Analyzing LBD Methods using a General Framework

A.K. Sehgal, X.Y. Qiu, and P. Srinivasan

Abstract This chapter provides a birds-eye view of the methods used for literature-based discovery (LBD). We study these methods with the help of a simple framework that emphasizes objects, links, inference methods, and additional knowledge sources. We consider methods from a domain independent perspective. Specifically, we review LBD research on postulating gene–disease connections, LBD systems designed for general purpose biomedical discovery goals, as well as LBD research applied to the web. Opportunities for new methods, gaps in our knowledge, and critical differences between methods are recognized when the "literature on LBD" is viewed through the scope of our framework. The main contributions of this chapter are in presenting open problems in LBD and outlining avenues for further research.

1 Introduction

Literature based discovery (LBD), also known as text mining and knowledge discovery from text (KDT), has garnered significant breadth and depth as a field of research and development. The field is vibrant as seen for instance by the growing number

A.K. Sehgal
Department of Computer Science, The University of Iowa, Iowa City, IA 52242, USA

X.Y. Qiu
Department of Management Sciences, The University of Iowa, Iowa City, IA 52242, USA

P. Srinivasan
Department of Computer Science, The University of Iowa, Iowa City, IA 52242, USA
and
Department of Management Sciences, The University of Iowa, Iowa City, IA 52242, USA
and
School of Library and Information Science, The University of Iowa, Iowa City, IA 52242, USA
psriniva@iowa.uiowa.edu

P. Bruza and M. Weeber (eds.), *Literature-based Discovery,*
Springer Series in Information Science and Knowledge Management 15.

of conferences, workshops, papers, commercial and free systems, and review papers (e.g. Weeber et al. [27]). There is also a growing variety of LBD methods, orientations and applications. We observe that in general, LBD strategies are designed to fit the problem at hand with methods selected or designed in a somewhat adhoc manner. And since the space of text mining problems is broad (and growing), the range of solutions proposed and applied is also broad. As a consequence, there is a bewildering array of methods in text mining. While this situation offers almost free rein to researchers, it also makes it challenging to determine what methods (or aspects about methods) are most successful or most appropriate for a given problem in a specific domain. Seemingly similar problems are sometimes addressed using significantly different approaches while certain LBD approaches exhibit broader appeal. At this point, what is needed is a "meta-level" examination of the major milestones in methods. Thus our goal is to begin such an examination by offering a bird's eye view of LBD research. In particular, we analyze methodologies in LBD papers using a general framework. Some of the expected outcomes from such a framework-based review are to be able to more effectively:

1. Compare and contrast research in LBD
2. Observe the gaps in research
3. Assess the prevalence of particular methods
4. Make comparisons across domains or type of text
5. Understand the relationship between LBD methods and problems being solved
6. Understand the evolution of ideas in LBD research

Although we select papers for review with a fairly broad brush, we do not claim comprehensiveness in coverage. Likely the selections will reveal our own inclinations and preferences. Despite these built-in limitations, this framework-based review, is to the best of our knowledge, a first attempt at *domain-independent* meta-analysis of LBD research with a significant emphasis on *methodology*. We offer it as a potentially useful starting point for discussion, extension and refinements by others.

2 A Framework for Analyzing LBD Research and Development

LBD refers to automatic or semi-automatic efforts supporting end user exploration of a text collection with the goal of generating or exploring new ideas. Specifically, LBD systems help *form* and/or *explore* hypotheses using large collections of texts. LBD takes off from an age old process fundamental to fields of intellectual endeavor such as the sciences, where ideas build upon prior published work. LBD systems are of interest given their potential to consider very large sets of documents as also documents from fields that a user would not normally study. Generating or exploring hypotheses within such large-scale and heterogeneous document collections typically implies effort well beyond human capacity. While offering these advantages, LBD systems are far from reflecting the human acuity involved in the

manual processes they try to emulate. In fact, LBD output is always tentative, requiring end user decisions on suggestions to take forward and suggestions to reject.

The kinds of hypotheses of interest in LBD are those that somehow relate at least a pair of entities. For example, an LBD system may suggest a financial connection between two individuals, or a link between a gene and a disease, or indicate potential interest in a product from the view point of an organization, or find communities of people related in some novel way. LBD is clearly akin to data-mining from *structured* data which also focuses on hypothesis formation and knowledge discovery. The power of LBD is seen especially in its capacity to generate novel ideas by bridging different areas of specializations represented in the text collection, thus reflecting a multidisciplinary perspective.

LBD research has strong and early roots based in the research of Swanson and Smalheiser (see chapter titled 'Literature Based Discovery? The Very Idea'). Their initial LBD efforts lead them to successfully postulate several hypotheses by linking evidence extracted from different documents. However, these were accomplished through significant manual effort. Since then a growing body of research, including Swanson and Smalheiser's own work with their ARROWSMITH systems,[1,2] aims at automating LBD. The overall approach is to try to automate as many of the key steps in LBD as possible, thereby minimizing human intervention. LBD strategies have been developed and applied to biomedicine in general and bioinformatics in particular. These efforts typically involve the MEDLINE[3] database with optionally allied sources such as Entrez Gene[4] and OMIM[5] and vocabularies such as the Gene Ontology [4]. LBD has also been applied to the humanities field as well as to knowledge discovery problems on the web.

2.1 LBD Framework

Based upon our own experiences in LBD [22–24], including work on Manjal[6], our prototype biomedical LBD system, and our understanding of the literature, we propose a simple framework for analyzing LBD methods. The framework has the four dimensions listed in Fig. 1. It allows us to understand and specify the key methodological choices made by the authors of the papers reviewed. It also allows us to objectively compare studies and suggest instances where alternative methods may also be beneficial.

Objects refer to the kinds of concepts (abstract or otherwise) that are the focus of the LBD effort. In some cases these may refer to entities of a specific type such as genes, perhaps even limited to genes of a specific species. Other studies may

[1] University of Chicago version: http://kiwi.uchicago.edu/

[2] University of Illinois – Chicago version: http://arrowsmith.psych.uic.edu

[3] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed

[4] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

[5] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

[6] http://sulu.info-science.uiowa.edu/Manjal.html

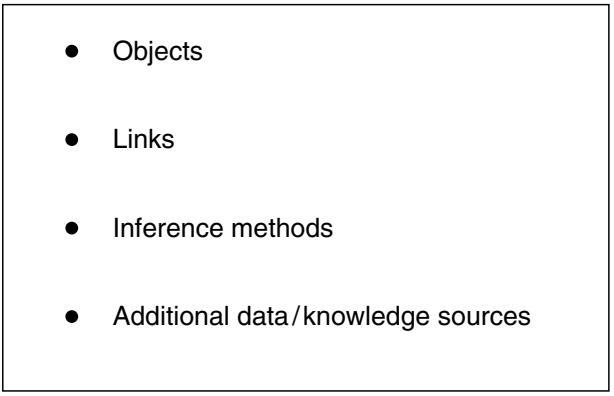

Fig. 1 LBD framework

involve multiple varieties of entity types such as persons, organizations and products. In still other cases, the LBD objects may be a collection of "topics" where topics may refer to PubMed queries (e.g. ("hypertension" AND ("2001/01/01"[PDAT]: "2006/12/31"[PDAT])). Given a particular kind of object (entity), say genes, studies may differ on what information is used to derive representations for each gene. One could use, for instance, MEDLINE records related to the gene as the representation, or the gene's sequence, or its MeSH profile, or the MEDLINE sentences in which the gene name or its alias appears. Even with a given source, say MEDLINE records, variations are possible. One may retrieve records from PubMed using the disjunction of the gene's various names. Or, one may use only those documents that provide evidence of GO based annotation for the gene. Additionally, weights are sometimes allocated to the different features in the representation. One document (or a sentence or a MeSH term) may be more central to the gene than another. And of course different studies may employ different weighting strategies, including none at all. Thus while analyzing the success (or failure) of LBD methods for specific problems, one has to pay careful attention to the kinds of objects and their representations used.

Links represent associations between objects of interest. Links may vary from straightforward co-occurrence based connections to similarity-based assessments to more semantically motivated relationships. These may be obtained in different ways, e.g., from curated or automatically generated databases or extracted from texts using pattern recognition or more advanced NLP methods. As with objects, links may also be weighted. Additionally these may be directed and/or labeled. Multiple links between objects may also be used. Each of these options and their various combinations offer different capabilities to an LBD system.

Inference methods refer to the reasoning strategies used to identify *implicit* connections between objects. In the simplest case, one may use a transitive relationship between two objects to infer a novel connection. Extensions of this idea lead to the classic strategies of Open and Closed discovery (see chapter titled 'Literature Based Discovery? The Very Idea'). Other methods are also available. For example,

connections may be inferred between two objects if their representations (retrieved documents, MeSH profiles) are very similar even if they do not co-occur.

The final dimension refers to whether *additional sources* are used within the LBD process. For example, sequence data or disease mapping to chromosomal regions may be used to constrain the LBD hypotheses generated regarding putative links between genes and diseases. Similarly, geographical co-location may be used to limit the potential customers of products suggested by the LBD process.

To close this section on our framework, it may be that two studies with different goals use methods that are quite similar. This would indicate the generalizability of the methods. On the other hand two methods applied to the same goal could look very different. This might imply flexibility in the problem. It may also call for or lead to direct comparisons of the two methods. More generally, the framework might also point to dimensions that are less well explored than others. Thus our goal is also to identify open areas for research on LBD.

We now analyze select LBD research using our framework. Reviews of LBD methods are done primarily based upon descriptions in published papers. We present our analysis in three parts. In Sect. 3 we analyze a set of papers that directly target the discovery of gene–disease connections. By focusing on this subset of LBD research we will highlight the variability across methods even when they have the same LBD goal. In Sect. 4 we study general purpose LBD systems in biomedicine. Finally in Sect. 5 we examine LBD applications on the web. Each part includes an analysis of methods covered. Following this, in Sect. 6 we present our conclusions.

3 LBD for Postulating Gene–Disease Connections

Postulating novel connections between genes and diseases is a major emphasis in bioinformatics text mining. In all papers reviewed in this section, gene–disease links are postulated without qualification as to the type of link. For each study reviewed (throughout the paper) we identify its major features in terms of the key dimensions of our framework. We present Objects, Links and Inference Methods in a table. Additional knowledge sources are described in the discussions. By default, links are considered weighted, symmetric and unlabeled. Otherwise, unweighted links are marked with a U in the Notes column, asymmetric weights with an A and labeled links with an L. These qualifications under Notes apply only to the links.

3.1 G2D (Perez-Iratxeta et al. 2002)

G2D [20] is a system that ranks candidate genes for genetically inherited diseases for which no underlying gene has yet been assigned. The key objects and link are shown in Table 1. Two types of links are core to their procedure. Although both involve MEDLINE as the source, records supporting the links are extracted in different ways. The first link type (L1) associates ‘pathological conditions’ and

Table 1 G2D – Perez-Iratxeta et al. (2002)

Type ID	Object type	Object representation/link derivation	Notes
O1	Disease	Disease manifestations (category C MeSH terms)	–
O2	Chemical	Chemical (category D MeSH terms)	–
O3	Annotation	GO term	–
O4	Gene sequence	From RefSeq	–
O5	Gene sequence	From chromosome region of disease	–
L1	O1, O2	Co-occurrence in MEDLINE records about disease of interest	
L2	O2, O3	Co-occurrence between O2 and O3 in MEDLINE records used as evidence to annotate sequences O4 with O3	
L3	O1, O3	Inferred through L1 and L2	
L4	O3, O4	O3 annotates sequence O4	L
L5	O4, O5	Homology	U
L6	O1, O4	L3 and L4	
L7	O1, O5	Inferred from L5 and L6	
IM	*Average of fuzzy scores representing best paths between disease and GO terms*		

'chemical terms'. Pathological conditions are represented by category 'C' MeSH terms while chemical terms are category 'D' MeSH terms. L1 strength is a symmetric weight and is calculated as the number of MEDLINE records with both terms divided by the number of records having either term. The second link type (L2) connects 'chemical terms' and GO terms describing protein function. L2 strength is also symmetric and is calculated as the number of records having the chemical term and also providing evidence supporting annotation by the GO term in RefSeq[7] divided by the number of records with either feature.

L3 is inferred between the pathological condition and GO term pairs. Since several chemical bridges are possible between a pair, the weight is a fuzzy score representing the best possible chemical bridge. It is symmetric and is calculated as the product of weights for the best chemical path. Given that a disease may be characterized by several pathological conditions, the L3 weight between a disease and a GO term is the highest weight calculated for any of its manifestations. They rank candidate sequences using the homology between RefSeq annotated sequences in the chromosomal region to which the disease is mapped (L5). Ranking of candidate sequences to a disease is by the average of the scores calculated for each of their GO terms and the disease.

Thus when we look closely at their methods at least five types of objects and seven link types may be identified. Notice also for example, that their approach looks for the best path connecting a GO term to a disease using disease pathological conditions and chemicals as bridges. However, the score for a candidate gene is not the best offered through its annotations, but the average. This score is then normalized as an R score to allow for standardized comparisons.

[7] http://www.ncbi.nlm.nih.gov/RefSeq/

3.2 eVOC (Tiffin et al. 2005)

The authors use the eVOC Anatomical System ontology[8] as a bridging vocabulary to select candidate disease genes [26]. (We refer to this system as the eVOC system.) Specifically they exploit information on the genes' expression profiles within tissues affected by the disease of interest. As described by them, researchers may mine associations between disease and affected tissues without having a clinical understanding of the disease. This connection may then be applied to the selection of candidate genes for the disease. The authors state that the eVOC anatomical terminology has the advantage of being simple and purely descriptive, without the interpretational bias that may be associated with functional annotation systems such as GO. Table 2 identifies the key objects and links in their approach.

The authors first identify the top ranked eVOC terms for a given disease (through L3). This is done by calculating a score that depends upon how frequently a term is associated with the disease in MEDLINE (L2), as well as upon how often the term is used to annotate RefSeq genes (L1). The former is an asymmetric weight calculated as the number of abstracts containing both the disease name and the eVOC term divided by the number of abstracts with the disease name. The later weight is also asymmetric and is calculated as the number of RefSeq genes annotated by the term divided by the number of annotated genes. Here annotation counts for an eVOC term include counts for descendent terms in the eVOC hierarchy. Finally the L3 weight between each eVOC anatomy term and disease name is calculated as $[2 * association\ weight + annotation\ weight]/2$. They then select the top scoring n eVOC terms as characterizing the disease. L5 which is the inferred link to new genes considers expression based annotation obtained from the Ensembl genomic database[9]. The final candidate gene list contains those annotated with at

Table 2 eVoC – Tiffin et al. (2005)

Type ID	Object Type	Object representation/link derivation	Notes
O1	Disease	Disease term (and retrieved MEDLINE set)	–
O2	eVoc term	eVoc term (and retrieved MEDLINE set)	–
O3	Gene	RefSeq entries	–
O4	Gene	Gene entries in Ensembl	–
L1	O2, O3	Frequency based annotation weight	A
L2	O1, O2	Frequency based association weight	A
L3	O1, O3	Score from L1 and L2	
L4	O1, O4	Expression of O3 in O2 tissues from Ensemble	L
L5	O1, O4	Inferred from L3 and L4	
IM	*Genes annotated with at least $n - m$ of the n top ranked eVOC terms characterizing the disease*		

[8] http://www.sanbi.ac.za/evoc

[9] http://www.ensembl.org/Homo_sapiens/martview

least $n - m$ of the eVOC terms characterizing the disease. m functions as a slack parameter on the degree of matching. Optimal values for n and m were determined using training data and then evaluated on an independent test dataset.

3.3 BITOLA (Hristovski et al. 2001, 2003)

BIOTLA [13, 14] is a text mining system designed for the biomedical domain in general. We include it here as it has also been extended to identify novel genes for diseases. Using association rules with confidence and support scores, BITOLA builds on Swanson and Smalheiser's Open and Closed discovery strategies. In their earlier work [14], objects of interest are represented only by MeSH concepts. In later work [13], tailored to the gene–disease problem, they also consider gene names (that are not necessarily MeSH terms). Links are derived from co-occurrence data and are weighted by *confidence* and *support* scores. The *confidence* in a link between two MeSH concepts X and Y is an asymmetric weight which is the number of records having both X and Y divided by the number of records with X alone (or Y alone depending upon the perspective). Given a starting concept, X, associated Y concepts are found. Z concepts that are in turn associated with Y are identified. Each X–Z combination defines a novel relationship if they are not already directly associated. Filters may be applied to constrain the nature of the bridging Y concepts to those belonging to specific UMLS[10] semantic types of interest. Similarly links may be filtered based on threshold values for confidence or support. In the later version, tailored to the gene–disease application, they also provide filters to constrain the gene and disease to the same chromosomal region. They were able to postulate FLNA as a candidate gene for Bilateral perisylvian PMG, a malformation of cortical development in the brain [13]. More details about BITOLA are available in the chapter titled 'Literature-Based Knowledge Discovery using Natural Language Processing'. In Table 3 we show the key objects and links in their approach when

Table 3 BITOLA – Hristovski et al. (2003)

Type ID	Object type	Object representation/link derivation	Notes
O1	Disease	Disease MeSH term (and corresponding MEDLINE records with this term)	–
O2	Cell function	MeSH term of UMLS semantic type "cell function" (and corresponding MEDLINE records with this MeSH term)	–
O3	Gene	Gene names or aliases (and corresponding MEDLINE set with any of these terms)	–
L1	O1, O2	Confidence in O2 given O1	A
L2	O2, O3	Confidence in O3 given O2	A
L2	O1, O3	Inferred from L1 and L2 with added chromosomal constraint	A
IM	*Association rules exploiting transitivity with calculated confidence and support scores*		

[10] http://www.nlm.nih.gov/research/umls/

terms with the UMLS semantic type *cell function* are chosen as bridges between diseases and genes.

One observation to make at this point is that final scores on suggestions do not incorporate weights calculated for the intermediate links through the bridging Y pathways. Thus although a suggested Z concept may have a high confidence connection to its Y concept, this Y in turn may have a low confidence connection to the starting X concept (relative to other Y terms on the list).

3.4 Adamic et al. (2002)

In this work [2] the authors identify gene symbol occurrences (official and alias symbols) in MEDLINE records retrieved for a given disease. They then calculate the strength of the relationship between the gene and the disease by comparing the observed number of documents and the expected number of documents in which the gene is mentioned. Expected frequencies are determined assuming a random distribution of the gene term. They state that their work could be used to maintain a list of known genes for a disease. They do not explicitly explore "new" connections. However we include their research here as this approach of identifying genes that exhibit a statistically significant occurrence pattern in the disease literature is at the foundation of several papers and systems (Table 4).

One example is the newly formed Autoimmune Disease Database [16]. In it occurrences of gene names in document sets retrieved for diseases or disease names in document sets retrieved for genes are assessed for significance. LitMiner [18] also uses co-occurrence as the basis for relating two entities. Several types of entities are considered including genes and diseases. Unlike the Adamic et al. effort their link weight is symmetric and is calculated as the observed co-occurrence frequency divided by that expected by chance alone. MedGene[11] [15] is our last example of a system that relies on co-occurrence data. After comparing several statistical methods such as chi-square and Fishers exact probabilities, the authors select a symmetric measure called the natural log of the product of frequency. This is the product of two ratios. One is the disease–gene co-occurrence frequency divided by the disease frequency. The other is the disease–gene co-occurrence frequency divided by the gene frequency.

Table 4 Adamic et al. (2002)

Type ID	Object type	Object representation/link derivation	Notes
O1	Disease	Disease search terms	–
O2	Gene	Names and aliases	–
L1	O1, O2	Occurrence of O2 in MEDLINE records retrieved by O1	A
IM	*Comparison of observed and expected occurrences of O2 in O1*		

[11] http://hipseq.med.harvard.edu/MEDGENE/login.jsp

3.5 Wilkinson and Huberman (2004)

Wilkinson and Huberman [28] take the notion of significant co-occurrences and expand it to get additional capabilities. Specifically they explore methods for finding communities of genes that are likely to be functionally related in a given context such as that defined by a particular disease. Given a set of documents for a disease, they first identify occurrences of gene and protein names in each article. They then limit the identified gene set to those that are statistically relevant to the topic (disease). This is again done by comparing the observed number of gene occurrences with the expected number estimated, assuming no correlation between the gene and the disease. Specifically, given that two uncorrelated terms co-occur according to a binomial distribution, they consider observed gene-disease co-occurrences of at least one standard deviation greater than the binomial expected value as statistically relevant (Table 5).

They then create a co-occurrence network of the relevant genes and apply a graph partitioning algorithm to identify communities. Links in the network are not weighted and simply indicate that the genes co-occur. Their graph partitioning algorithm is based on the concept of the "betweenness centrality" of an edge. The betweenness of an edge AB is defined as the number of shortest paths between pairs of other vertices that contain AB. The edge with the highest betweenness is likely an intercommunity edge and is removed, thus breaking up the network into two or more connected components. This process iterates till certain stopping conditions are met. At each iteration the betweenness scores are recomputed. At the end, connected gene sets are declared to form a "community".

When applied to the disease topic 'colon cancer', the authors show that functionally unrelated genes tend to be placed in separate communities even if they exhibit some co-occurrence. Their method is offered as an approach for summarizing available information. The communities also indicate new directions for research based on connections among genes that may otherwise be overlooked or that would require much time and effort to be found manually. Their paper presents an analysis of select gene communities found for colon cancer. For example, they show that COX-1 and COX-2, isoforms of cyclooxygenases, are correctly placed in different communities as they are involved in different mechanisms. They also suggest possible connections between a set of phospholipase A2 genes and the gene FACL4 in the context of this disease.

Table 5 Wilkinson and Huberman (2004)

Type ID	Object type	Object representation/link derivation	Notes
O1	Disease	Disease terms (and corresponding MEDLINE records with these terms)	–
O2	Gene	Occurrences of names and aliases in O1 documents	–
L1	Set of O2	Membership in a common community	
IM	*Identify connected components by splitting the network using criteria based on "betweenness"*		

3.6 Analysis of Gene-Disease LBD Approaches

Objects: Given our selection of papers, the key objects are in all cases genes and diseases. However, we observe differences in representation across the studies. Representing a disease by its MeSH term or by its free-text terms can make a significant difference. For example, *alopecia areata* retrieves 1,768 records when limited to MeSH whereas it retrieves 17% more records (2,069) when searched without any limits. Also diseases are sometimes represented by their document set and sometimes by the documents sets corresponding to their pathological conditions. With genes we see an even greater variety in representation – from the occurrences of gene names and aliases in MEDLINE to gene sequence and gene expression data. Note also that the set of genes considered itself may vary. For example, G2D considers only those sequences that map to the disease chromosomal region while BITOLA allows this as an option.

Links: Even greater variability is seen in the types of links utilized. First, it is not surprising to see co-occurrence used for predicting disease–gene connections (Perez-Iratxeta et al., Adamic et al.) as co-occurrence is widely used in information extraction and text mining research. Examples include the efforts on predicting gene–gene relationships [18], transcription factor associations [19], and protein–protein interactions [7]. In each of the studies reviewed here some statistical assessment is undertaken to gauge significance of the proposed relationship. Typically this is some comparison of the observed co-occurrence frequency and the frequency expected assuming that the two objects are randomly paired.

Consider now the approach that takes advantage of intermediate conceptual bridges (links) such as through chemical terms (G2D), through eVOC (anatomical) terms and through terms representing cellular functions (BITOLA). In effect, these methods require particular varieties of semantics to tie the disease and the gene. An open question is how to determine the advantages gained by these semantic constraints when compared with co-occurrence based efforts. Likely false positives drop due to these requirements. However, is it the case that text mining, designed to moor on the fringe of the known, is better served by less constrained methods? Possibly this question can only be answered empirically.

Comparing the studies that use intermediate links also begs the question as to which type of connector is more effective. A point to note in this regard is that in G2D the disease and gene are at least four steps apart as its logic takes one from a disease to its pathological conditions to chemicals to GO annotations to RefSeq sequences to homologous sequences in a chromosomal region. Whereas in the eVOC approach only three steps are involved. Now is it the case that with every additional step there is an added risk of error? More fundamentally how do these different connectors, chemical and functional links as in G2D and anatomical links as in the eVOC system, compare? Could the GO cellular component vocabulary be used as effectively in G2D? Note that one could also use genes themselves as connectors between diseases and other genes. For example, in Chilibot [8] this strategy has been used to discover connections between phenomenon (such as long-term potentiation)

and other genes. Genes as connectors are also implicit in the research by (Wilkinson and Huberman) on finding gene communities for diseases.

Clearly one direction of research is to explore more vocabularies as potential connectors. A related research direction is also to use vocabularies in combination which would require the study of evidence combination models. The choice of connector(s) could also be problem specific, i.e., depending on what is know about the particular disease. In that case a more general approach leaving the choice of conceptual bridge to the user as in BITOLA might be the best. This is another area where more research could be done. Moreover, as we accumulate experience with vocabularies for connectors, we might even begin to identify preferred characteristics (in addition to semantics). For instance, it may be that the more specific term subsets of a GO vocabulary are of greater value. Or, perhaps terms with high usage are more important. Observe that weights in the eVOC system are directly related to frequency of annotation with the term.

Inference methods: Differences in how measures exploit co-occurrence data are obvious but probably not as significant as other differences that may be observed. For example, inference methods that rely on a single path (BITOLA and G2D for links between GO terms and diseases) are categorically different from those that favour multiple paths. The eVOC system expects to find at least $m - n$ bridging anatomical concepts out of the n characterizing a disease. That is a tighter constraint. In a sense an extension of this notion is found in the research of Wilkinson and Huberman where the level of interest in a gene depends upon the 'community' to which it belongs.

A second major aspect to consider is one that is almost never addressed in text mining sytems. This aspect arises in the context of symmetric versus asymmetric methods. Given a system, can one expect to get consistent results whether we start from a disease seeking genes or we start from a gene seeking diseases? Take for example BITOLA that uses an asymmetric measure. Given the way in which confidence scores are computed, it is not clear if compatible results will be obtained. These scores are conditional probabilities that rely on the starting condition and the condition at each node of the path leading to the target. Thus directionality will matter. Of course, there is a natural perspective on a given problem, namely, the perspective of the user. However, it may be the case that for a given disease D, a gene G is identified as most interesting. Whereas from the perspective of the same gene, D may not be the most interesting disease. Perhaps one possible approach with asymmetric strategies would be to traverse both directions for a given problem and take the intersection of the top ranking suggestions.

Additional Knowledge Sources: While making explicit the additional sources used, our framework also suggests alternative designs. For example, the genes (O3 in Table 2) in the eVOC study could be represented by MEDLINE searches and L1 could be the co-occurrence in MEDLINE of eVOC terms and gene names. Assessing this strategy would at least tell us about the added value of using expression data from Ensemble. Alternatively, diseases could be represented by their description in

OMIM (and optionally also the linked documents) and the eVOC terms could be identified in these descriptions. As a final example, one could perhaps extend both the G2D system and the eVOC study to consider communities of genes where the edges between genes are drawn as a function of their common tissue expression patterns or their sequence similarities.

4 General Purpose Biomedical LBD Systems

We now consider LBD systems designed as general purpose biomedical text mining systems. Such systems are not tied to specific applications such as the discovery of gene–disease associations or protein–protein interactions. ARROWSMITH, BITOLA, IRIDESCENT, LitLinker and Manjal are some of the key domain independent systems available for use on the Web. In addition there are the research efforts of Weeber et al. [27] and Gordon and Lindsay [10], among others, which have extended and explored LBD strategies. Except for Manjal, the LBD systems listed above are (likely to be) described in detailed in other chapters of this book. Hence they will only be briefly reviewed here. In addition we present Manjal, a general purpose LBD system that we have developed at the University of Iowa.

BITOLA [13, 14] has also been described in the context of gene–disease links. As shown in Table 6 the objects of interest in BITOLA are topics represented by MeSH terms. The later version (2003) expands this to include genes as represented by their names and aliases. Open and Closed discovery are offered but the greater emphasis appears to be on using Open discovery to identify indirect relationships. BITOLA computes support and confidence from the association rules formalism to gauge the association strength between concept pairs. ARROWSMITH is likely the oldest general purpose LBD system implemented. There are presently two versions of ARROWSMITH, viz., University of Chicago version and University of Illinois – Chicago version. Both implement Closed discovery. The University of Chicago version allows one to upload two sets of retrieved MEDLINE records corresponding to two topics. These sets are then compared to find the list of intersecting title words, phrases and MeSH terms. These intersections are presented to the user as a ranked list where the ranking strategy also considers the common MeSH terms between the two starting query topics. The key difference in the University of Illinois – Chicago version is that the literature search step is integrated into the discovery system.

As described on their web site, LitLinker[12], considers objects represented by MeSH terms and implements Open discovery. According to the description in [21], they exploit correlations between terms calculated using the Apriori algorithm [3]. Finally there is the commercial system IRIDESCENT [29, 30]. It includes genes; diseases, disorders, syndromes or phenotypes; chemical compounds and small molecules; and drug names as objects. Although not truly a general purpose system, we include it here given its variety of objects and, we believe, its extensibility. Across a few papers they experiment with different probabilistic measures

[12] http://litlinker.ischool.washington.edu/

Table 6 General purpose biomedical LBD systems

Type ID	Object type	Object representation/link derivation	Notes
		BITOLA	
O1	Topic	MeSH concept and retrieved MEDLINE records	–
L1	Pairs of topics	Co-occurrence	A
L2	Pairs of topics	Implicit: connections through intermediate MeSH terms of specified semantic type	A
IM	*Open discovery with confidence and support weights*		
		ARROWSMITH (U. Chicago and U. Illinois – Chicago)	
O1	Topic	PubMed search and retrieved MEDLINE records	–
L1	Pairs of topics	Co-occurrence	
L2	Pairs of topics	Implicit: connections through shared title words, phrases, MeSH terms	
IM	*Closed discovery with frequency based weights*		
		LitLinker	
O1	Topic	MeSH term and retrieved MEDLINE records	–
L1	Pairs of topics	Co-occurrence	A
L2	Pairs of topics	Implicit: connections through intermediate MeSH terms of specified semantic type	A
IM	*Open discovery with weights calculated using support*		
		IRIDESCENT	
O1	Disease	OMIM entries and retrieved MEDLINE records	–
O2	Genes	Entrez Gene entries and retrieved MEDLINE records	–
O3	Chemicals	MeSH concepts and retrieved MEDLINE records	–
O4	Drugs	(from FDA) and retrieved MEDLINE records	–
O5	GO terms	(from GO) and retrieved MEDLINE records	–
L1	Pairs of objects	Co-occurrence	
L2	Pairs of objects	Implicit: connections through other objects	
IM	*Open discovery with probabilistic weights*		
		Manjal	
O1	Topic	PubMed search and topic profile from retrieved MEDLINE records	–
O2	Topic	MeSH concept and topic profile from retrieved MEDLINE records	–
L1	Pairs of topics	Co-occurrence	
L1	Pairs of topics	Profile similarity	
L3	Pairs of topics	Implicit: related through other topics	A
IM	*Open discovery, closed discovery, multi-topic analysis, bipartite topic analysis TFIDF weights*		

of association that may be used to gauge the relatedness between a pair. In [30], for example, they study mutual information and extend these in two ways to assess the value of implicit relationships identified using an Open discovery model. The extensions consider the different pathways connecting the two objects. Using IRIDESCENT they found, for example, that the drug chlorpromazine, which is normally used to treat problems such as psychotic disorders and also severe hiccups, would also reduce the progression of cardiac hypertrophy.

4.1 Manjal

Manjal, available on the Web[13] offers a variety of LBD options for mining MEDLINE. In each option a user specifies one or more topics, where a topic is any valid PubMed search. For each input topic provided, Manjal first retrieves records from MEDLINE after which it builds its profile, which is simply a weighted vector of terms. Terms are from the MeSH or/and RN fields of the records. Profile terms are assigned TF*IDF weights with cosine normalization. All the text mining functions offered in Manjal operate on top of these profiles. In essence, these functions can make use of similarities calculated between topic profiles, employ MeSH terms and their profiles as bridges between topics and relate topics based on shared MeSH terms. In addition Manjal also offers co-occurrence based analysis.

Manjal users may conduct Open discovery runs starting with a single topic. The end result is a ranked list of MeSH terms, organized by semantic type. Each ranked MeSH term represents a topic that might have an interesting (and implicit) connection to the starting topic. Bridging topics are also presented. In the Closed discovery option two topics are provided as input and their MeSH profiles are created. The MeSH terms they share and their combined weights provide the tentative bridges between the two starting topics. A third function extends the notion of the two-topic Closed discovery function to work with larger sets of topics. Profiles are built for each topic and the neighborhood of any given topic may be explored. Neighborhoods may be selected on the basis of profile similarity or on the basis of co-occurrence frequency. Manjal's user interface is graphical and interactive. Both nodes and links may be clicked to obtain further details including for example, the corresponding set of PubMed documents.

An upgraded beta version of Manjal (not public, access available by request) offers additional functions. For example, it allows analysis of bipartite topic sets. This is appropriate when the problem naturally breaks down into two groups of topics. For example, the two sets of topics could be a set of diseases and a set of genes, or a set of environmental toxins and a set of diseases etc. Using this function one may for example, rank members of one set in terms of their association with members of the other set.

In all of the above functions the user may constrain the text mining process by specifying the types of connections desired. Indeed it is desirable to do so as otherwise the process could generate an overwhelming amount of information. This is done by allowing only terms from certain UMLS semantic types to participate in the process. For example, in Open discovery the intermediate terms may be restricted to those of type *Cell Function* or *Gene or Genome*.

Manjal has tested successfully on a set of "benchmark" LBD problems that derive from the research of Swanson and Smalheiser [23]. This replication study is the most extensive performed thus far. Manjal has also been used to propose a beneficial relationship between the dietary substance *Curcumin Longa* also known as turmeric and disorders such as retinal diseases, Crohn's disease and problems of the spinal

[13] Manjal: http://sulu.info-science.uiowa.edu/Manjal.html

cord [24]. The postulated connections were through biochemical pathways involving several genes such as inflammatory genes. Interestingly, a recent pilot study [12] has been published where a pure curcumin preparation was administered to patients with ulcerative proctitis and patients with Crohn's disease. The authors conclude that the results encourage follow-up double-blind placebo-controlled studies.

4.2 Analysis of the General Purpose LBD Systems

Objects: We observe significant differences in how topics are conceptualized in these systems. With ARROWSMITH and Manjal users may start with any search that is legitimate in the PubMed system. However, the difference between these two is that in Manjal intermediate topics are defined by MeSH terms whereas with ARROWSMITH these are terms from the free-text fields. The remaining systems either limit themselves to MeSH for topic specification or to predefined objects whose names (or aliases) appear somewhere in the MEDLINE records. These differences are fundamental. For example, when the user is constrained to MeSH as input, complex queries such as *erythromycin AND antihistamines AND hypertension* cannot be considered. More generally, the space of input topics is unbounded with ARROWSMITH and Manjal. Whereas, with BITOLA and LitLinker, these are bounded by the MeSH vocabulary. A possible extension that remains consistent with the parameters of these systems is to allow for combinations of MeSH concepts as input topics. This would certainly remove some of the constraints, albeit at the cost of having to calculate various frequency based statistics at run time.

Links: As seen in LBD systems exploring gene–disease connections, there are differences in the way co-occurrence is used (or not) to define links. However, what is more interesting is the remarkable absence of "semantic" links. For example, although IRIDESCENT identifies disease sets, gene sets etc. from curated resources, it appears to ignore the links between the two object types available from say OMIM. The larger question to address concerns the extent to which these LBD systems may benefit from the inclusion of expert acknowledged relationships as available in curated databases. One option may be to utilize known relationships harvested from sources such as Entrez Gene and OMIM to build a network of associated objects. This could then be the basis of Open and Closed discovery algorithms. A second option could be to use a hybrid approach that allows one to smoothly incorporate both semantic relationships along with co-occurrence based information. Exploring evidence combination models is then an important research direction.

Points raised earlier about identifying implicit relationships from single paths versus multiple paths also apply here. IRIDESCENT, for example, probabilistically assesses the strength of the association between the target topic (in Open discovery) and the collection of intermediate topics connecting to the starting topic. Manjal also calculates a weight that is a function of the number and importance rating of intermediate paths. In contrast BITOLA, for example, offer single link

paths. Their relative merits remain unknown though at the intuitive level strategies favouring multiple paths may be more reliable.

Inference Methods: The final aspect to (re-)consider is that of symmetric versus asymmetric inferencing strategies. Again, it is not clear as to the role of direction in all of these systems. Whether one starts with a gene topic looking for novel diseases or vice versa, it is unclear if compatible results are obtained. This too remains an open research area in the context of these general purpose LBD systems.

5 Predicting Relationships from the Web

We now turn our attention to efforts on discovering novel links from the Web. The main emphasis has been on discovering connections between people. However, some effort has been devoted to finding connections involving companies and industries as also between web pages.

5.1 Adamic and Adar (2003)

Adamic and Adar [1] aim to predict relationships between students at MIT and Stanford based on the similarity in characteristics extracted from their home pages. Specifically, the authors use the text, hyperlinks (inlinks and outlinks), and mailing list subscriptions on the web pages to "profile" students. Each individual feature is weighted by the inverse log of its frequency. Profile similarity is computed as the sum of the weights of the features in common. The authors also analyze predictions based on the individual feature types and find that the text of the home pages acts as the best predictor of a relationship. Table 7 represents the key objects and links.

Table 7 Adamic and Adar (2003)

Type ID	Object type	Object representation/link derivation	Notes
O1	Students at Stanford	(i) Text words in Home Page (ii) Inlinks in Home Page (iii) Outlinks in Home Page (iv) Mailing lists in Home Page (v) Composite of above	–
O2	Students at MIT	(i) Text words in Home Page (ii) Inlinks in Home Page (iii) Outlinks in Home Page (iv) Mailing lists in Home Page (v) Composite of above	–
L1	O1, O2	Sum of weights of features in common	
IM	*Similarity in profiles*		

Table 8 BenDov et al. (2004)

Type ID	Object type	Object representation/link derivation	Notes
O1	Person	Name and retrieved news items as identified by ClearForest software	–
L1	Pairs of people	Co-occurrence in sentence	U
L2	Pairs of people	(from L1) Implicitly via other people	U
IM	*Transitive relationships*		

Further analysis of the predictions made for the Stanford students can be found online[14]. Note that this research does not rely on co-occurrence as the home pages of two individuals are unlikely to overlap. We regard this research as LBD albeit working off non-traditional "documents".

5.2 Ben-Dov et al. (2004)

Working off approximately 9,100 documents from four news sites: CNN, BBC, CBS and Yahoo, the authors of this paper [5] identify novel relationships between person entities. Two entities are explicitly connected if they co-occur in a sentence. Two entities are implicitly connected if they form part of a transitive structure with an intermediate entity. For example they connect Osama Bin Laden and Pope John Paul via Ramzi Yousef. Bin Laden is connected to Yousef in several ways. For example one document mentions that Yousef stayed in Bin Laden's house. Two documents mention a book by Simon Reeve called "The New Jackals: Ramzi Yousef, Osama bin Laden and Future of Terrorism". Yousef is connected to the Pope by reports on an attempted assassination. They identify entities in the news articles using an information extraction tool (ClearForest[15]). They also extract semantic links between entities using NLP-based methods such as by identifying patterns involving noun phrases, verbs, etc. However, they do not use semantic links for knowledge discovery. Table 8 shows the details.

5.3 Cory (1997)

Although the research described in this paper [9] addresses the humanities domain in general, the author also focuses on relationships between people. This work is a direct application of Swanson's Open discovery approach to humanities data obtained

[14] http://www.hpl.hp.com/research/idl/papers/web10/frequency.html
[15] http://www.clearforest.com/

Table 9 Cory (1997)

Type ID	Object type	Object representation/link derivation	Notes
O1	Writer	Name and retrieved records from humanities index	–
L1	Pairs of writers	Co-occurrence	U
L2	Pairs of writers	(from L1) Implicitly linked via other writers	U
IM	*Transitive relationships*		

from the humanities index (under the WILS database)[16]. The aim here is to identify new analogies. The author says that this application domain is difficult because the language of humanities articles is not as structured and formal as that of medical articles. The author notes that person names fits the bill and uses those. Starting from a person (A) Cory retrieves documents from the humanities index and then identifies people names (Bs) in the titles other than A itself. Semantics for the relations between the Bs and A are manually established from the documents and interesting Bs identified. For each of these Bs, Cory conducts fresh searches and identifies names (Cs) in the titles of the new articles retrieved and again relation semantics are manually established. Then via a transitive analysis the author connects A with interesting Cs. A relationship is novel if a query containing both A and C does not retrieve any documents (Table 9). Cory finds a novel analogy for the twentieth century writer Robert Frost in the form of a classical second century BCE Greek writer, Carneades, via a nineteenth century writer, William James (1842–1910).

5.4 Kumar et al. (1999)

In this paper [17] the authors describe an approach to identify online communities. They concentrate on "new" communities that are not yet established or are implicitly defined. By this they mean communities at a finer level of detail such as the community of turkish student organizations in the US. These communities are typically not yet listed on any web portal. Their operational definition of a community is a densely connected bipartite subgraph known as a 'core'. A core consists of fans that are pages with outlinks and centers that are pages with inlinks. Fans can be thought of as specialized hubs and centers are the pages with the required information. The authors define an iterative procedure that consists of many pruning steps. Starting from a large set of nodes they keep pruning until they identify a community (or core) in which both the fans and centers have a minimum number of outlinks and inlinks respectively. Using data crawled by Alexa, consisting of over 200 million web pages, they identify communities such as the Australian Fire Brigade Services. They also explore temporal analysis to verify how many communities,

[16] We acknowledge that we have stretched our definition of Web based LBD works to include this WILS database research. We do this given the innovativeness of the work and its direct use of LBD.

out of a random sample of 400 communities identified, survive for more than 18 months. Interestingly they found that approximately 70% of the communities were still alive (Table 10).

5.5 Tan and Kumar (2001)

Tan and Kumar [25] propose an approach to identify novel links between web pages from sequences in user session data. The aim is to help in restructuring web sites so as to better conform to the navigational behavior of users. They do this by identifying sequential and non-sequential indirect associations from user session data. They first identify all the frequent itemsets from the data (using algorithms such as the Apriori algorithm [3]). In the non-sequential case they postulate indirect associations between pairs of unrelated pages if they share intermediate sets of pages, called Mediator sets, that are frequently associated with them. In the sequential case they identify intermediate sequences, called mediator sequences, and infer indirect connections between pairs of unrelated pages that share mediator sequences via three kinds of inference mechanisms, viz., convergence, divergence and transitivity. These indirect connections suggest more optimal ways to structure web sites (Table 11).

5.6 Bernstein et al. (2002)

Bernstein et al. [6] address the goal of identifying relationships between companies, more specifically similarities, using a large corpus of business news. They combine information extraction techniques with network analysis and statistical approaches

Table 10 Kumar et al. (1999)

Type ID	Object type	Object representation/link derivation	Notes
O1	Web pages	URL address	–
L1	Pair of O1 objects	URL based connections	U
L2	Community of O1 objects	(from L1) 'Cores' with particular features	U
IM	*Presence in 'core' after pruning*		

Table 11 Tan and Kumar (2001)

Type ID	Object type	Object representation/link derivation	Notes
O1	Web page	URL	–
L1	Pair of O1 objects	Support based on co-occurrence frequency	
L2	Pair of O1 objects	Implicit: through intermediate sets	
L3	Pair of O1 objects	Implicit: through intermediate sequences	A
IM	*Convergence, divergence, and transitivity*		

Table 12 Bernstein et al. (2002)

Type ID	Object type	Object representation/link derivation	Notes
O1	Company	Company name and vector of co-occurring companies	–
O2	Industry	Average of company vectors	–
L1	Pairs of O1 objects	Similarity of vectors	
L2	O1, O2	Similarity of vectors	
L3	Pairs of O2 objects	Similarity of vectors	
IM	*Cosine similarity*		

to extract knowledge of company interrelationships. Distinct company names are identified in a news collection from a 4-month period. Companies, represented as nodes, are displayed in a co-occurrence network to provide a visual overview of their distribution and connectivity. Going further, they represent each company by a vector of its co-occurring companies and calculate cosine similarities between vectors. Note that in this approach two companies may not co-occur and yet show high similarity. They also use the same principles to explore the relationship between individual companies and different industries as well as between industries. An industry is regarded as a cluster of companies. An industry vector is defined as the average of the vectors of the companies that belong to it. Table 12 abstracts from their work the key features of their methods.

5.7 Analysis of Web Based LBD

The key objects considered are students, web pages, companies and industries. Students for example, were represented by their home pages optionally augmented with their inlinks and/or their outlinks. Several alternative representations may be considered. One could use their entries in blogs, the set of papers presented as seen in conference web sites or the web pages of related individuals such as professors and co-authors. Each variety of representation provides a different perspective on the individual student that may be useful in determining novel relationships.

With links we see an interesting variety, indicative of the broad potential with the Web. For example, in addition to exploiting URL-based links, we see relationships inferred from user-access data for Web sites. Similarly, one can imagine search logs being a good source of implicit relations between pages, web sites, products, organizations and possibly also between web users. However, the anonymous nature of search logs makes the detection of user connections difficult. With companies we see that in addition to generating a co-occurrence based network Bernstein et al. uses a second-order strategy that groups companies by calculating similarities based on company co-occurrence feature vectors. This allows for two companies to be very similar without co-occurring.

Beyond these object and link specific points, several general observations may be made. The first is that there is little research on the web involving methods that are the same or analogous to LBD. This is particularly striking when compared to the level of activity in the biomedical domain. There are several potential explanations for this discrepancy. First, the kinds of goals that may be targeted are not as easily specifiable on the web as in biomedicine. In biomedicine, especially in bioinformatics, the key entities are widely understood to include genes, diseases, proteins, chemicals etc. This understanding is reflected in the kinds of curated resources that have been created. Thus to look for novel associations, such as those between diseases and genes, or drugs and genes follows naturally from the set of key entities. Despite the heterogeneity of the Web, very few key entity types have been explored for LBD.

A second possible reason for the paucity of LBD research in the web domain is the ambiguity challenge. For example, a straightforward application of Open discovery would involve starting with a search on an input topic. Even when focussed on a person as the input topic, we will need to filter the retrieved set in order to disambiguate between the multiple individuals likely to share the same name. We note that this sub-problem is itself being directly addressed (e.g., [11]). Although ambiguity resolution is also required in biomedical LBD, the problem is far more pronounced on the Web, given its heterogeneity and especially given its much faster pace of growth.

A third possible explanation for the lack of LBD research on the Web is a very practical one, which is the non-availability of appropriate datasets. Researchers in the biomedical domain may easily avail of the MEDLINE database, PubMed APIs and the UMLS vocabularies. This has created an incredibly hospitable environment for LBD research. Added to this are the many curated resources such as Entrez Gene and OMIM, typically with an option for data downloads. In contrast, the Web is close to being inhospitable to LBD research. For example, API's to search systems such as Google or Yahoo! limit users to only 1,000 and 5,000 daily searches, respectively. Moreover, each search is limited to the top few results. Also these APIs do not provide all the search options that are implemented on the respective web sites. For example, the Google API does not allow blog search. Avoiding these search systems implies the need to crawl the web and develop ones own Web datasets. Collections such as Alexa crawls[17], available at a fairly low cost, are certainly in the right direction. But the real power of LBD is in identifying *novel* hypotheses which implies working with information that is *current*. Thus although working off predefined collections may help in refining methods, it is unlikely to be of real value to end users.

To counter these challenges, the web, with information on almost every type of entity, offers excellent opportunities for existing LBD methods. Consider the kinds of problems addressed in the papers reviewed. A key emphasis is on finding implicit relations between people: students from two universities (Adamic and Adar), authors across time (Cory), and individuals mentioned in news articles

[17] http://www.amazon.com/b/ref=sc_fe_c_0_239513011_1/103-4334540-2295806ie=UTF8&node=12782661& no=239513011&me=A36L942TSJ2AJA

(Ben-Dov et al.). There is also some emphasis on individual pages as key entities in terms of defining emerging communities of web pages and toward reshaping websites. There is surprisingly little work on companies. However, these are just the beginnings in terms of LBD on the Web. For example, even within the space of individuals, we have an open research forum given the different classifications of professions, affiliations etc.

The Web also offers excellent opportunities for developing new LBD methods. In fact, the development of new methods is almost inevitable given that each document (web page) is readily characterized not only by its content but also by its inlinks and outlinks. Interestingly the methods proposed by Tan and Kumar and by Kumar et al. consider mainly the URLs and links. It may be the case, for example, that by considering page content based similarities as well, more cohesive cores are identified by the latter's method. This is also somewhat indicated by the Adamic and Adar research, where the text of the home page is determined to be the best predictor of a relationship. LBD methods exploiting URLs may also contribute to biomedical LBD given the availability of fast growing full-text collections such as PubMed Central. Thus citations to and from biomedical articles may eventually be exploited for LBD.

6 Conclusions

We presented an overview of literature-based discovery methods using a common framework for analysis. The framework focusses on the key objects, links, inference methods and knowledge sources used. The analysis was presented in three parts. The first part was constrained to a single theme of finding novel gene–disease connections. In the second part we analyzed general purpose LBD systems in biomedicine. The third part analyzed the few papers that use LBD or analogous methods on the Web.

The framework allowed us to perform a focussed comparison and analysis of LBD methods. In the process several open questions and directions for research were identified. For example, in the gene–disease context an important question is on the relative merits of single link discovery paths versus multilink paths. Another important angle for research is on the design of evidence-combination models that consider multiple intermediate vocabularies. With general-purpose biomedical LBD systems an example open research direction is on incorporating semantic links from curated databases into the process. Links of interest include not only those extracted from texts but also those readily available in curated resources. Despite the prevalence of LBD research in biomedicine we still do not know the relative merits of implicit connections that are co-occurrence based versus those derived from more semantic/conceptual links. Also needed is research studying the implications of symmetric versus asymmetric LBD strategies. We believe that this question has been given little or no attention in the literature. As a consequence, there is the risk of underrating or overrating a hypothesis given the chosen direction of the LBD

analysis. This chapter also compares LBD on the web with LBD in biomedicine. It is clear that LBD on the web is at a very early stage. However, LBD opportunities are abundant, especially if we can cross a few of the key hurdles. Moreover, methods such as URL-based LBD strategies, developed on the Web have the potential to influence methods for biomedicine.

There are several limitations of the analysis presented in this chapter. As said initially this chapter is not a comprehensive review of LBD research. Thus for example, we ignored interesting problems such as identifying implicit drug–disease, protein–protein interactions. In the general-purpose LBD research, we reviewed only LBD systems as opposed to papers that presented strategies without having a freely accessible system. Also we did not focus on the types of experiments and the results obtained in each paper. Instead we considered primarily the key methodological details.

To conclude, our framework-based review provides a better understanding of the similarities and differences across LBD systems and methods. Through this endeavor, our own knowledge on the evolution of LBD research in different domains and some of the key hurdles has greatly improved. This chapter also raises several questions and identifies avenues for extending LBD research. Hopefully these will guide the efforts of the LBD research and development community.

Acknowledgements This material is partly based upon work supported by the National Science Foundation under Grant No. 0312356 award to Srinivasan. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003
2. Lada A. Adamic, Dennis Wilkinson, Bernardo A. Huberman, and Eytan Adar. A Literature Based Method for Identifying Gene-Disease Connections. In *Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB 2002)*, pp. 109–117, 2002
3. Rakesh Agrawal and Ramakrishnan Srikant. Mining Sequential Patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, pp. 3–14, 1995
4. Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J Micheal Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25:2529, 2000
5. Moty Ben-Dov, Wendy Wu, Ronan Feldman, and Paul A. Cairns. Improving Knowledge Discovery by Combining Text-Mining and Link-Analysis Techniques. In *Proceedings of the SIAM International Conference on Data Mining*, 2004
6. Abraham Bernstein, Scott Clearwater, Shawndra Hill, Claudia Perlich, and Foster Provost. Discovering Knowledge from Relational Data Extracted from Business News. In *Proceedings of Workshop on Multi-Relational Data Mining (MRDM 2002)*, 2002

7. Peter M. Bowers, Matteo Pellegrini, Mike J. Thompson, Joe Fierro, Todd O. Yeates, and David Eisenberg. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biology*, 5(R35), 2004
8. Hao Chen and Burt M. Sharp. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, 5(147), 2004
9. Kenneth A. Cory. Discovering hidden analogies in an online humanities database. *Computers and the Humanities*, 31:1–12, 1997
10. Michael D. Gordon, Robert K. Lindsay, and Weiguo Fan. Literature-based discovery on the World Wide Web. *ACM Transactions on Internet Technologies (TOIT)*, 2(4):261–275, 2002
11. Ramanathan V. Guha and A. Garg. Disambiguating People in Search. Technical Report, Stanford University, 2004
12. Peter R. Holt, Seymour Katz, and Robert Kirshoff. Curcumin therapy in inflammatory bowel disease: a pilot study. *Digestive Diseases and Sciences*, 50(11):2191–2193, 2005
13. Dimitar Hristovski, Borut Peterlin, Joyce A. Mitchell, and Susanne M. Humphrey. Improving literature based discovery support by genetic knowledge integration. *Studies in Health Technology and Informatics*, 95:68–73, 2003
14. Dimitar Hristovski, Janez Stare, Borut Peterlin, and Saso Dzeroski. Supporting discovery in medicine by association rule mining in medline and UMLS. *Medinfo*, 10(Pt 2):1344–1348, 2001
15. Yanhui Hu, Lisa M. Hines, Haifeng Weng, Dongmei Zuo, Miguel Rivera, Andrea Richardson, and Joshua LaBaer. Analysis of genomic and proteomic data using advanced literature mining. *Journal of Proteome Research*, 2:405–12, 2003
16. Thomas Karopka, Juliane Fluck, Heinz-Theodor Mevissen, and Änne Glass. The Autoimmune Disease Database: a dynamically compiled literature-derived database. *BMC Bioinformatics*, 7(325), 2006
17. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the Web for Emerging Cyber-Communities. In *Proceedings of the Eighth International Conference on World Wide Web (WWW-8)*, pp. 1481–1493, 1999
18. Holger Maier, Stefanie Döhr, Korbinian Grote, Sean O'Keeffe, Thomas, Werner, Martin Hrabé de Angelis, and Ralf Schneider. LitMiner and WikiGene: identifying problem-related key players of gene regulation using publication abstracts. *Nucleic Acids Research*, 33:W779–W782, 2005
19. Hong Pan, Li Zuo, Vidhu Choudhary, Zhuo Zhang, Shoi H. Leow, Fui T. Chong, Yingliang Huang, Victor W.S. Ong, Bijayalaxmi Mohanty, Sin L. Tan, S.P.T. Krishnan, and Vladimir B. Bajic. Dragon TF Association Miner: a system for exploring transcription factor associations through text-mining. *Nucleic Acids Research*, 32:W230–W234, 2004
20. Carolina Perez-Iratxeta, Peer Bork, and Miguel A. Andrade. Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, 31(3):316–319, 2002
21. Wanda Pratt and Meliha Yetisgen-Yildiz. Litlinker: Capturing Connections Across the Biomedical Literature. In *Proceedings of the International Conference on Knowledge Capture (K-CAP 2003)*, pp. 105–112, 2003
22. Aditya K. Sehgal, Xing Y. Qiu, and Padmini Srinivasan. Mining MEDLINE Metadata to Explore Genes and their Connections. In *Proceedings of the 2003 SIGIR Workshop on Text Analysis and Search for Bioinformatics*, 2003
23. Padmini Srinivasan. Text mining: generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology (JASIST)*, 55(5):396–413, 2004
24. Padmini Srinivasan and Bisharah Libbus. Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*, Suppl. 1:I290–I296, 2004
25. Pang-Ning Tan and Vipin Kumar. Mining Indirect Associations in Web Data. In *Proceedings of the Workshop on Mining Logdata Across All Customer Touchpoints (WEBKDD '01)*, pp. 145–166, 2001
26. Nicki Tiffin, Janet F. Kelso, Alan R. Powell, Hong Pan, Vladimir B. Bajic, and Winston A. Hide. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Research*, 33(5):1544–1552, 2005

27. Marc Weeber, Jan A. Kors, and Barend Mons. Online tools to support literature-based discovery in the life sciences. *Briefings in Bioinformatics*, 6(3):277–286, 2005
28. Dennis M. Wilkinson and Bernardo A. Huberman. A Method for Finding Communities of Related Genes. In *Proceedings of the National Academy of Sciences of the United States of America*, 101:5241–5248, 2004
29. Jonathan D. Wren. *The IRIDESCENT System: An Automated Data-Mining Method to Identify, Evaluate, and Analyze Sets of Relationships Within Textual Databases*. PhD thesis, University of Texas Southwestern Medical Center, 2003
30. Jonathan D. Wren, Raffi Bekeredjian, Jelena A. Stewart, Ralph V. Shohet, and Harold R. Garner. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, 20(3):389–398, 2004