grouping of output terms, making this a particularly important research area.

8.2.4. Word sense disambiguation

Applying Word Sense Disambiguation (WSD) to LBD has recently received attention, but more work is warranted. Biomedical documents are highly ambiguous and this ambiguity leads to spurious connections [95]. Tools such as MetaMap and SemRep help reduce ambiguity, but these tools alone are not sufficient. Analysis of the 2009 Medline data shows that there are 1,072,902 terms in Medline that exist in the UMLS of which 35,013 are ambiguous, and 2979 have two or more senses with the same semantic type, therefore semantic type filtering alone is not sufficient. Zhang et al. [9] finds that ambiguity is a problem with SemRep, and their solution is to eliminate all ambiguous predications generated by SemRep. They note that this severely reduces the number of generated predication. In fact, they found ambiguity to be such a problem that they opt to use the 2006 version of the UMLS over the 2012 due to the growth of ambiguity (in particular ambiguity of protein names). Preiss et al. [95] show that performing WSD as a preprocessor for LBD improves results. Wren et al. [79] use the Acronym Resolving General Heuristic (ARGH) to resolve acronyms. Development of more effective WSD algorithms, particularly for MetaMap and SemRep will help LBD systems.

9. Conclusion

This paper covered the fundamental methodologies and components of modern and historical LBD systems. Today, a wide variety of systems from all methodologies exist. Table 3 shows several current and historical systems and key differences between them. Several of the systems have evolved over time, but the categorization is based on the system at the time of the publication provided. Recently there has been a trend towards integrating semantic parsers [96,48,40,56]. This allows for more precision when extracting relationships. Since relationships have labeled types, relationship filters may be applied, and LBD output can be better explained.

User interaction studies [76] have revealed how systems are being used by researchers, and user-interaction based systems [47,57] are becoming increasingly popular. There are benefits and drawbacks to each system, and selecting appropriate methodologies is application specific. Since evaluation is a difficult task, selecting appropriate components is difficult, and there is no definitive “best” method for LBD. It is still an evolving area of research. Fig. 3 shows contributions to LBD over time to provide an idea of how the field has progressed. Even with all of LBD's unanswered questions, it is being applied in biomedical applications today; drug development, drug repurposing, and adverse drug event prediction are popular application areas. As LBD techniques become more refined, they will likely become essential tools for these applications and others.

Conflicts of interest

We have no conflicts of interest.

References

- [1] L. Hunter, K.B. Cohen, Biomedical language processing: what's beyond pubmed?, *Molecular Cell* 21 (5) (2006) 589–594.
- [2] D.R. Swanson, Medical literature as a potential source of new knowledge, *Bull. Med. Library Assoc.* 78 (1) (1990) 29.
- [3] R.N. Kostoff, Literature-related discovery (Ird): potential treatments for cataracts, *Technol. Forecast. Social Change* 75 (2) (2008) 215–225.
- [4] R.N. Kostoff, M.B. Briggs, T.J. Lyons, Literature-related discovery (Ird): potential treatments for multiple sclerosis, *Technol. Forecast. Social Change* 75 (2) (2008) 239–255.
- [5] R.N. Kostoff, M.B. Briggs, Literature-related discovery (Ird): potential treatments for parkinson's disease, *Technol. Forecast. Social Change* 75 (2) (2008) 226–238.
- [6] P. Srinivasan, B. Libbus, Mining medline for implicit links between dietary substances and diseases, *Bioinformatics* 20 (Suppl. 1) (2004) i290–i296.
- [7] C.B. Ahlers, D. Hristovski, H. Kilicoglu, T.C. Rindflesch, Using the literature-based discovery paradigm to investigate drug mechanisms, in: *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, 2007.
- [8] D. Hristovski, A. Kastrin, B. Peterlin, T.C. Rindflesch, Combining semantic relations and dna microarray data for novel hypotheses generation, in: *Linking Literature, Information, and Knowledge for Biology*, Springer, 2010, pp. 53–61.
- [9] R. Zhang, M.J. Cairelli, M. Fiszman, H. Kilicoglu, T.C. Rindflesch, S.V. Pakhomov, G.B. Melton, Exploiting literature-derived knowledge and semantics to identify potential prostate cancer drugs, *Cancer Inform. (Suppl. 1)* (2014).
- [10] Y. Hu, L.M. Hines, H. Weng, D. Zuo, M. Rivera, A. Richardson, J. LaBaer, Analysis of genomic and proteomic data using advanced literature mining, *J. Proteome Res.* 2 (4) (2003) 405–412.
- [11] S.N. Dettreos, C. Andronis, E.J. Friedla, A. Persidis, A. Persidis, Drug repurposing and adverse event prediction using high-throughput literature analysis, *Wiley Interdisc. Rev.: Syst. Biol. Med.* 3 (3) (2011) 323–334.
- [12] T. Cohen, D. Widdows, C. Stephan, R. Zinner, J. Kim, T. Rindflesch, P. Davies, Predicting high-throughput screening results with scalable literature-based discovery methods, *CPT: Pharmacometr. Syst. Pharmacol.* 3 (10) (2014) 1–9.
- [13] H.-T. Yang, J.-H. Ju, Y.-T. Wong, I. Shmulevich, J.-H. Chiang, Literature-based discovery of new candidates for drug repurposing, *Briefings Bioinform.* 18 (3) (2017) 488–497.
- [14] M. Rastegar-Mojarad, R.K. Elayavilli, D. Li, R. Prasad, H. Liu, A new method for prioritizing drug repositioning candidates extracted by literature-based discovery, in: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2015, pp. 669–674.
- [15] M. Rastegar-Mojarad, R.K. Elayavilli, L. Wang, R. Prasad, H. Liu, Prioritizing adverse drug reaction and drug repositioning candidates generated by literature-based discovery, in: *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, ACM, 2016, pp. 289–296.
- [16] N.C. Baker, Methods in literature-based drug discovery, Ph.D. thesis, University of North Carolina at Chapel Hill, 2010.
- [17] R. Banerjee, Y. Choi, G. Piyush, A. Naik, I. Ramakrishnan, Automated suggestion of tests for identifying likelihood of adverse drug events, in: *IEEE International Conference on Healthcare Informatics*, Citeseer, 2014, pp. 170–176.
- [18] N. Shang, H. Xu, T.C. Rindflesch, T. Cohen, Identifying plausible adverse drug reactions using knowledge extracted from the literature, *J. Biomed. Inform.* 52 (2014) 293–310.
- [19] D. Hristovski, A. Kastrin, D. Dinevski, A. Burgun, L. Žiberna, T.C. Rindflesch, Using literature-based discovery to explain adverse drug effects, *J. Med. Syst.* 40 (8) (2016) 1–5.
- [20] J. Mower, D. Subramanian, N. Shang, T. Cohen, Classification-by-analogy: using vector representations of implicit relationships to identify plausibly causal

- drug/side-effect prediction, in: Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium, 2016.
- [21] R.N. Kostoff, J.L. Solka, R.L. Rushenberg, J.A. Wyatt, Literature-related discovery (Ird): water purification, *Technol. Forecast. Social Change* 75 (2) (2008) 256–275.
 - [22] M.D. Gordon, N.F. Awad, The tip of the iceberg: the quest for innovation at the base of the pyramid, in: *Literature-Based Discovery*, Springer, 2008, pp. 23–37.
 - [23] D.R. Swanson, N.R. Smalheiser, A. Bookstein, Information discovery from complementary literatures: categorizing viruses as potential weapons, *J. Am. Soc. Inform. Sci. Technol.* 52 (10) (2001) 797–812.
 - [24] E. Aamot, Literature-based discovery for oceanographic climate science, in: *European Chapter of the Association for Computational Linguistics (EACL)*, 2014, pp. 1–10.
 - [25] D. Hristovski, A. Kastrin, T.C. Rindflesch, Semantics-based cross-domain collaboration recommendation in the life sciences: preliminary results, in: *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, 2015, pp. 805–806.
 - [26] D.R. Swanson, N.R. Smalheiser, An interactive system for finding complementary literatures: a stimulus to scientific discovery, *Artif. Intell.* 91 (2) (1997) 183–203.
 - [27] M. Weeber, H. Klein, L. de Jong-van den Berg, R. Vos, et al., Using concepts in literature-based discovery: Simulating swanson's raynaud–fish oil and migraine–magnesium discoveries, *J. Am. Soc. Inform. Sci. Technol.* 52 (7) (2001) 548–557.
 - [28] P. Srinivasan, Text mining: generating hypotheses from medline, *J. Am. Soc. Inform. Sci. Technol.* 55 (5) (2004) 396–413.
 - [29] C.C. van der Eijk, E.M. van Mulligen, J.A. Kors, B. Mons, J. van den Berg, Constructing an associative concept space for literature-based discovery, *J. Am. Soc. Inform. Sci. Technol.* 55 (5) (2004) 436–444.
 - [30] M.D. Gordon, S. Dumais, Using latent semantic indexing for literature based discovery, *J. Am. Soc. Inform. Sci.* 49 (8) (1998) 674–685.
 - [31] I. Petrič, T. Urbančič, B. Cestnik, M. Macedoni-Lukšič, Literature mining method rajolink for uncovering relations between biomedical concepts, *J. Biomed. Inform.* 42 (2) (2009) 219–227.
 - [32] M.D. Gordon, R.K. Lindsay, Toward discovery support systems: a replication, re-examination, and extension of swanson's work on literature-based discovery of a connection between raynaud's and fish oil, *J. Am. Soc. Inform. Sci.* 47 (2) (1996) 116–128.
 - [33] P. Bruza, D. Song, R. McArthur, Abduction in semantic space: towards a logic of discovery, *Logic J. IGPL* 12 (2) (2004) 97–109.
 - [34] J. Preiss, M. Stevenson, R. Gaizauskas, Exploring relation types for literature-based discovery, *J. Am. Med. Inform. Assoc.* (2015) ocv002.
 - [35] M. Yetisgen-Yildiz, W. Pratt, Using statistical and knowledge-based approaches for literature-based discovery, *J. Biomed. Inform.* 39 (6) (2006) 600–611.
 - [36] D. Hristovski, J. Stare, B. Peterlin, S. Dzeroski, Supporting discovery in medicine by association rule mining in medline and umls, *Stud. Health Technol. Inform.* (2) (2001) 1344–1348.
 - [37] S. Thaicharoen, T. Altman, K. Gardiner, K.J. Cios, Discovering relational knowledge from two disjoint sets of literatures using inductive logic programming, in: *IEEE Symposium on Computational Intelligence and Data Mining, 2009 (CIDM'09)*, IEEE, 2009, pp. 283–290.
 - [38] T.C. Rindflesch, M. Fiszman, The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text, *J. Biomed. Inform.* 36 (6) (2003) 462–477.
 - [39] D. Hristovski, C. Friedman, T.C. Rindflesch, B. Peterlin, Exploiting semantic relations for literature-based discovery, in: *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, 2006.
 - [40] D. Cameron, O. Bodenreider, H. Yalamanchili, T. Danh, S. Vallabhaneni, K. Thirunarayan, A.P. Sheth, T.C. Rindflesch, A graph-based recovery and decomposition of swanson's hypothesis using semantic predications, *J. Biomed. Inform.* 46 (2) (2013) 238–251.
 - [41] A. Fader, S. Soderland, O. Etzioni, Identifying relations for open information extraction, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2011, pp. 1535–1545.
 - [42] M.-C. De Marneffe, B. MacCartney, C.D. Manning, Generating typed dependency parses from phrase structure parses, in: *Proceedings of Language Resources and Evaluation Conference (LREC)*, vol. 6, 2006, pp. 449–454.
 - [43] J. Preiss, Seeking informativeness in literature based discovery, *ACL* 2014, 2014, p. 112.
 - [44] D. Hristovski, A. Kastrin, D. Dinevski, R. Thomas, Towards implementing semantic literature-based discovery with a graph database, in: *The Seventh International Conference on Advances in Databases, Knowledge, and Data Applications*, 2015.
 - [45] P. Bruza, R. Cole, D. Song, Z. Bari, Towards operational abduction from a cognitive perspective, *Logic J. IGPL* 14 (2) (2006) 161–177.
 - [46] M. Symonds, P. Bruza, L. Sitbon, The efficiency of corpus-based distributional models for literature-based discovery on large data sets, in: *Proceedings of the Second Australasian Web Conference-vol. 155*, Australian Computer Society, Inc., 2014, pp. 49–57.
 - [47] T. Cohen, G.K. Whitfield, R.W. Schvaneveldt, K. Mukund, T. Rindflesch, Epiphanet: an interactive tool to support biomedical discoveries, *J. Biomed. Disc. Collab.* 5 (2010) 21–49.
 - [48] T. Cohen, D. Widdows, R. Schvaneveldt, T.C. Rindflesch, Finding schizophrenia's prozac emergent relational similarity in predication space, in: *International Symposium on Quantum Interaction*, Springer, 2011, pp. 48–59.
 - [49] J. Stegmann, G. Grohmann, Factor analytic approach to transitive text mining using medline descriptors, in: *Literature-Based Discovery*, Springer, 2008, pp. 115–131.
 - [50] R.J. Cole, P.D. Bruza, A bare bones approach to literature-based discovery: an analysis of the raynaud's/fish-oil and migraine-magnesium discoveries in semantic space, in: *International Conference on Discovery Science*, Springer, 2005, pp. 84–98.
 - [51] T. Cohen, R.W. Schvaneveldt, T.C. Rindflesch, Predication-based semantic indexing: permutations as a means to encode predications in semantic space, in: *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, 2009.
 - [52] T. Cohen, D. Widdows, T. Rindflesch, Expansion-by-analogy: a vector symbolic approach to semantic search, in: *International Symposium on Quantum Interaction*, Springer, 2014, pp. 54–66.
 - [53] B. Wilkowski, M. Fiszman, C.M. Miller, D. Hristovski, S. Arabandi, G. Roseblat, T.C. Rindflesch, Graph-based methods for discovery browsing with semantic predications, *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, vol. 2011, American Medical Informatics Association, 2011, p. 1514.
 - [54] T. Cohen, D. Widdows, R.W. Schvaneveldt, T.C. Rindflesch, Logical leaps and quantum connectives: forging paths through predication space, in: *AAAI Fall Symposium: Quantum Informatics for Cognitive, Social, and Semantic Processes*, 2010.
 - [55] J.C. Goodwin, T. Cohen, T. Rindflesch, Discovery by scent: discovery browsing system based on the information foraging theory, in: *2012 IEEE International Conference Bioinformatics and Biomedicine Workshops (BIBMW)*, IEEE, 2012, pp. 232–239.
 - [56] T.E. Workman, M. Fiszman, T.C. Rindflesch, D. Nahl, Framing serendipitous information-seeking behavior for facilitating literature-based discovery: a proposed model, *J. Assoc. Inform. Sci. Technol.* 65 (3) (2014) 501–512.
 - [57] T.E. Workman, M. Fiszman, M.J. Cairelli, D. Nahl, T.C. Rindflesch, Spark, an application based on serendipitous knowledge discovery, *J. Biomed. Inform.* 60 (2016) 23–37.
 - [58] R.N. Kostoff, M.B. Briggs, J.L. Solka, R.L. Rushenberg, Literature-related discovery (Ird): methodology, *Technol. Forecast. Social Change* 75 (2) (2008) 186–202.
 - [59] I. Petrič, B. Cestnik, N. Lavrač, T. Urbančič, Bisociative knowledge discovery by literature outlier detection, in: *Bisociative Knowledge Discovery*, Springer, 2012, pp. 313–324.
 - [60] W. Pratt, M. Yetisgen-Yildiz, Litlinker: capturing connections across the biomedical literature, in: *Proceedings of the 2nd International Conference on Knowledge Capture*, ACM, 2003, pp. 105–112.
 - [61] T. Cohen, D. Widdows, R.W. Schvaneveldt, P. Davies, T.C. Rindflesch, Discovering discovery patterns with predication-based semantic indexing, *J. Biomed. Inform.* 45 (6) (2012) 1049–1065.
 - [62] R.K. Lindsay, M.D. Gordon, Literature-based discovery by lexical statistics, *J. Assoc. Inform. Sci. Technol.* 50 (7) (1999) 574.
 - [63] X. Hu, X. Zhang, I. Yoo, Y. Zhang, A semantic approach for mining hidden links from complementary and non-interactive biomedical literature, in: *SDM*, SIAM, 2006, pp. 200–209.
 - [64] M. Yetisgen-Yildiz, 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, 2006.
 - [65] A.T. McCray, A. Burgun, O. Bodenreider, Aggregating umls semantic types for reducing conceptual complexity, *Stud. Health Technol. Inform.* 84 (0 1) (2001) 216.
 - [66] D. Hristovski, S. Dzeroski, B. Peterlin, A. Rožić, Supporting discovery in medicine by association rule mining of bibliographic databases, in: *European Conference on Principles of Data Mining and Knowledge Discovery*, Springer, 2000, pp. 446–451.
 - [67] J.D. Wren, Extending the mutual information measure to rank inferred literature relationships, *BMC Bioinform.* 5 (1) (2004) 1.
 - [68] M. Yetisgen-Yildiz, W. Pratt, A new evaluation methodology for literature-based discovery systems, *J. Biomed. Inform.* 42 (4) (2009) 633–643.
 - [69] D. Hristovski, B. Peterlin, J.A. Mitchell, S.M. Humphrey, Using literature-based discovery to identify disease candidate genes, *Int. J. Med. Inform.* 74 (2) (2005) 289–298.
 - [70] D.R. Swanson, N.R. Smalheiser, V.I. Torvik, Ranking indirect connections in literature-based discovery: the role of medical subject headings, *J. Am. Soc. Inform. Sci. Technol.* 57 (11) (2006) 1427–1439.
 - [71] J. Stegmann, G. Grohmann, Hypothesis generation guided by co-word clustering, *Scientometrics* 56 (1) (2003) 111–135.
 - [72] D. Hristovski, T. Rindflesch, B. Peterlin, Using literature-based discovery to identify novel therapeutic approaches, *Cardiovasc. Hematol. Agents Med. Chem. (Formerly Curr. Med. Chem.-Cardiovasc. Hematol. Agents)* 11 (1) (2013) 14–24.
 - [73] D. Cameron, R. Kavuluru, T.C. Rindflesch, A.P. Sheth, K. Thirunarayan, O. Bodenreider, Context-driven automatic subgraph creation for literature-based discovery, *J. Biomed. Inform.* 54 (2015) 141–157.
 - [74] J.D. Wren, The 'open discovery' challenge, in: *Literature-Based Discovery*, Springer, 2008, pp. 39–55.

- [75] M.M. Skeels, K. Henning, M.Y. Yildiz, W. Pratt, Interaction design for literature-based discovery, in: CHI'05 Extended Abstracts on Human Factors in Computing Systems, ACM, 2005, pp. 1785–1788.
- [76] N.R. Smalheiser, V.I. Torvik, A. Bischoff-Grethe, L.B. Burhans, M. Gabriel, R. Homayouni, A. Kashef, M.E. Martone, G.A. Perkins, D.L. Price, Collaborative development of the arrowsmith two node search interface designed for laboratory investigators, *J. Biomed. Disc. Collab.* 1 (1) (2006) 8.
- [77] M. Weeber, H. Klein, A.R. Aronson, J.G. Mork, L. De Jong-van Den Berg, R. Vos, Text-based discovery in biomedicine: the architecture of the dad-system, in: Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium, American Medical Informatics Association, 2000, p. 903.
- [78] D. Hristovski, B. Peterlin, J.A. Mitchell, S.M. Humphrey, L. Sitbon, I. Turner, Improving literature based discovery support by genetic knowledge integration, *Stud. Health Technol. Inform.* (2003).
- [79] J.D. Wren, R. Bekereditian, J.A. Stewart, R.V. Shohet, H.R. Garner, Knowledge discovery by automated identification and ranking of implicit relationships, *Bioinformatics* 20 (3) (2004) 389–398.
- [80] D.R. Swanson, Fish oil, raynaud's syndrome, and undiscovered public knowledge, *Perspect. Biol. Med.* 30 (1) (1986) 7–18.
- [81] T. Urbančič, I. Petrič, B. Cestnik, M. Macedoni-Lukšič, Literature mining: towards better understanding of autism, in: Conference on Artificial Intelligence in Medicine in Europe, Springer, 2007, pp. 217–226.
- [82] R.N. Kostoff, C.G. Lau, Combined biological and health effects of electromagnetic fields and other agents in the published literature, *Technol. Forecast. Social Change* 80 (7) (2013) 1331–1349.
- [83] R.N. Kostoff, U. Patel, Literature-related discovery and innovation: chronic kidney disease, *Technol. Forecast. Social Change* 91 (2015) 341–351.
- [84] R.N. Kostoff, Literature-related discovery (Ird): introduction and background, *Technol. Forecast. Social Change* 75 (2) (2008) 165–185.
- [85] R.N. Kostoff, J.A. Block, J.L. Solka, M.B. Briggs, R.L. Rushenberg, J.A. Stump, D. Johnson, T.J. Lyons, J.R. Wyatt, Literature-related discovery (Ird): lessons learned, and future research directions, *Technol. Forecast. Social Change* 75 (2) (2008) 276–299.
- [86] R.A. DiGiacomo, J.M. Kremer, D.M. Shah, Fish-oil dietary supplementation in patients with raynaud's phenomenon: a double-blind, controlled, prospective study, *Am. J. Med.* 86 (2) (1989) 158–164.
- [87] R. Frijters, M. Van Vugt, R. Smeets, R. Van Schaik, J. De Vlieg, W. Alkema, Literature mining for the discovery of hidden connections between drugs, genes and diseases, *PLoS Comput. Biol.* 6 (9) (2010) e1000943.
- [88] E. Lekka, S.N. Deftereos, A. Persidis, A. Persidis, C. Andronis, Literature analysis for systematic drug repurposing: a case study from biovista, *Drug Discovery Today: Therapeutic Strategies* 8 (3) (2012) 103–108.
- [89] M. Yetisgen-Yildiz, W. Pratt, Evaluation of literature-based discovery systems, in: *Literature-Based Discovery*, Springer, 2008, pp. 101–113.
- [90] A. Ahmed, S.M. Alhashmi, A metric for literature-based discovery methodology evaluation, in: 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), IEEE, 2015, pp. 1–5.
- [91] E.J. Matthews, A.A. Frid, Prediction of drug-related cardiac adverse effects in humans: creation of a database of effects and identification of factors affecting their occurrence, *Regul. Toxicol. Pharmacol.* 56 (3) (2010) 247–275.
- [92] T. Bekhuis, Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy, *Biomed. Digital Libraries* 3 (1) (2006) 2.
- [93] N.C. Baker, D. Fourches, A. Tropsha, Drug side effect profiles as molecular descriptors for predictive modeling of target bioactivity, *Molec. Informat.* 34 (2–3) (2015) 160–170.
- [94] M. Symonds, P. Bruza, G. Zuccon, B. Koopman, L. Sitbon, I. Turner, Automatic query expansion: a structural linguistic perspective, *J. Assoc. Inform. Sci. Technol.* 65 (8) (2014) 1577–1596.
- [95] J. Preiss, M. Stevenson, The effect of word sense disambiguation accuracy on literature based discovery, in: Proceedings of the ACM Ninth International Workshop on Data and Text Mining in Biomedical Informatics, ACM, 2015, 1–1.
- [96] D. Hristovski, C. Friedman, T.C. Rindflesch, B. Peterlin, Literature-based knowledge discovery using natural language processing, in: *Literature-Based Discovery*, Springer, 2008, pp. 133–152.