Chapter 7
Literature-Based Discovery

Abstract Literature-Based Discovery (LBD) refers to a range of approaches that take a body of scientific literature as the input, apply a series of computational, manual, or a hybrid processes, and finally generate hypotheses that are potentially novel and meaningful for further investigations. This chapter introduces the origin of LBD, its major landmark studies, available tools, and resources. In particular, we explain the design and application of PKD4J to illustrate the principles and analytic decisions one typically needs to make. We highlight the recent developments in this area and outline remaining challenges.

Swanson's Pioneering Work

Swanson's work on Raynaud disease/fish-oil discovery exemplified the problem of mining undiscovered public knowledge from biomedical literature (Swanson 1986a). According to Swanson (1986a, b). LBD (a.k.a. UDPK) can be public, yet undiscovered, if independently created fragments of knowledge and information are logically related but never retrieved, interpreted, and studied together. In other words, when considered together, two complementary and non-interactive literature sets of articles (independently created fragments of knowledge) can reveal useful information of scientific interest not apparent in either of the two sets alone (Swanson 1986a, b).

Swanson formalizes the procedure to discover UPK from biomedical literatures as follows: Consider two separate literature sets, CL and AL, where the documents in CL discuss concept C and documents in AL discuss concept A. Both of these two literature sets discuss their relationship with some intermediate concepts B (also called bridge concepts). However, their possible connection via the concepts B is not discussed together in any of these two literature sets as shown in Fig. 7.1.

Swanson's UPK (or ABC) model can be described as the process to induce "A implies C", which is derived from both "A implies B" and "B implies C"; the

C. Chen and M. Song, *Representing Scientific Knowledge*,
https://doi.org/10.1007/978-3-319-62543-0_7

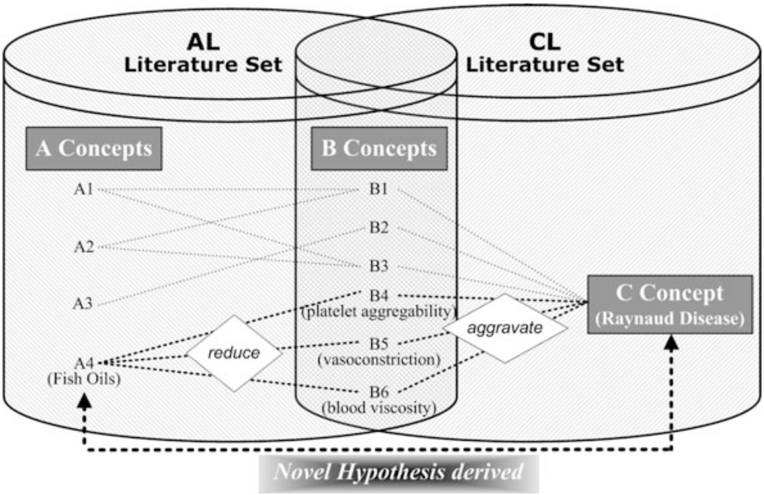


Fig. 7.1 Swanson's UPK model—the connection of fish oils and Raynaud disease

derived knowledge or relationship "A implies C" is not conclusive but hypothetical. For example, Swanson tried to uncover novel suggestions for what (B) causes Raynaud disease (C) or what (B) are the symptoms of the disease, and what (A) might treat the disease as shown in Fig. 7.1. Through analyzing the document set that discusses Raynaud disease he found that Raynaud disease (C) is a peripheral circulatory disorder aggravated by high platelet aggregation (B), high blood viscosity (B) and vasoconstriction (B). Then he searched these three concepts (B) against MEDLINE to collect a document set relevant to them. With the analysis on the document set he found out those articles show the ingestion of fish oils (A) can reduce these phenomena (B); however, no single article from both document sets mentions Raynaud disease (C) and fish oils (A) together. Putting these two separate literatures together, Swanson hypothesized that fish oils (A) may be beneficial to people suffering from Raynaud disease (C). This hypothesis that Raynaud disease might be treated by fish oil was hidden in the biomedical literature until Swanson uncovered through literature-based discovery. This novel hypothesis was later clinically confirmed by DiGiacomo et al. (1989). Later on, Swanson used the same approach to uncover 11 connections of migraine and magnesium (Swanson 1988).

One of the drawbacks of Swanson's method is that the method requires large amount of manual intervention and very strong domain knowledge, especially in the process of qualifying the intermediate concepts that Swanson names the "B" concepts. In order to reduce dependence on domain knowledge and human intervention and to automate the whole process as much as possible, several approaches have been developed to automate this discovery process based on Swanson's method (Lindsay and Gordon 1999; Pratt and Yetisgen-Yildiz 2003; Srinivasan 2004; Weeber et al. 2003). They have not only successfully replicated the Raynaud disease-fish-oil and migraine-magnesium discoveries, but also discovered new treatments for other diseases such as thalidomide (Weeber et al. 2003).

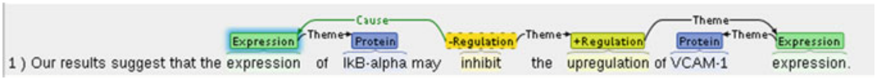
These research works have produced valuable insights into new hypothesis. On the other hand, substantial manual intervention is required to reduce the number of possible connections. We describe a fully automated approach for mining undiscovered public knowledge from biomedical literature. Our approach replaces ad hoc manual pruning with semantic knowledge from biomedical ontologies. We use semantic information to manage and filter the sizable branching factor in the potential connections among a huge number of medical concepts.

To efficiently find novel hypotheses efficiently and effectively from a huge search space of possible connections among the biomedical concepts, we need to first solve the problem of ambiguous biomedical terms. We utilize biomedical ontologies, namely UMLS and MeSH for this purpose. Our method requires minimal human intervention. Unlike other approaches (Hristovski et al. 2001; Pratt and Yetisgen-Yildiz 2003; Srinivasan 2004), our method only requires the user to specify the possible semantic relationships between the starting concept and the to-be-discovered target concepts rather than possible semantic types of the target concepts and the bridge concepts. Our method utilizes semantic knowledge (e.g., semantic types, semantic relations and semantic hierarchy) on bridge concepts and the target concepts to filter out irrelevant concepts and meaningless connections between concepts. Since there could be many plausible relationships between the bridge concepts and the target concepts, our method uses semantic relations to filter those relationships to identify desirable ones.

Major Trends of LBD

Swanson's pioneering work provides the framework on which almost all subsequent research in LBD is based (Cameron et al. 2013, Cohen et al. 2010, Malhotra et al. 2013, Spangler et al. 2014). The initial approach proposed by Swanson requires a laborious, time-intensive, manual process. The follow-up studies attempted to overcome these challenges by developing processes to make LBD easier and faster to perform and more automatic overall. Those studies proposed different techniques for concept extraction, computation of results, and sizes and types of input data. In LBD, human experts continue to play a significant role. New systems essentially follow Swanson's ABC model of discovery.

A recent trend in LBD is that more works has focused specifically on, and provided advancements in, automation of the LBD process. Using more advanced Natural Language Processing (NLP) techniques while at the same time exploiting metadata (e.g., from UMLS) has led to a reduction in the role of human experts (Wilkowski et al. 2011). Another trend is to use more advanced methods to capture important correlations between concepts. Hristovski et al. (2001) and Pratt and Yetisgen-Yildiz (2003) used an unsupervised machine learning algorithm (association rule mining) along with support and confidence metrics. In contrast,



1) Our results suggest that the expression of IkB-alpha may inhibit the upregulation of VCAM-1 expression.

Fig. 7.2 An example of Brat visualization of entity and relation

Wren et al. (2004) used statistical techniques to distinguish significant correlations. A related trend is the application of visualization. van der Eijk et al. (2004) differs from other work by giving a visual output directly to the user without the intermediate steps requiring human expert guidance. Overall, reducing reliance on human experts by increasing the degree of automation is an important recent trend in LBD research. The development of web-based visualization such as D3.js[1] and Brat[2] makes visualization of LBD scalable and accessible via web. The example of visualization with a PubMed sentence by Brat is shown in Fig. 7.2.

LBD Systems

We outline the design and functionality of three examples of LBD systems, namely the ArrowSmith developed in late 1990s, the BITOLA systems in mid 2000s, and the more recent Hypothesis Generator in 2015.

ArrowSmith

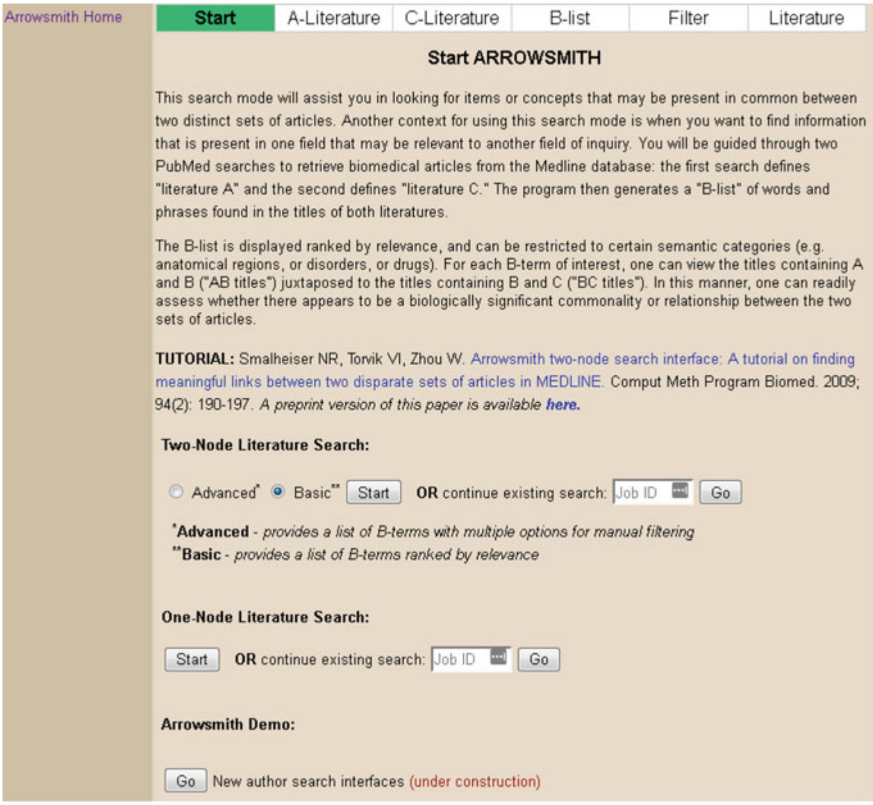
ArrowSmith is the very first LBD tool introduced by Swanson and Smalheiser (1997), which is publicly avaliable.[3] ArrowSmith provides a two-mode discovery method. The simple PubMed search function is available for the users to input two PubMed queries in order to define the two sets of articles A and C (Fig. 7.3).

To retrieve MEDLINE records corresponding to user queries in a fast mode, a local MEDLINE database was created. When a query is entered, the article ID numbers are downloaded from PubMed and the full MEDLINE records are retrieved from the local database, including a tokenized result of each article title after stopwords were removed. If articles are not found in the local database, then they are downloaded from PubMed as XML files, processed and stored in the local database. B-terms and their feature values are computed in a parallel mode by processing the sets of tokenized titles in chunks, and merging the results later on when each process is done. B-term features were pre-computed and stored in the term database for fast look-up.

[1] https://d3js.org/.

[2] http://brat.nlplab.org/features.html.

[3] http://arrowsmith.psych.uic.edu.



Start ARROWSMITH

This search mode will assist you in looking for items or concepts that may be present in common between two distinct sets of articles. Another context for using this search mode is when you want to find information that is present in one field that may be relevant to another field of inquiry. You will be guided through two PubMed searches to retrieve biomedical articles from the Medline database: the first search defines "literature A" and the second defines "literature C." The program then generates a "B-list" of words and phrases found in the titles of both literatures.

The B-list is displayed ranked by relevance, and can be restricted to certain semantic categories (e.g. anatomical regions, or disorders, or drugs). For each B-term of interest, one can view the titles containing A and B ("AB titles") juxtaposed to the titles containing B and C ("BC titles"). In this manner, one can readily assess whether there appears to be a biologically significant commonality or relationship between the two sets of articles.

TUTORIAL: Smalheiser NR, Torvik VI, Zhou W. Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. Comput Meth Program Biomed. 2009; 94(2): 190-197. *A preprint version of this paper is available* ***here.***

Two-Node Literature Search:

○ Advanced* ● Basic** Start **OR** continue existing search: Job ID Go

*__Advanced__ - *provides a list of B-terms with multiple options for manual filtering*
**__Basic__ - *provides a list of B-terms ranked by relevance*

One-Node Literature Search:

Start **OR** continue existing search: Job ID Go

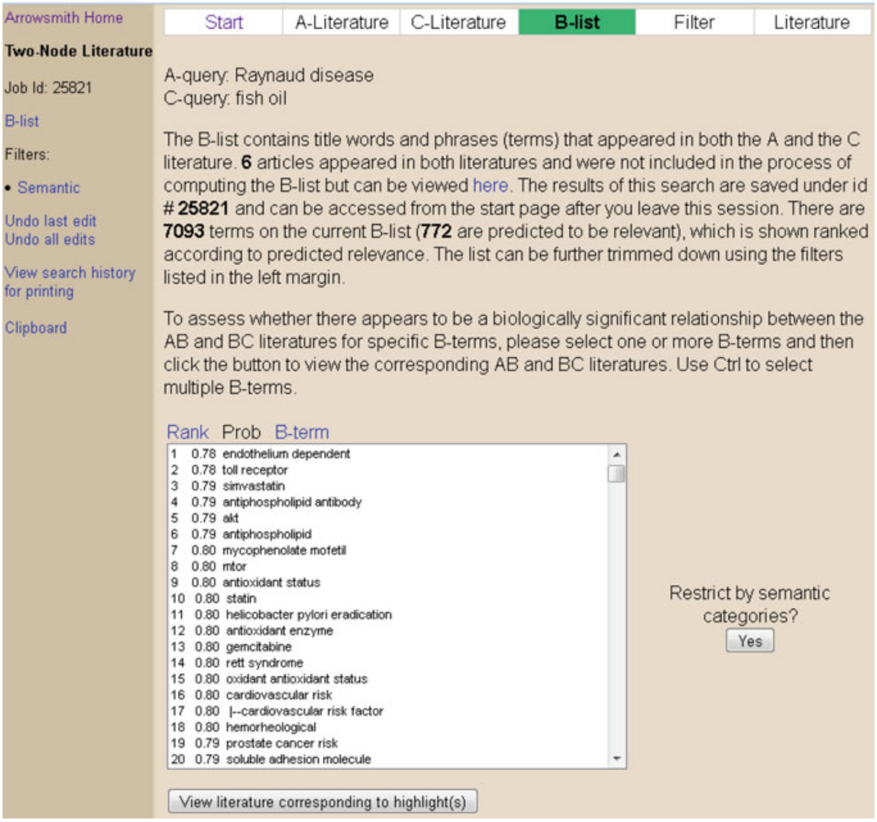
Arrowsmith Demo:

Go New author search interfaces (under construction)

Fig. 7.3 The homepage of ArrowSmith

For instance, if we choose “Raynaud’s disease” as the A-literature term and “Fish Oil” as the C-literature term, ArrowSmith returns the list of B terms after couple of minutes’ execution time. With “Raynaud’s disease” and “Fish oil” as A and C, ArrowSmith generates a total of 7093 B-terms that do not appear in both A and C literature (six articles that appeared in both A and C were excluded in the resulting b-term list). The list of B-terms is shown in the inner box of Fig. 7.4, which is sorted in order of predicted relevance score of a B-term that indicates a biological significance between the AB and BC literatures.

We can filter out the resulting B-term list by semantic types provided in UMLS. For instance, if we want to restrict the B-terms to the two semantic types, Activities & Behaviors and Anatomy, you can simply select the check box next to those two types once you click on “Restrict by semantic categories” button. It will result in the 730 B-terms that passed the filtering criteria (Fig. 7.5). Before clicking the button, you may want to scroll down the list to see if there are any non-highlighted B-terms that you want to keep. Use Ctrl to select additional B-terms.



A-query: Raynaud disease
C-query: fish oil

The B-list contains title words and phrases (terms) that appeared in both the A and the C literature. **6** articles appeared in both literatures and were not included in the process of computing the B-list but can be viewed here. The results of this search are saved under id # **25821** and can be accessed from the start page after you leave this session. There are **7093** terms on the current B-list (**772** are predicted to be relevant), which is shown ranked according to predicted relevance. The list can be further trimmed down using the filters listed in the left margin.

To assess whether there appears to be a biologically significant relationship between the AB and BC literatures for specific B-terms, please select one or more B-terms and then click the button to view the corresponding AB and BC literatures. Use Ctrl to select multiple B-terms.

Rank	Prob	B-term
1	0.78	endothelium dependent
2	0.78	toll receptor
3	0.79	simvastatin
4	0.79	antiphospholipid antibody
5	0.79	akt
6	0.79	antiphospholipid
7	0.80	mycophenolate mofetil
8	0.80	mtor
9	0.80	antioxidant status
10	0.80	statin
11	0.80	helicobacter pylori eradication
12	0.80	antioxidant enzyme
13	0.80	gemcitabine
14	0.80	rett syndrome
15	0.80	oxidant antioxidant status
16	0.80	cardiovascular risk
17	0.80	|--cardiovascular risk factor
18	0.80	hemorheological
19	0.79	prostate cancer risk
20	0.79	soluble adhesion molecule

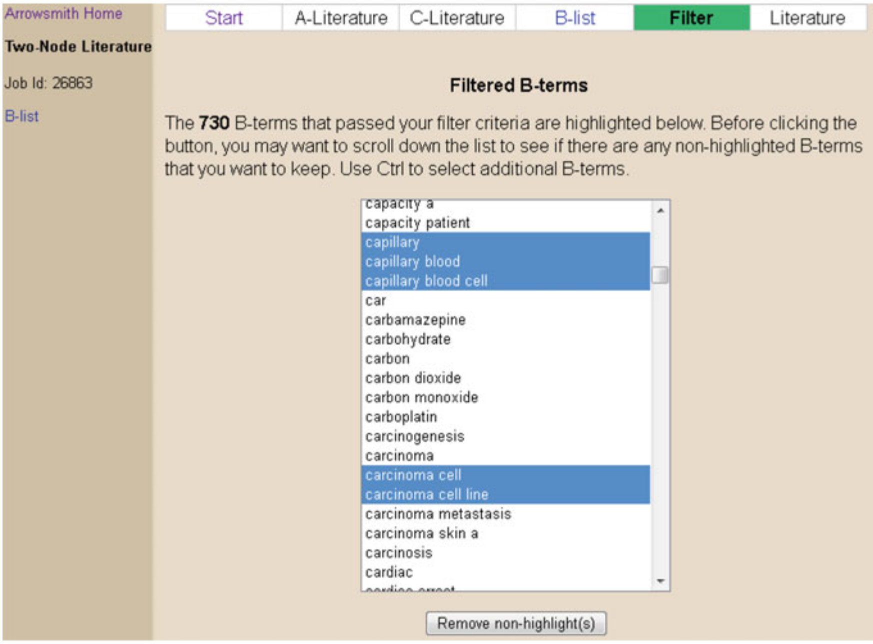
Fig. 7.4 The resulting B-term list for "Raynaud disease" and "Fish oil"

BITOLA

BITOLA is an web-based LBD system that has been around for about a decade (Hristovski et al. 2003), which is publicly available at.[4] The purpose of BITOLA is to help the biomedical researchers make new discoveries by discovering potentially new relations between biomedical concepts. The set of concepts contains MeSH and human genes from HUGO. BITOLA provides two discovery options: closed and open.

Open discovery allows the input of a single concept, then categories for first-order relatives of that concept, then categories for relatives of those first order concepts. Discovery algorithm for discovering new relations between medical concepts consists of the following five steps (Hristovski et al. 2001):

[4] http://arnika.mf.uni-lj.si/pls/bitola2/bitola.



Filtered B-terms

The **730** B-terms that passed your filter criteria are highlighted below. Before clicking the button, you may want to scroll down the list to see if there are any non-highlighted B-terms that you want to keep. Use Ctrl to select additional B-terms.

capacity a
capacity patient
capillary
capillary blood
capillary blood cell
car
carbamazepine
carbohydrate
carbon
carbon dioxide
carbon monoxide
carboplatin
carcinogenesis
carcinoma
carcinoma cell
carcinoma cell line
carcinoma metastasis
carcinoma skin a
carcinosis
cardiac
cardiac arrest

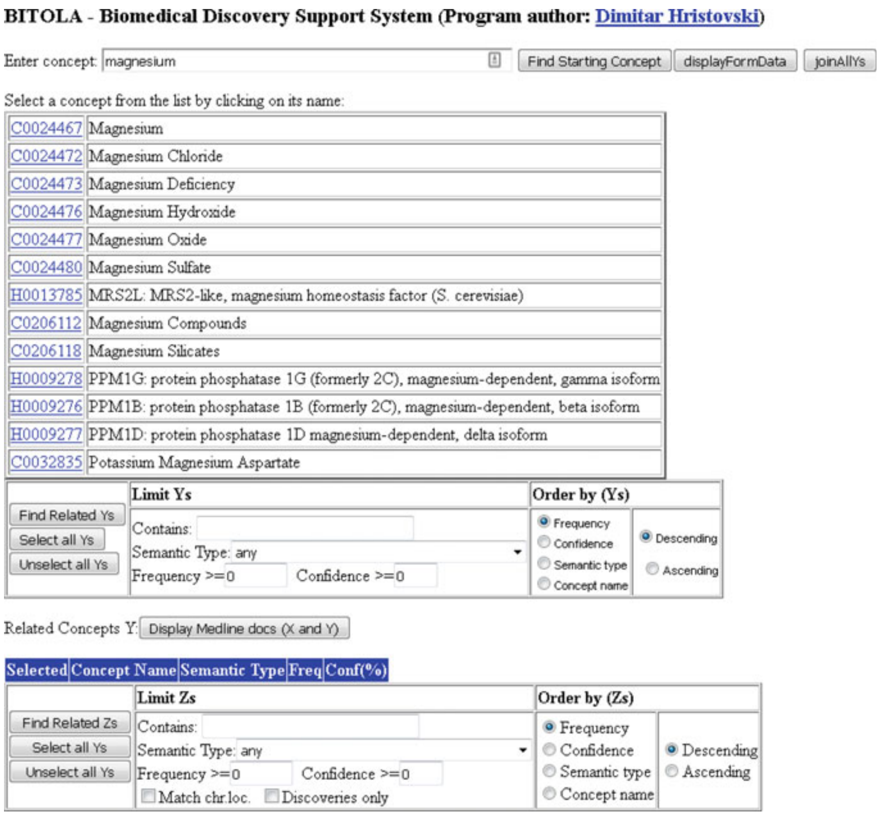
Fig. 7.5 Filtered B-terms

1. Given a starting concept of interest X
2. Find all concepts Y such that there is an association rule X → Y
3. Find all concepts Y such that there is an association rule Y → Z
4. Eliminate those Z for which an association X → Z already exists
5. The remaining concepts Z are candidates for an new relation between X and Z.

Because in MEDLINE each concept can be associated with many other concepts, the possible number of X → Z combinations can be extremely large. In order to deal this combinatorial problem, BITOLA applies filtering (limiting) and ordering functions to the discovery algorithm. The related concepts can be limited by the semantic type to which they belong and final possibility for limiting the number of related concepts or false related concepts is by setting thresholds on the support and confidence measures of the association rules. The main goal of the ordering is to present best candidates first to make human review as easy as possible (Hristovski et al. 2001).

For example, if Magnesium is the interest of search, type Magnesium and click on Find Starting Concept X in the BITOLA system, which will return a list of terms relevant to the query. As shown in Fig. 7.6, the query found 13 terms.

From the generated list, choose the very top one Magnesium, and BITOLA will fill in CUI (C0024467), the semantic type, and the chromosomal location automatically (if exists). Click on the button Find Related Zs, BITOLA will generate the



BITOLA - Biomedical Discovery Support System (Program author: Dimitar Hristovski)

Enter concept: magnesium

Select a concept from the list by clicking on its name:

C0024467	Magnesium
C0024472	Magnesium Chloride
C0024473	Magnesium Deficiency
C0024476	Magnesium Hydroxide
C0024477	Magnesium Oxide
C0024480	Magnesium Sulfate
H0013785	MRS2L: MRS2-like, magnesium homeostasis factor (S. cerevisiae)
C0206112	Magnesium Compounds
C0206118	Magnesium Silicates
H0009278	PPM1G: protein phosphatase 1G (formerly 2C), magnesium-dependent, gamma isoform
H0009276	PPM1B: protein phosphatase 1B (formerly 2C), magnesium-dependent, beta isoform
H0009277	PPM1D: protein phosphatase 1D magnesium-dependent, delta isoform
C0032835	Potassium Magnesium Aspartate

Limit Ys
Contains:
Semantic Type: any
Frequency >=0 Confidence >=0

Order by (Ys)
● Frequency ○ Confidence ○ Semantic type ○ Concept name
● Descending ○ Ascending

Related Concepts Y:

Selected	Concept Name	Semantic Type	Freq	Conf(%)

Limit Zs
Contains:
Semantic Type: any
Frequency >=0 Confidence >=0
□ Match chr.loc. □ Discoveries only

Order by (Zs)
● Frequency ○ Confidence ○ Semantic type ○ Concept name
● Descending ○ Ascending

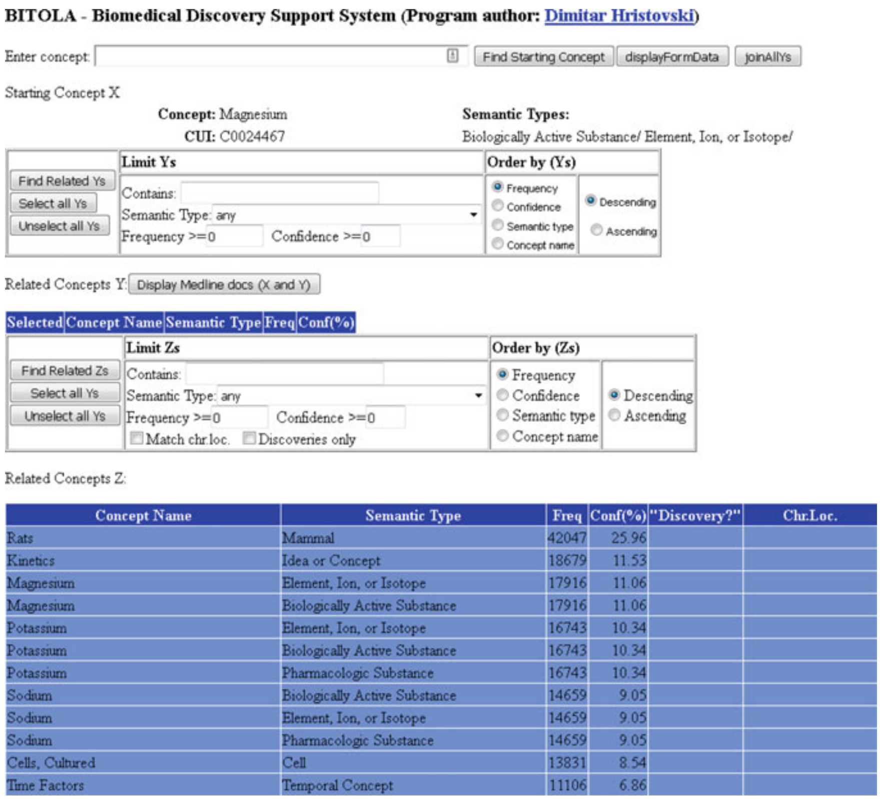
Fig. 7.6 The search results for the query *magnesium*

results, containing concept name, semantic type, frequency, confidence level, discovery, and chromosomal location (see Fig. 7.7).

Once a list of related concepts Zs is displayed, click the button Find Intermediate Ys, which will generate a list of substance terms that have been linked to Magnesium in some articles. See Fig. 7.8.

From this list of related concepts Ys, selecting the term Potassium with the semantic type of Pharmacologic Substance and clicking on the button Display Medline docs (X and Y) will display the two articles in PubMed about both Magnesium (X) and Potassium (Y). The user can explore other links, or re-run the query with other categories, so as to explore domains and chemicals that are linked to both Magnesium and Potassium.

In addition to the Closed Discovery option of BITOLA, the Open Discovery option of BITOLA allows the users to expand their inquiry into one node basis discovery. The Open Discovery option works quite similarly as the Closed Discovery one. The only difference is the structure. With closed discovery the user nominates X and Z then search for Y (limiting categories, if desired). With open



BITOLA - Biomedical Discovery Support System (Program author: Dimitar Hristovski)

Enter concept:

Starting Concept X

Concept: Magnesium
CUI: C0024467

Semantic Types:
Biologically Active Substance/ Element, Ion, or Isotope/

Limit Ys
Contains:
Semantic Type: any
Frequency >=0 Confidence >=0

Order by (Ys)
● Frequency
○ Confidence
○ Semantic type
○ Concept name
● Descending
○ Ascending

Related Concepts Y:

Selected	Concept Name	Semantic Type	Freq	Conf(%)

Limit Zs
Contains:
Semantic Type: any
Frequency >=0 Confidence >=0
□ Match chr.loc. □ Discoveries only

Order by (Zs)
● Frequency
○ Confidence
○ Semantic type
○ Concept name
● Descending
○ Ascending

Related Concepts Z:

Concept Name	Semantic Type	Freq	Conf(%)	"Discovery?"	Chr.Loc.
Rats	Mammal	42047	25.96		
Kinetics	Idea or Concept	18679	11.53		
Magnesium	Element, Ion, or Isotope	17916	11.06		
Magnesium	Biologically Active Substance	17916	11.06		
Potassium	Element, Ion, or Isotope	16743	10.34		
Potassium	Biologically Active Substance	16743	10.34		
Potassium	Pharmacologic Substance	16743	10.34		
Sodium	Biologically Active Substance	14659	9.05		
Sodium	Element, Ion, or Isotope	14659	9.05		
Sodium	Pharmacologic Substance	14659	9.05		
Cells, Cultured	Cell	13831	8.54		
Time Factors	Temporal Concept	11106	6.86		

Fig. 7.7 The results of the related concepts Z to "Magnesium"

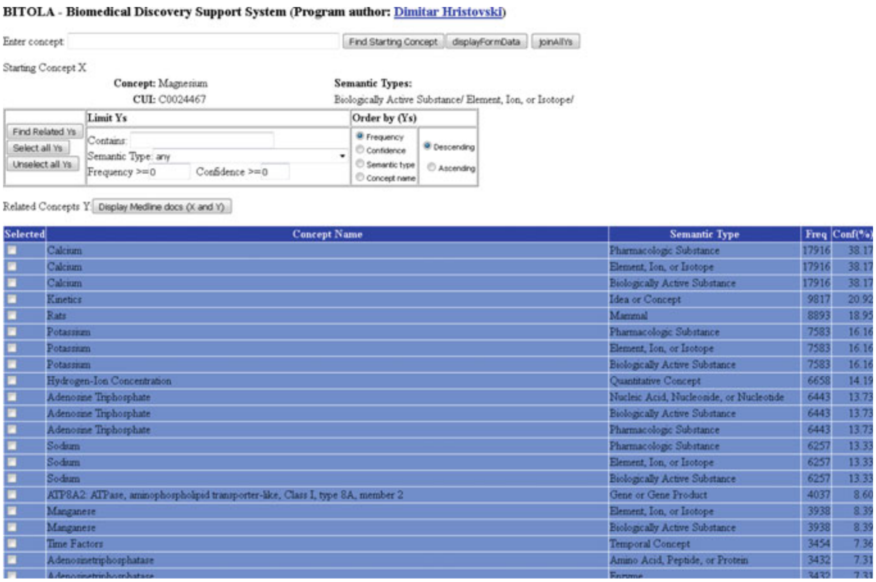
discovery, the user nominates X, then search for Y (limiting categories, if desired), then search for Z (limiting categories, if desired).

Hypothesis Generator

Hypothesis Generator is a recently developed LBD system that is based on PKDE4J (Song et al. 2015) for entity and relation extraction (Baek et al. 2017). Hypothesis Generator was originally developed to examine how lactosylceramide is associated with arterial stiffness. However, due to the flexibility of the system, hypothesis generator can serve as the general LBD system.

A brief instruction for hypothesis generator is as follows. First, the user types in one or more search terms, for example, "Raynaud disease" (Fig. 7.9).

The search function is backed by the Apache Lucene information retrieval system. Hypothesis generator indexed the 2015 version of MEDLINE records with



BITOLA - Biomedical Discovery Support System (Program author: Dimitar Hristovski)

Enter concept:

Starting Concept X

Concept: Magnesium **Semantic Types:**

CUI: C0024467 Biologically Active Substance/ Element, Ion, or Isotope/

Limit Ys

Contains:
Semantic Type: any
Frequency >=0 Confidence >=0

Order by (Ys)

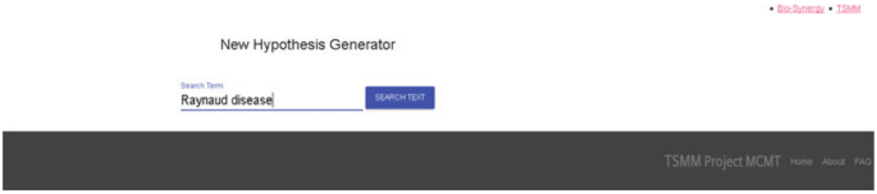
● Frequency ○ Confidence ○ Semantic type ○ Concept name

● Descending ○ Ascending

Related Concepts Y

Selected	Concept Name	Semantic Type	Freq	Conf(%)
□	Calcium	Pharmacologic Substance	17916	38.17
□	Calcium	Element, Ion, or Isotope	17916	38.17
□	Calcium	Biologically Active Substance	17916	38.17
□	Kinetics	Idea or Concept	9817	20.92
□	Rats	Mammal	8893	18.95
□	Potassium	Pharmacologic Substance	7583	16.16
□	Potassium	Element, Ion, or Isotope	7583	16.16
□	Potassium	Biologically Active Substance	7583	16.16
□	Hydrogen-Ion Concentration	Quantitative Concept	6658	14.19
□	Adenosine Triphosphate	Nucleic Acid, Nucleoside, or Nucleotide	6443	13.73
□	Adenosine Triphosphate	Biologically Active Substance	6443	13.73
□	Adenosine Triphosphate	Pharmacologic Substance	6443	13.73
□	Sodium	Pharmacologic Substance	6257	13.33
□	Sodium	Element, Ion, or Isotope	6257	13.33
□	Sodium	Biologically Active Substance	6257	13.33
□	ATP8A2: ATPase, aminophospholipid transporter-like, Class I, type 8A, member 2	Gene or Gene Product	4037	8.60
□	Manganese	Element, Ion, or Isotope	3938	8.39
□	Manganese	Biologically Active Substance	3938	8.39
□	Time Factors	Temporal Concept	3454	7.36
□	Adenosinetriphosphatase	Amino Acid, Peptide, or Protein	3432	7.31
□	Adenosinetriphosphatase	Enzyme	3432	7.31

Fig. 7.8 The list of related concepts Y to the target term "Magnesium"



New Hypothesis Generator

Search Term: Raynaud disease

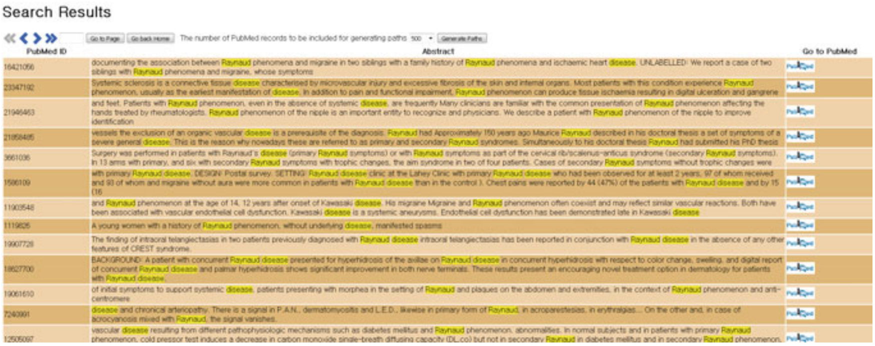
Fig. 7.9 The search homepage of the hypothesis generator

Lucene. The search term is highlighted in either the title or the abstract field (see Fig. 7.10).

PubMed ID for each result will be shown on the left and a direct link to the article is given on the right. The user can choose the number of PubMed records to be included for generating the paths.

On the search result page, the user can choose the number of PubMed records to extract entities from. This step is necessary since the current version of hypothesis generator extracts entities on the fly. In the future, extraction of entities will be done offline and stored in the database. If that is in place, this step will be eliminated. Once the number of records is chosen, you can click on the "generate paths" button, which will result in the follow result (Fig. 7.11).

The left panel shows the list of extracted entities and you can pick any two entities that you are interested into see the relation between two. Type in the entities that you want to conduct path analysis from the list of entity names. The left will be

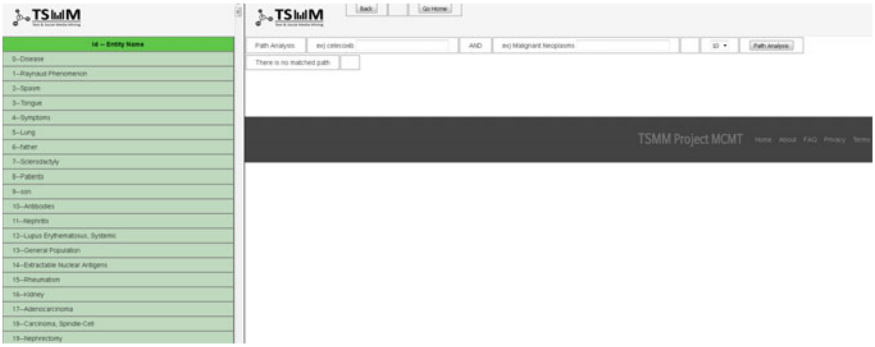


Search Results

The number of PubMed records to be included for generating paths 500

PubMed ID	Abstract	Go to PubMed
16421056	documenting the association between Raynaud phenomena and migraine in two siblings with a family history of Raynaud phenomena and ischaemic heart disease. UNLABELLED: We report a case of two siblings with Raynaud phenomena and migraine, whose symptoms	PubMed
23347192	Systemic sclerosis is a connective tissue disease characterised by microvascular injury and excessive fibrosis of the skin and internal organs. Most patients with this condition experience Raynaud phenomenon, usually as the earliest manifestation of disease. In addition to pain and functional impairment, Raynaud phenomenon can produce tissue ischaemia resulting in digital ulceration and gangrene	PubMed
21946463	and feet. Patients with Raynaud phenomenon, even in the absence of systemic disease, are frequently Many clinicians are familiar with the common presentation of Raynaud phenomenon affecting the hands treated by rheumatologists. Raynaud phenomenon of the nipple is an important entity to recognize and physicians. We describe a patient with Raynaud phenomenon of the nipple to improve identification	PubMed
21858485	vessels the exclusion of an organic vascular disease is a prerequisite of the diagnosis. Raynaud had Approximately 150 years ago Maurice Raynaud described in his doctoral thesis a set of symptoms of a severe general disease. This is the reason why nowadays these are referred to as primary and secondary Raynaud syndromes. Simultaneously to his doctoral thesis Raynaud had submitted his PhD thesis	PubMed
3661036	Surgery was performed in patients with Raynaud's disease (primary Raynaud symptoms) or with Raynaud symptoms as part of the cervical rib/scalenus-anticus syndrome (secondary Raynaud symptoms). In 13 arms with primary, and six with secondary Raynaud symptoms with trophic changes, the aim syndrome in two of four patients. Cases of secondary Raynaud symptoms without trophic changes were	PubMed
1586109	with primary Raynaud disease. DESIGN: Postal survey. SETTING: Raynaud disease clinic at the Lahey Clinic with primary Raynaud disease who had been observed for at least 2 years, 97 of whom received and 93 of whom and migraine without aura were more common in patients with Raynaud disease than in the control). Chest pains were reported by 44 (47%) of the patients with Raynaud disease and by 15 (16	PubMed
11903548	and Raynaud phenomenon at the age of 14, 12 years after onset of Kawasaki disease. His migraine Migraine and Raynaud phenomenon often coexist and may reflect similar vascular reactions. Both have been associated with vascular endothelial cell dysfunction. Kawasaki disease is a systemic aneurysms. Endothelial cell dysfunction has been demonstrated late in Kawasaki disease	PubMed
1119826	A young women with a history of Raynaud phenomenon, without underlying disease, manifested spasms	PubMed
19907728	The finding of intraoral telangiectasias in two patients previously diagnosed with Raynaud disease intraoral telangiectasias has been reported in conjunction with Raynaud disease in the absence of any other features of CREST syndrome.	PubMed
18627700	BACKGROUND: A patient with concurrent Raynaud disease presented for hyperhidrosis of the axillae on Raynaud disease in concurrent hyperhidrosis with respect to color change, swelling, and digital report of concurrent Raynaud disease and palmar hyperhidrosis shows significant improvement in both nerve terminals. These results present an encouraging novel treatment option in dermatology for patients with Raynaud disease.	PubMed
19061610	of initial symptoms to support systemic disease, patients presenting with morphea in the setting of Raynaud and plaques on the abdomen and extremities, in the context of Raynaud phenomenon and anti-centromere	PubMed
7240991	disease and chronical arteriopathy. There is a signal in P.A.N., dermatomyositis and L.E.D., likewise in primary form of Raynaud, in acroparestesias, in erythralgias... On the other and, in case of acrocyanosis mixed with Raynaud, the signal vanishes.	PubMed
12505097	vascular disease resulting from different pathophysiologic mechanisms such as diabetes mellitus and Raynaud phenomenon, abnormalities. In normal subjects and in patients with primary Raynaud phenomenon, cold pressor test induces a decrease in carbon monoxide single-breath diffusing capacity (DL,co) but not in secondary Raynaud in diabetes mellitus and in secondary Raynaud phenomenon.	PubMed

Fig. 7.10 The search result page for the query “Raynaud disease”

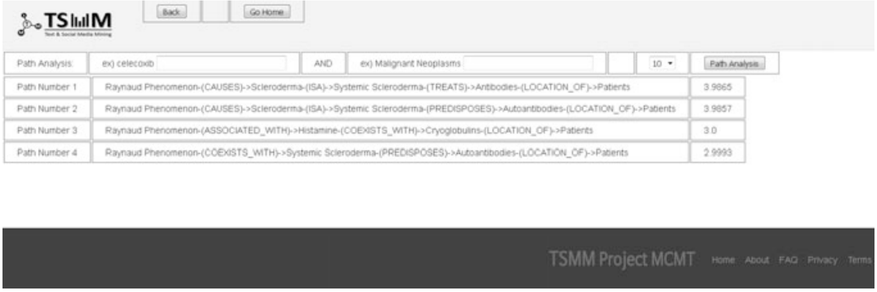


Id -- Entity Name
0--Disease
1--Raynaud Phenomenon
2--Spasm
3--Tongue
4--Symptoms
5--Lung
6--father
7--Sclerodactyly
8--Patients
9--son
10--Antibodies
11--Nephritis
12--Lupus Erythematosus, Systemic
13--General Population
14--Extractable Nuclear Antigens
15--Rheumatism
16--Kidney
17--Adenocarcinoma
18--Carcinoma, Spindle-Cell
19--Nephrectomy

Path Analysis: ex) celecoxib AND ex) Malignant Neoplasms 10

There is no matched path

Fig. 7.11 The results of extracted entities (left) and the path analysis start page (right)



Path Analysis:	ex) celecoxib	AND	ex) Malignant Neoplasms		10	Path Analysis

Path Number 1	Raynaud Phenomenon-(CAUSES)->Scleroderma-(ISA)->Systemic Scleroderma-(TREATS)->Antibodies-(LOCATION_OF)->Patients	3.9865
Path Number 2	Raynaud Phenomenon-(CAUSES)->Scleroderma-(ISA)->Systemic Scleroderma-(PREDISPOSES)->Autoantibodies-(LOCATION_OF)->Patients	3.9857
Path Number 3	Raynaud Phenomenon-(ASSOCIATED_WITH)->Histamine-(COEXISTS_WITH)->Cryoglobulins-(LOCATION_OF)->Patients	3.0
Path Number 4	Raynaud Phenomenon-(COEXISTS_WITH)->Systemic Scleroderma-(PREDISPOSES)->Autoantibodies-(LOCATION_OF)->Patients	2.9993

Fig. 7.12 The path analysis result

the ‘A-term’ and the right will be the ‘C-term’ of your path. The user can choose the number of path you want to analysis as shown. For instance, if “Raynaud Phenomenon” is chosen as the A-term and “Patients” as the C-term, then the ‘Path Analysis’ will generate the results as shown in Fig. 7.12.

For “Raynaud Phenomenon” and “Patients” as A and C, respectively, the system returns four paths. The relation type between the entities is shown in the parenthesis. Importance of each path is determined by the overall semantic relatedness score. The overall relatedness score is computed by the average of a Phenomenon and Scleroderma. Pair 2 is Scleroderma and Systemic Scleroderma. Pair 3 is Systemic Scleroderma and Antibodies. Pair 4 is Antibodies and Patients. The relation type between Systemic Scleroderma and Antibodies is CAUSES. The relation type between Scleroderma and Systemic Scleroderma is IS-A. The relation type between Systemic Scleroderma and Antibodies is TREATS. The relation type between Antibodies and Patients is LOCATION_OF.

PKD4J: A Scalable and Flexible Engine

PKDE4J stands for Public Knowledge Discovery Engine for Java, is a scalable, flexible text mining system for public knowledge discovery (Song et al. 2015). The main task of PKDE4J is to extract entities and their relations from the unstructured text. PKDE4J extends Stanford CoreNLP written in Java (Manning et al. 2014). PKDE4J addresses the information overload problem that modern text mining systems promise to solve by automating the process of understanding the relevant parts of the scientific literature. Key tasks pertinent to the information overloading problem include increasing the efficiency of searching for information, facilitating the creation of large-scale models of the relationships of biomedical entities, and allowing for automated inference of new information as well as hypothesis generation to guide biomedical research.

Design Principle

The primary design principle is to make PKDE4J as scalable and flexible as possible. Song et al. (2015) used the pipeline architecture for developing PKDE4J. Unlike other text mining systems for LBD, PKDE4J is a configuration based system so that various different combinations of text processing components are readily enabled for different tasks. For example, for the problem of drug-disease interaction, we can use SIDER (http://sideeffects.embl.de/) for drug dictionary and KEGG (http://www.genome.jp/kegg/disease/) for disease dictionary. Another layer of flexibility is that entities can be extracted either by exact or approximate match. On top of the exact matching based entity extraction, bio entity can be extracted either by approximate matching-based, supervised-learning only, or a mixture of supervised-learning and dictionary. PKDE4J overcomes the problems of the dictionary-based approach by applying regular expression rules and N-gram to extract entities. Second, PKDE4J is a flexible extraction system that can be applied to different extraction tasks such as multi-class entity extraction, Protein-Protein Interaction (PPI), trigger extraction, etc.

Most of the current approaches are focused heavily on a specific application to solve a specific kind of problem. PKDE4J is designed to address the aforementioned issue by developing an extensible rule engine based on dependency parsing for relation extraction. It provides a rule configuration file that contains 17 rules to identify whether relation exists in a sentence and determine its relation type. Since a relation extraction task requires an unique set of extraction rules, one single optimized prediction model is only effective in a certain condition. For instance, a different model is required for the task of whether a sentence contains relation or not from the task of event extraction. In such scenario, supervised learning may not be the best option since for each task, a different classification model needs to be built. Thus, a flexible, plug-and-play module for a rule engine is the best option for different extraction tasks in an efficient manner.

Architecture

PKDE4J consists of four major components. The overall architecture of PKDE4J illustrates the connections between these components (Fig. 7.13).

The first component is preprocessing of input text. PKDE4J supports a verity of text formats, which includes PubMed in XML, PubMed Central in XML, ClincalTrials.gov in XML, and text data in CSV. The second component is entity extraction, including dictionary-based, supervised learning-based, a combination of dictionary with ontology like UMLS, and a combination of supervised learning-based with UMLS. The third component is relation extraction, which is based on a dependency tree-based rules. The fourth component is the storage and retrieval of the results from the entity and relation extraction components. The results are stored in a relational database in the format that can be used for visualization.

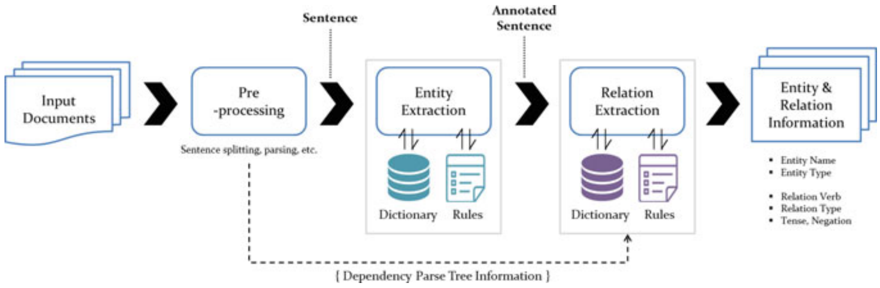


Fig. 7.13 The overall architecture of PKDE4J. *Source* Song et al. (2015)

Preprocessing

The preprocessing component covers various text processing tasks. The first one is tokenization. PKDE4J uses the Penn Treebank 3 (PTB) tokenization implemented in Stanford CoreNLP. PTBTokenizer is based on JFlex for an efficient, fast, and deterministic tokenization.

The second preprocessing task is sentence boundary detection. PKDE4J uses a Maximum Entropy model trained with the GENIA corpus for sentence splitting.

The third task is Part-Of-Speech (POS) tagging. PKDE4J uses the Stanford POS tagging algorithm for this task. The Stanford POS tagging algorithm is based on a flexible statistical CRF model.

The fourth task is lemmatization aided by Stanford CoreNLP. The fifth task is normalization of tokens. Token normalization is required since text contains various non-alphanumeric characters which may hinder the quality of entity extraction. The sixth task is n-gram matching. PKDE4J adopts the Apache Lucene ShingleWrapper algorithm, which constructs n-gram tokens from a token stream. The seventh task is approximate string matching. Approximate string matching may be needed when input text contains many spelling variations for the same entity name. PKDE4J extends the Soft-TFIDF algorithm that is a hybrid similarity measure introduced by Cohen et al. (2003).

Entity Extraction

Figure 7.14 shows the overall architecture of entity extraction component that consists of several steps. Step 1 is to load dictionaries. Dictionary loading is required when you choose the dictionary-based approach for entity extraction over other approaches. Depending on the target entities to be extracted, a list of

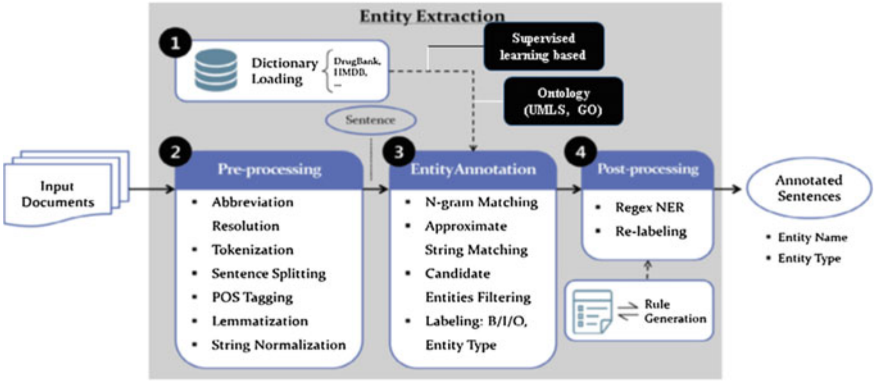


Fig. 7.14 Entity extraction component. An extended version of Song et al. (2015)

dictionaries are determined. Step 2 is preprocessing, which was described in the preprocessing component. Step 3 is entity annotation where the entity matching takes place between tokenized n-grams and dictionary entries. In entity annotation, there are four different options: (1) dictionary only, (2) a combination of dictionary with ontology, (3) supervised learning only, and (4) a combination of supervised learning with ontology. Step 4 is post-matching. For further improvement of extraction quality, PKDE4J uses the regular expressions to match the entities that are not found by dictionary. The regular expression rules define cascaded patterns over token sequences, which provides a flexible extension of the traditional regular expression language defined over strings.

Relation Extraction

The relation extraction component relies heavily on a set of dependency parsing based rules. Dependency parse trees provide a useful structure for the sentences by annotating edges with dependency types, e.g. subject, auxiliary, modifier. Dependency parse trees embed various information of dependencies within sentences, i.e. between words that are far apart in a sentence. The relation extraction module consists largely of three steps (See Fig. 7.15).

Step 1 is loading couple of dictionaries that contain biologically meaning verbs such as up-regulate, down-regulate, simptomize, etc. and nominalized terms like expression. Biologically meaningful verbs are classified into several categories and each category may have a few types (Table 7.1). The relation extraction component detects biologically meaningful verbs from sentences and map them to either categories or types, depending on the configuration setting.

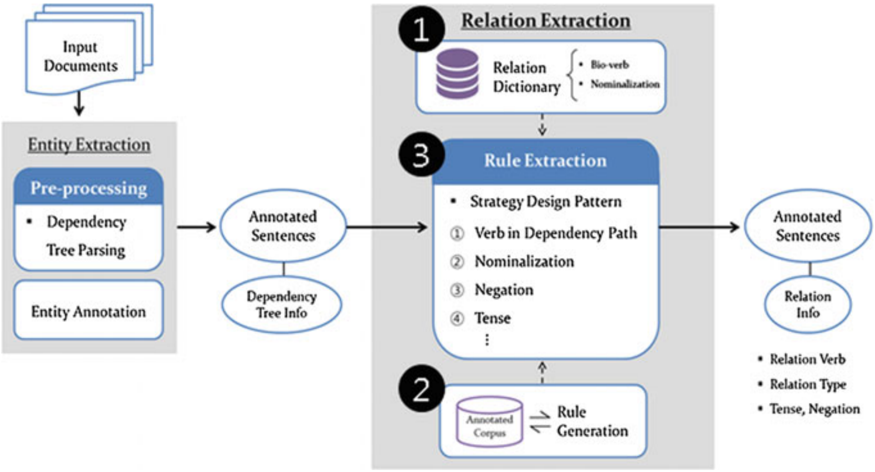


Fig. 7.15 Relation extraction component. *Source* Song et al. (2015)

Table 7.1 Classification of the biologically meaningful verb list

Category	Type	Verb example
Positive	Increase	Activate, promote, stimulate
	Transmit	Transport, link
	Substitute	Replace
Negative	Decrease	Inactivate, inhibit, block, arrest
	Remove	Breakbond, release, omit
Neutral	Contain	Embed, include, constitute
	Modify	Reconstitute, mutate, oxidize
	Method	Bleach, precipitate, coprecipitate
	Report	Prove, suggest, compare
Plain	Plain	Acquire, underlie, fix

Step 3 applies a set of relation rules to parsed dependency trees. After preprocessing, PKDE4J traverses the resulting dependency tree in postorder to find the relation triplets by using predefined set of relation rules for a dependency tree. In PKDE4J, each rule is called a strategy, which echoes the strategy design pattern adopted from Object-oriented system development. A strategy design pattern is particularly useful for creating objects which represent various strategies and a context object whose behavior varies as per its strategy object. In PKDE4J, a strategy represents a dependency tree-based relation rule. By applying a predefined set of strategies to each sentence, PKDE4J applies 17 predefined rules to the sentence, which generates a set of relation features such as relation type, tense, and negation for any given two entities located in the sentence (See Table 7.2).

Storing the Results of Extraction

At the last stage of pipeline, PKDE4J generates two major outputs. The first output is the extracted entities and the second output is the extracted relations. These outputs are stored in the relational database for further analysis. Table 7.3 shows the example of extracted entities. The example is a simplified version of output that only show PMID, entity name, entity type, and sentence where the entity is located

Table 7.2 A list of strategies that characterize relation between two entities

① Verb in dependency path	⑩ Number entities between entities
② No verb in dependency path	⑪ Entities in between
③ Detect nominalization	⑫ Surface distance
④ Weak nominalization	⑬ Entity counts
⑤ Negation	⑭ Same head
⑥ Tense (active/passive)	⑮ Entity order
⑦ Contain clause	⑯ Full tree path
⑧ Clause distance	⑰ Path length
⑨ Negation clause	

Table 7.3 Example of output of extracted entities

PMID	Entity	Type	Sentence
28482223	Phentolamine	DRUG	Phentolamine is one of the most representative nonselective αadrenoreceptor blocking agents, which have been proved to be owned various pharmacological actions
28482223	protein	FOOD	With the aid of multiple biophysical techniques, this scenario was to detailed explore the potential biorecognition between phentolamine and the hemeprotein in the cytosol of erythrocytes, and the influences of dynamic characters of protein during the bioreaction
28482223	protein	FOOD	Biorecognition can induce fairly structural transformation (selfregulation) of protein conformation

in. In addition to those four attributes, there are other attributes available such as beginning and ending position of entity as the results of entity extraction.

The second output is the relation extraction result shown in Table 7.4. The output consists of PMID, relation type, left entity name, left entity type, right entity name, right entity type, verb, voice, negation, and sentence where two entities are located in.

Table 7.4 Example of output of extracted relations

Field	Value 1	Value 2
PMID	8447197	27983686
Relation Type	PLAIN	RESULT_OF
Entity 1	Alcohol	Dairy
Entity 1 Type	FOOD	FOOD
Entity 2	Alcoholic	Drink
Entity 2 Type	FOOD	FOOD
Verb	Play	Containing
Tense	ACTIVE	ACTIVE
Negation	POSITIVE	POSITIVE
Sentence	Many variables, aside from the amount and duration of alcohol consumption, play a role in the development and progression of alcoholic liver disease (ALD)	In a placebo controlled, randomized, crossover study, 35 healthy males received either six placebo gelatin capsules consumed with 200 mL of water, six capsules with 800 mg polyphenols derived from red wine and grape extracts, or the same dose of polyphenols incorporated into 200 mL of either pasteurized dairy drink, soy drink (both containing 3.4% proteins) or fruit flavored protein free drink

Recent Developments and Remaining Challenges

Recently, LBD research has paid attention to deep learning as an effort to improve the quality of discovery. Rather et al. (2017) applied a word embedding technique called Word2Vec to the LBD problem. They used the MRDEF subset of UMLS Metathesaurus to train the Word2Vec model and reported a 23% overlap between their approach and MRREL. Deep learning has also been applied to the task of phenotyping (Che et al. 2015) used to identify patient subgroups based on individual clinical markers. Žitnik et al. (2013) conducted a study on non-negative matrix factorization techniques for fusing various molecular data to uncover disease-disease associations and show that available domain knowledge can help reconstruct known and obtain novel associations. Despite the recent interests in deep learning, it is still premature. More advanced studies of the applications of deep learning to the LBD problems are needed to evaluate how deep learning can advance LBD research.

There are several remaining challenges in LBD. The first challenge is how to implement a comprehensive procedure to obtain manually labeled samples. Although state-of-the-art machine learning methods have been utilized to automate the process, current approaches still observe degraded performance in the face of limited availability of labeled samples that are manually annotated by medical experts. Another major challenge is the convergence of multi-disciplinary teams that are pertinent to LBD. Although collaboration among researchers from various different fields is prevalent in LBD, it is often observed that development is separated from evaluation and end-usage of the tool developed. The third challenge is the standardization of evaluation. Evaluation in LBD is often ad hoc based and no general guidelines are established for LBD researchers to follow. Although there is a movement of standardization such as PubAnnotation,[5] we still need to put much effort into setting up the guidelines for LBD research.

References

Baek SH, Lee D, Kim M, Lee JH, Song M (2017) Enriching plausible new hypothesis generation in PubMed. PLoS ONE 12(7):e0180539

Cameron D, Bodenreider O, Yalamanchili H, Danh T, Vallabhaneni S, Thirunarayan K, Sheth AP, Rindflesch TC (2013) A graph-based recovery and decomposition of Swanson's hypothesis using semantic predications. J Biomed Inform 46:238–251. doi:10.1016/j.jbi.2012.09.004

Che Z, Kale D, Li W, Bahadori MT, Liu Y (2015) Deep computational phenotyping. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 507–516 (ACM, 2015)

Cohen T, Whitefield GK, Schvaneveldt RW, Mukund K, Rindflesch T (2010) EpiphaNet: an interactive tool to support biomedical discoveries. J Biomed Discov Collab 5:21–49

[5] http://pubannotation.org/.

Cohen WW, Ravikumar P, Fienberg SE (2003) A comparison of string metrics for matching names and records. In: Paper Presented at the International Conference on Knowledge Discovery and Data Mining (KDD) 09, Workshop on Data Cleaning, Record Linkage, and Object Consolidation

DiGiacomo RA, Kremer JM, Shah DM (1989) Fish-oil dietary supplementation in patients with Raynaud's pheomenon: a double blind, controlled, prospective study. Am J Med 86:158–164

Hristovski D, Peterlin B, Džeroski S, Stare J (2001) Literature based discovery support system and its application to disease gene identification. In: Proceeding AMIA Symposium 928

Hristovski D, Peterlin B, Mitchell JA, Humphrey SM (2003) Improving literature based discovery support by genetic knowledge integration. Stud Health Technol Inform 95:68–73

Lindsay RK, Gordon MD (1999) Literature-based discovery by lexical statistics. J Am Soc Inf Sci 50(7):574–587

Malhotra A, Younesi E, Gurulingappa H, Hofmann-Apitius M (2013) 'HypothesisFinder:' a strategy for the detection of speculative statements in scientific text. PLoS Comput Biol 9(7): e1003117. doi:10.1371/journal.pcbi.1003117

Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014) The stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp 55–60

Pratt W, Yetisgen-Yildiz M (2003) LitLinker: capturing connections across the biomedical literature, K-CAP'03, pp 105–112, Sanibel Island, FL, 23–25 Oct 2003

Rather NN, Patel CO, Khan SA (2017) Using deep learning towards biomedical knowledge discovery. Int J Math Sci Comput (IJMSC) 3(2):1–10. doi:10.5815/ijmsc.2017.02.01

Song M, Kim WC, Lee DH, Heo GE, Kang KY (2015) PKDE4J: entity and relation extraction for public knowledge discovery. J Biomed Inform 57:320–332

Spangler S, Wilkins AD, Bachman BJ, Nagarajan M, Dayaram T, Haas P, Regenbogen S, Pickering CR, Comer A, Myers JN, Stanoi I, Kato L, Lelescu A, Labrie JJ, Parikh N, Lisewski AM, Donehower L, Chen Y, Lichtarge O (2014) Automated hypothesis generation based on mining scientific literature. In: Paper Presented at the Proceedings of the 20th ACM SIGKDD International Conference on Knowledge discovery and data mining, New York, NY, USA

Srinivasan P (2004) Text mining: generating hypotheses from MEDLINE. J Am Soc Inf Sci 55 (4):396–413

Swanson DR (1986a) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 30:7–18

Swanson DR (1986b) Undiscovered public knowledge. Libr Q 56(2):103–118

Swanson DR (1988) Migraine and magnesium: eleven neglected connections. Perspect Biol Med 31(4):526–557

Swanson DR, Smalheiser NR (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif Intell 91:183–203

van der Eijk C, Van Mulligen E, Kors JA, Mons B, Van den Berg J (2004) Constructing an associative concept space for literature-based discovery. J Am Soc Inf Sci Technol 55(5):436–444

Weeber M, Vos R, Klein H, de Jong-Van den Berg LT, Aronson AR, Molema G (2003) Generating hypotheses by discovering implicit associations in the literature: a case report for new potential therapeutic uses for Thalidomide. J Am Med Inf Assoc 10(3):252–259

Wilkowski B, Fiszman M, Miller CM, Hristovski D, Arabandi S, Rosemblat G, Rindflesh TC (2011) Graph-based methods for discovery browsing with semantic predications. In: AMIA Annual Symposium Proceedings, pp 1514–1523

Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR (2004) Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics 20(3):389–398

Žitnik M, Janjić V, Larminie C, Zupan B, Pržulj N (2013) Discovering disease-disease associations by fusing systems-level molecular data. Sci Rep 3