Discovering Relational Knowledge from Two Disjoint Sets of Literatures Using Inductive Logic Programming

Supphachai Thaicharoen, Tom Altman, Katheleen Gardiner, and Krzysztof J. Cios

Abstract— **Literature-based discovery for hypothesis generation is a subarea of text mining that aims to discover novel or previously-unknown knowledge from two complementary but disjoint (CBD) sets of literatures. The discovery approach is based on Swanson's discovery models where indirect connections between two disjoint sets of literatures A and C could be found through a set of common terms B extracted from A and C. In this paper, we report an application of an Inductive Logic Programming (ILP), specifically the WARMR algorithm, to the field of literature-based discovery. The application extends Swanson's closed discovery model to uncover potentially meaningful knowledge in forms of relational frequent patterns that may exist after the connections between the two sets of literatures are found. We conducted an experiment between two pairs of topics: Raynaud's disease and fish oils, and Down syndrome and cell polarity. The experimental results demonstrate that our method can be used to enhance a literature-based discovery approach by providing potentially meaningful knowledge in addition to the indirect connections.**

I. Introduction

LITERATURE-based discovery for hypothesis generation is a subarea of text mining that aims to discover novel or previously-unknown knowledge from literature. It was first introduced in 1986 by Swanson, who discovered links between two sets of unrelated literatures by finding common terms or concepts present in both sets [1]. The key feature of the method is stated as: given two unrelated topics *A* and *C*, *A* is indirectly connected with *C* if there exists at least a term *B* that commonly occurs in retrieved literature set **A** and retrieved literature set **C**.

A number of interesting studies have been conducted that extended Swanson's ABC discovery methods. Lindsay and Gordon used lexical statistics for extracting and ranking terms in the common *B*-term list [2]. Pratt and Yetisgen-Yildiz employed biomedical concepts instead of words or phrases [3]. Srinivasan proposed a concept profile for knowledge discovery [4]. Jin and Srihiri introduced concept chains [5]. Hristovski *et al.* exploited frequent item sets and association rules in their knowledge discovery approach [6], whereas Stegmann and Grohmann utilized co-word clustering for hypothesis generation [7]. Finally, Wren extended the mutual information for ranking terms in the common *B*-term list [8]. In this paper, we present a further enhancement of Swanson's method by discovering additional meaningful knowledge after the indirect connections are found. Mining knowledge from two sets of literatures can thus be transformed into a problem of relational data mining.

Relational data mining is an area of data mining that is used to extract knowledge or patterns from a relational database. It could be typically attained by the use of Inductive Logic Programming (ILP). ILP is an inductive machine learning technique at the intersection of machine learning and logic programming [9]. Given examples and background knowledge as inputs, an ILP system generates output as a relational description of the examples using predicate logic. There are two main categories of ILP tasks, predictive and descriptive. A predictive ILP system generates hypothesis in terms of relational rules that can be used for classification of new examples. A descriptive ILP system aims to discover regularities in a set of examples. The advantages of ILP over typical data mining methods are its expressive data representations and ability to incorporate background knowledge. PROGOL [10], FOIL [11], and WARMR [12], [13] are among the existing important ILP algorithms.

In this study, we are interested in extracting knowledge in forms of relational frequent patterns which correspond to frequent Prolog queries. Discovery of relational frequent patterns was introduced by Dehaspe and Toivonen [12]. Several studies and applications have used this approach. Liakata and Pulman employed WARMR to automatically learn domain theories in terms of association rules for company succession events from a parsed text corpus, and then used Finite State Automata (FSA) to construct a graphical representation of the learned theories [14]. Blaťák implemented a relational rule mining algorithm called dRAP for mining long first-order frequent patterns from distributed text data [15]. He transformed the mined patterns into propositional representation before solving two text mining tasks: context-sensitive text correction of English language and morphological disambiguation of the Czech language. King *et al.* applied WARMR to find frequent patterns from a database of chemical compounds tested for carcinogenicity in rodents [16]. In their study, each compound was described by a combination of various numbers of the following five logical relations, atom-bond description, generic structural groups, genotoxicity, mutagenicity and structural indicators. King *et al.* encoded the frequent patterns generated by WARMR as attributes and used them for classification. They found that methods which used these encoded WARMR outputs as

Supphachai Thaicharoen is a Ph.D. student in Computer Science and Engineering Department at University of Colorado Denver. (email: supphachai.thaicharoen@email.ucdenver.edu).

Tom Altman is a professor in Computer Science and Engineering Department at University of Colorado Denver. (email: tom.altman@ucdenver.edu).

Katheleen Gardiner is a professor in Department of Pediatrics at University of Colorado Denver. Her research interests are Down syndrome and chromosome 21. (email: katheleen.gardiner@ucdenver.edu).

Krzysztof J. Cios is a professor at Virginia Commonwealth University and affiliated with IITiS Polish Academy of Sciences. His main research interests are bioinformatics and data mining. (email: kcios@vcu.edu).

978-1-4244-2765-9/09/$25.00 ©2009 IEEE

attributes in their predictive models were among the top three most accurate classifiers.

We propose here an extension of literature-based discovery for hypothesis generation using WARMR to find relational frequent patterns between two sets of disjoint literatures. With our approach, two tables were generated from a pair of complementary but disjoint (CBD) sets of retrieved literatures, and a third table was built from the set of common terms extracted from the two literature sets. First-order logic representation was then used to represent attributes in a table. We performed experiments using two pairs of topics: Raynaud's disease and fish oils, and Down syndrome and cell polarity. The experimental results demonstrate that our approach is complementary to a typical literature-based discovery method, and can be used to discover potentially meaningful knowledge in addition to indirect connections.

The paper is organized as follows. We present the technical background in Section II, namely, the WARMR algorithm and Swanson's ABC discovery models. Then, we describe the data and the pre-processing tasks, including our method, in Section III. Next, we present results from performed experiments and discuss them in Section IV. Finally, concluding remarks are given in Section V.

II. Technical background

A. WARMR

WARMR is an ILP data mining algorithm introduced by Dehaspe and Toivonen [12], [13]. It is an extension of the APRIORI association rule mining algorithm [17] for discovering frequent queries in a relational database. Given a relational database $\mathcal{D}$ and a declarative language bias definition $\mathcal{L}$, WARMR discovers queries that have the number of returned records greater than or equal to a user-specified minimum support (*minsup*), and generates relational association rules with a confidence value greater than or equal to a user-specified minimum confidence (*minconf*). The declarative language bias definition is used for defining admissible query patterns, which limits the search space of WARMR.

In WARMR, Prolog is used to represent queries, patterns, rules and database, and a term *query extension* is used for relational association rule.

A *query extension* is an existentially quantified implication. Given a Prolog clause,

$$?\text{-}\ l_1, ..., l_m. \rightarrow ?\text{-}\ l_1, .., l_m, l_{m+1}, ..., l_n.,$$

where $1 \leq m < n$, a *query extension* can be formulated by Equation 1.

$$?\text{-}\ l_1, ..., l_m \rightsquigarrow l_{m+1}, ..., l_n., \tag{1}$$

where $?\text{-}\ l_1, ..., l_m.$ is called the body, $l_{m+1}, ..., l_n.$ the head and $?\text{-}\ l_1, .., l_m, l_{m+1}, ..., l_n.$ the conclusion of the query extension.

WARMR is a level-wise algorithm based on breath-first search in the lattice spanned by a specialization operator or a subsumption relation $\preceq$. A pattern "p_1 is more general than p_2" or "p_2 is more specific than p_1", if $p_1 \preceq p_2$. In addition, $p_1 \preceq p_2$ if and only if $p_2 \models p_1$, where $\models$ is the logical implication operator.

Similar to APRIORI, WARMR iterates between candidate generation and candidate evaluation phases. In the candidate generation phase, candidates are derived from frequent patterns generated from the previous steps using a refinement function. These new candidates are valid if they (i) conform to the declarative language bias definition $\mathcal{L}$, (ii) are not more specific than infrequent patterns, and (iii) are not the same as already-generated frequent queries. In the candidate evaluation phase, frequencies of candidates are calculated with respect to the database. Candidates with frequencies greater than or equal to the user-specified minimum support are added to the frequent pattern set. Otherwise, they are added to infrequent pattern set.

In contrast to APRIORI algorithm that employs a subset relation, WARMR utilizes the notion of θ-subsumption relation for pruning candidates. θ-subsumption is a stronger variants of $\preceq$ relation. A query q_1 θ-subsumes q_2 if and only if there exits a set of variable/value pairs, $\{X_1/x_1, X_2/x_2, ..., X_n/x_n\}$ of the query q_2 such that every atom in the query q_2 resulted from variable/value substitutions occurs in query q_1. Note that if the query $q_1 \supseteq q_2$, then the query $q_1\theta$-subsumes q_2. However, if the query $q_1\theta$-subsumes q_2, it does not always mean that $q_1 \supseteq q_2$.

B. Swanson's ABC discovery models

Swanson and Smalheiser introduced two literature-based discovery models, *open discovery* and *closed discovery*, for hypothesis generation from two complementary but disjoint sets of literatures [18].

In the *open discovery* model, a starting topic a is used as a keyword search to retrieve a set of relevant literatures from a database. This set is called the *first literature set*. Subsequently, a set of terms $\mathbf{B} = \{b_1, b_2, ..., b_Q\}$ is extracted from the *first literature set*, and then each b_q is used as a keyword search to retrieve another set of literatures called the *second literature set of* b_q. Next, a set of terms $\mathbf{C}^{b_q} = \left\{c_1^{b_q}, c_2^{b_q}, ..., c_W^{b_q}\right\}$ are extracted from the *second literature set of* b_q for all terms b_q in $\mathbf{B}$. Finally, the starting term a is used together with each term $c_w^{b_q}$ in $\mathbf{C}^{b_q}$ as a keyword search for $q = \{1, ..., Q\}$ and $w = \{1, ..., W\}$. If there is no literature returned, a new association between a and $c_w^{b_q}$ through b_q is discovered, $a \rightarrow b_q \rightarrow c_w^{b_q}$.

In the *closed discovery* model, topics a and c are assumed to be unrelated. It means that there is no literature returned when a and c are used together as a keyword search. Two literature sets, $\mathbf{A}$-set and $\mathbf{C}$-set, are retrieved from a database. The $\mathbf{A}$-set uses a as a keyword search, and the $\mathbf{C}$-set uses c as a keyword search. A new association is discovered if there exists a term b that belongs to terms extracted from both the $\mathbf{A}$-set and the $\mathbf{C}$-set. The open discovery model is usually used for generating hypotheses and the closed discovery model is used for testing hypotheses.

The notion of two complementary but disjoint sets of literatures was introduced by Swanson [19]. Two separate sets of literatures are complementary, if, when combined, the relationship between them gives some important inferences and insights. In addition, two separate sets of literatures are disjoint, if there are no common articles between them and they do not cite or mention each other.

III. Data and methods

A. Data sets

Inspired by Swanson's discovery of the indirect connections between Raynaud's disease and fish oils, we retrieved sets of biomedical abstracts from NCBI PubMed using the following search strategy:

- Raynaud's disease and fish oils
  Search keyword: "raynaud disease", Publication date: from 1980 to 1985, Field: Text Word and Number of returned abstracts: 902
  Search keyword: "fish oils", Publication date: from 1980 to 1985, Field: Text Word and Number of returned abstracts: 187

In addition, based on our interest in Down syndrome, we used a second pair of topics, Down syndrome and cell polarity, and retrieved sets of biomedical abstracts using the following search strategy:

- Down syndrome and cell polarity
  Search keyword: "down syndrome", Publication date: 2008, Field: Text Word and Number of returned abstracts: 881
  Search keyword: "cell polarity", Publication date: 2008, Field: Text Word and Number of returned abstracts: 930

B. Methods

Discovery of relational knowledge from a pair of CBD sets of literatures can be achieved using a relational data mining approach where data are usually stored in a relational database.

We constructed a relational database in first-order logic. A table was created for each CBD set of literatures where a record corresponds to an abstract, and the PubMed ID of the abstract is used as a primary key. To find links between two tables, we extracted terms that are common to the two literature sets and used them as foreign keys. Accordingly, for a pair of CBD sets, $\mathbf{A}$ and $\mathbf{C}$, and a set of common terms $\mathbf{B}$, three tables are constructed: *Topic_A* table, *Topic_C* table and *Common_Term_B* table. The relational knowledge to be discovered is frequent patterns of common terms that link to biomedical abstracts containing co-sentences between two terms associated with a set of user-selected verbs. For two terms to be co-sentenced, one term must be in a subject part and another term must be in the object part of a sentence, where the subject and object parts are separated by a verb. Details follow.

1) Pre-processing:

- Given two unrelated topics, retrieve two sets of biomedical abstracts from NCBI PubMed, one set for each topic. Two topics are assumed to be unrelated if they have never been co-mentioned in the same article.
- Eliminate identical abstracts if any exists to ensure disjointedness.
- Perform sentence boundary detection and part-of-speech tagging on each abstract using Lingpipe (available at http://alias-i.com/lingpipe) and Genia Tagger [20], [21], respectively.
- Extract important terms (noun phrases) from each set of abstracts tagged by Genia Tagger [20], [21], and carried out stop word removal using a combined set of English and PubMed stop words.
- Compute weights of each extracted terms using a term weighting scheme such as document frequency, term frequency, or $TF \cdot IDF$.
- Rank terms in each set in descending order based on their numerical weights, and select a subset of terms using a threshold. Most single-word phrases with high ranks are too general: therefore, eliminate all single-word phrases.
- Extract common terms between the two sets of terms.
- Compute and ranked the common terms based on relative document frequency formulated by Equation 2. This is similar to Lindsay and Gordon's study [2].

$$RDF_t = \frac{DF_t^1 + DF_t^2}{N^1 + N^2}, \tag{2}$$

  where RDF_t is the relative frequency of a common term t, DF_t^1 is the document frequency of term t in the first literature set, DF_t^2 is the document frequency of term t in the second literature set, N^1 is the total number of documents in the first literature set, and N^2 is the total number of documents in the second literature set.
- Construct a relational database in first-order logic as shown in Fig. 1.

2) Co-sentence in first-order logic: We employed first-order logic for representing our relational database. Based on King *et al.* [16], we represented each abstract by a set of co-sentence predicates. In other words, a topic of interest is analogous to a compound, each retrieved abstract is analogous to atoms constituting the compound, and the co-sentence predicates extracted from the abstract are analogous to the properties of atoms that constituted the compound. Our motivation to use co-sentence, rather than co-mentioned, in the same abstract is based on Wren *et al.* [22] who found that the relatedness between two entities co-mentioned in the same sentence is higher than that co-mentioned in the same abstract.

In addition, since our goal is to uncover co-sentence information in an abstract in which at least one common term is present, we assume that when a common term occurs in an abstract, it relates to every sentence in that abstract in the same way that it relates to the abstract.

To construct a co-sentence predicate we divided a sentence into three parts based on verbs that occur in the sentence: subject phrase, verb phrase and object phrase. Let np be a noun phrase and v be a verb,

Sentence 1: $np_1\ v_1\ np_2$.
$\Longrightarrow$ Subj(np_1), Verb(v_1), Obj(np_2).

Then, if a subject phrase or object phrase contained more than one noun phrase as shown in Sentence 2, an enumeration of all different subject/verb/object combinations was generated.

Sentence 2: $np_1\ v_1\ np_2\ np_3$.
$\Longrightarrow$ Subj(np_1), Verb(v_1), Obj(np_2).
$\Longrightarrow$ Subj(np_1), Verb(v_1), Obj(np_3).

Similarly, if a sentence contained more than one verb as shown in Sentence 3, all possible different combinations of subjects, verbs, and objects were generated.

Sentence 3: $np_1\ v_1\ np_2\ v_2\ np_3$.
$\Longrightarrow$ Subj(np_1), Verb(v_1), Obj(np_2).
$\Longrightarrow$ Subj(np_1), Verb(v_1), Obj(np_3).
$\Longrightarrow$ Subj(np_1), Verb(v_2), Obj(np_3).
$\Longrightarrow$ Subj(np_2), Verb(v_2), Obj(np_3).

Finally, a relational table for a set of literatures was constructed using PubMed abstract IDs as primary keys and co-sentence predicates as attributes.

In addition, based on the notion of co-sentence, types of noun phrases, subject or object phrases, within the same sentence are not important. In other words, a co-sentence of Subj(np_1), Verb(v_1), and Obj(np_2) is the same as that of Obj(np_2), Verb(v_1), and Subj(np_1). As a result, noun phrases are alphabetically ordered in a co-sentence predicate. Moreover, verbs in the passive voice are converted into the base form.

Example: Given an abstract in **A**-set with PubMed ID *1234567* containing a a sentence $np_1\ v_1\ np_2$, if np_2 is alphabetically preceding np_1, a co-sentence predicate of this sentence can be expressed as follows.
Co-sentence predicate:
topicA_co_sentence('1234567', 'np_2', 'np_1', 'v_1').

3) Relational database in first-order logic: For a pair of complementary but disjoint sets of literatures $\mathbf{A} = \{a_1, a_2, ..., a_p, ..., a_P\}$ and $\mathbf{C} = \{c_1, c_2, ..., c_r, ..., c_R\}$, whose abstract contains one or more sentence patterns, and a set of common terms $\mathbf{B} = \{b_1, b_2, ..., b_q, ..., b_Q\}$, three tables in first-order logic can be created as follows:

Let $\mathbf{T} = \{t_1, t_2, ..., t_i, t_j, ..., t_N\}$ be a set of terms, $\mathbf{V} = \{v_1, v_2, ..., v_s, ..., v_S\}$ be a set of verbs extracted from all abstracts of both sets of literatures, and $\mathbf{B} \subseteq \mathbf{T}$

Topic A table:
topicA_co_sentence('a_1', 't_i', 't_j', 'v_s').
topicA_co_sentence('a_1', 't_i', 't_j', 'v_s').
topicA_co_sentence('a_1', 't_i', 't_j', 'v_s').
...
topicA_co_sentence('a_2', 't_i', 't_j', 'v_s').
topicA_co_sentence('a_2', 't_i', 't_j', 'v_s').
...
...
topicA_co_sentence('a_p', 't_i', 't_j', 'v_s').
topicA_co_sentence('a_p', 't_i', 't_j', 'v_s').
topicA_co_sentence('a_p', 't_i', 't_j', 'v_s').
...
...
topicA_co_sentence('a_P', 't_i', 't_j', 'v_s').
topicA_co_sentence('a_P', 't_i', 't_j', 'v_s').
...

Topic C table:
topicC_co_sentence('c_1', 't_i', 't_j', 'v_s').
topicC_co_sentence('c_1', 't_i', 't_j', 'v_s').
topicC_co_sentence('c_1', 't_i', 't_j', 'v_s').
...
topicC_co_sentence('c_2', 't_i', 't_j', 'v_s').
topicC_co_sentence('c_2', 't_i', 't_j', 'v_s').
...
...
topicC_co_sentence('c_r', 't_i', 't_j', 'v_s').
topicC_co_sentence('c_r', 't_i', 't_j', 'v_s').
topicC_co_sentence('c_r', 't_i', 't_j', 'v_s').
...
...
topicC_co_sentence('c_R', 't_i', 't_j', 'v_s').
topicC_co_sentence('c_R', 't_i', 't_j', 'v_s').
...

Common term B table:
common_term('b_1').
common_term('b_2').
common_term('b_3').
...
...
common_term('b_q').
...
...
common_term('b_Q').

Note that within the same co_sentence predicate, $t_i \neq t_j$, and across different co_sentence predicates, t_i or t_j in one predicate may or may not be the same as t_i or t_j in other predicates. Similarly, v_s in one predicate may or may not be the same as v_s in other predicates. In other words, each predicate is unique within a biomedical abstract. A graphical description of a relational database constructed from a pair of complementary but disjoint sets of literatures is shown in Fig. 1.

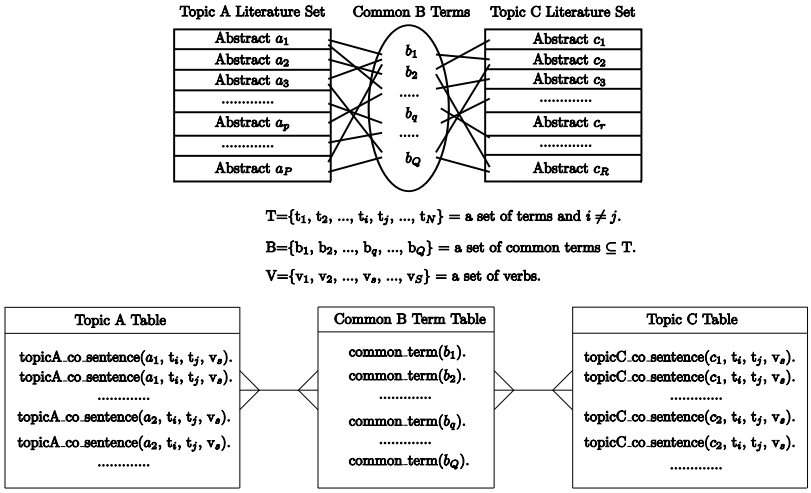

Fig. 1. A relational database in first-order logic built from two complementary but disjoint sets of literatures.

4) ACE data mining system: We used WARMR in ACE data mining system for searching for relational frequent patterns and rules. ACE is a data mining software tool that provides a common interface to a set of relational data mining algorithms. It was developed by the data mining research group at Katholieke Universiteit Leuven [23]. As of version 1.2.14, it consists of a number of useful algorithms such as TILDE (the relational version of the decision tree learner C4.5), WARMR (the relational version of APRIORI association rule mining algorithm) and ICL (the relational version of CN2 rule learner). ACE data mining system can be run in a Prolog system such as SWI-prolog (available at http://www.swi-prolog.org/) under the Linux operating system.

IV. Experiments and results

To search for frequent relational patterns using WARMR, a key pattern to be counted must first be specified. We used the *common_term* predicate as the key pattern. For the ACE data mining system, two input files are necessary: knowledgebase (.kb) file which contains predicates of relational database and settings (.s) file which contains user-defined key predicates and other necessary parameters such as minimum support and confidence. An input file, background knowledge (.bg) file used to provide knowledge about the domain, is optional.

To create a knowledgebase file, the database needs to be formatted in a way suitable for the data mining system used. Two possible formats are available in ACE data mining system: key format and models format. In the key format, every predicate must contain a key pattern as one of its argument. For the models format, the key pattern is specified by the *begin* and *end* predicates. We used the key format for constructing the knowledgebase.

We constructed a knowledgebase file using the top 15 common terms listed in Table 1 and 25 user-selected action verbs listed in Table 2 as constraints. These user-selected common terms and action verbs serve as two functions: (i) they are used for extracting only knowledge of interest and (ii) they are used to reduce the problem space.

An example of knowledgebase file for Down syndrome and cell polarity is shown below.

common_term('epithelial_cell').
common_term('plasma_membrane').
common_term('gene_expression').
...
...
common_term('essential_role').
cell_polarity_co_sentence('18256323', 'key_role',
'itch_pathology', 'presence', 'correlate').
...
...
down_syndrome_co_sentence('17517397', 'gene_expression',
'cytochrome_p450_aromatase_activity',
'mullerian_inhibiting_substance', 'inhibit').
...
...

The first argument in the *cell_polarity_co_sentence* and *down_syndrome_co_sentence* predicates corresponds to the PubMed ID of the abstract that contains the common term in the second argument, the third argument to the first noun phrase, the fourth argument to second noun phrase and finally the fifth argument to a verb that separates these noun phrases in the same sentence.

The experimental results are presented in the rest of this section.

TABLE I

LIST OF THE TOP 15 MOST COMMON TERMS.

No.	Raynaud's disease and fish oils	Down syndrome and cell polarity
1	rheumatoid_arthristis	epithelial_cell
2	platelet_aggregation	plasma_membrane
3	blood_pressure	gene_expression
4	vascular_disease	molecular_mechanism
5	platelet_function	actin_cytoskeleton
6	blood_viscosity	cell_division
7	peripheral_vascular_disease	cell_migration
8	beneficial_effect	mental_retardation
9	plasma_viscosity	cell_proliferation
10	platelet_count	transcription_factor
11	systolic_blood_pressure	key_role
12	connective_tissue	key_regulator
13	essential_hypertension	crucial_role
14	autoimmune_disease	early_stage
15	epidemiological_study	essential_role

TABLE II

LIST OF 25 USER-SELECTED ACTION VERBS.

No.	Action verbs
1	activate
2	affect
3	associate
4	cause
5	control
6	correlate
7	decrease
8	enhance
9	exhibit
10	improve
11	increase
12	induce
13	inhibit
14	influence
15	interact
16	involve
17	lead
18	prevent
19	promote
20	raise
21	reduce
22	regulate
23	stimulate
24	suppress
25	treat

A. Raynaud's disease and fish oils

A subset of relational patterns generated by WARMR on Raynaud's disease and fish oils data set with the minimum support set to 0.2 is described below.

- *freq(2, 3, [common_term(A), fish_oils_co_sentence(B, A, blunted_circulatory_response, membrane_phospholipid, associate)], 0.266666666666667).*
  The format of this frequent pattern output is described as follows. The *freq* means "frequent query" or "frequent pattern". The first argument corresponds to the tree depth level. In this case, the tree depth level is 2. The second argument is the frequent pattern number generated by WARMR. Accordingly, this frequent pattern is the third pattern generated in level 2 by WARMR. The third argument is the frequent pattern which consists of *common_term/1* and *fish_oils_co_sentence/5* predicates. Finally, the fourth argument is the support of this frequent pattern. Since the minimum support is set to 0.2 and the support of this pattern is 0.267, the pattern is considered frequent.
  In addition, this output result can be interested as follows: 26.67% of the common terms link to the fish oils abstracts that contain the co-sentence pattern describing "blunted circulatory response either associates with or is associated with membrane phospholipid."
- *freq(2, 5, [common_term(A), fish_oils_co_sentence(B, A, blunted_circulatory_response, omega_3, associate)], 0.266666666666667).*
  26.67% of the common terms link to the fish oils abstracts that contain the co-sentence pattern describing "blunted circulatory response either associates with or is associated with omega-3."
- *freq(2, 28, [common_term(A), raynaud_disease_co_sentence(B, A, aggregation_platelet_retention, collagen, induce)], 0.2).*
  20% of the common terms link to the Raynaud's disease abstracts that contain the co-sentence pattern describing "aggregation platelet retention either induces or is induced by collagen."
- *freq(2, 26, [common_term(A), fish_oils_co_sentence(B, A, adp, platelet_aggregation,induce)], 0.2).*
  20% of the common terms link to the fish oils abstracts that contain the co-sentence pattern describing "adp induces or is induced by platelet aggregation."

B. Down syndrome and cell polarity

A subset of relational patterns generated by the WARMR on Down syndrome and cell polarity data set with the minimum support set to 0.2 is shown below.

- *freq(2, 16, [common_term(A), cell_polarity_co_sentence(B, A, increase_cell_migration, intercellular_interaction, correlate)], 0.2).*
  20% of the common terms link to the cell polarity abstracts that contain the co-sentence pattern describing "increase cell migration either correlates with or is correlated with intercellular interaction."
- *freq(2, 33, [common_term(A), cell_polarity_co_sentence(B, A, apicolateral_tight_junction, cell_proliferation, affect)], 0.2).*
  20% of the common terms link to the cell polarity abstracts that contain the co-sentence pattern describing "apicolateral tight junction either affects or is affected by cell proliferation."
- *freq(2, 293, [common_term(A), down_syndrome_co_sentence(B, A, dyrk1a_dose_sensitive_reduction, early_endodermal_mesodermal_differentiation, cause)], 0.2).*
  20% of the common terms link to the Down syndrome abstracts that contain the co-sentence pattern describing "dyrk1a dose sensitive reduction either causes or is caused by early endodermal mesodermal differentiation."
- *freq(2, 306, [common_term(A), down_syndrome_co_sentence(B, A, premature_expression, undifferentiated_trisomy_es_cell, cause)], 0.2).*
  20% of the common terms link to the Down syndrome abstracts that contain the co-sentence pattern describing "premature expression either causes or is caused by undifferentiated trisomy es cell."

In addition to relational frequent patterns, WARMR can be used to generate relational association rules. We modified a setting file by allowing verbs to be a variable with input/output mode, and set both minimum support and confidence to 0.1. We then reran WARMR on Raynaud's disease and fish oils data set. A subset of rules generated by WARMR and their interpretations is shown as follows.

- *rules(([fish_oils_co_sentence(B, A, blunted_circulatory_response, omega_3, C)] :- [common_term(A)]), bodyfreq(1.0), sup(0.266666666666667), conf(0.266666666666667), lift(none)).*
  The generated rule consists of two parts: head and body, which is separated by a ":-" symbol. The head part corresponds to the *fish_oils_co_sentence/5* predicate and the body part corresponds to the *common_term/1* predicate. It can be interpreted as follows. For all common terms A it holds that if a fish oil abstract that contains a common term A, it will also contain a co-sentence between blunted circulatory response and omega 3 in which a verb C is a separation between them.
- *rules(([raynaud_disease_co_sentence(B, A, aggregation_platelet_retention, collagen, C)] :- [common_term(A)]), bodyfreq(1.0), sup(0.2), conf(0.2), lift(none)).*
  For all common terms A it holds that if a Raynaud's disease abstract that contain the common term A contains a common term A will also contain a co-sentence between aggregation platelet retention and collagen in which a verb C is a separation between them.
- *rules(([fish_oils_co_sentence(D, A, blunted_circulatory_response, omega_3,E)] :- [common_term(A), raynaud_disease_co_sentence(B, A, treatment, vessel_patency_rate, C)]), bodyfreq(0.2), sup(0.133333333333333), conf(0.666666666666667), lift(none)).*
  For all common terms A, it holds that if a Raynaud's disease abstract contains a co-sentence between treatment and vessel patency rate, that a fish oils abstract will also contains a co-sentence between blunted circulatory response and omega 3.
- *rules(([fish_oils_co_sentence(D, A, omega_3, reactive_platelet, E)] :-[common_term(A), raynaud_disease_co_sentence(B, A, treatment, vessel_patency_rate, C)]), bodyfreq(0.2), sup(0.133333333333333), conf(0.666666666666667), lift(none)).*
  For all common terms A, it holds that if a Raynaud's disease abstract that contains the common term A contains a co-sentence between treatment and vessel patency rate, that a fish oils abstract that contains the common term A will also contain a co-sentence between omega 3 and reactivate platelet.

V. Conclusions

In the paper we report an application of the Inductive Logic Programming technique, the WARMR algorithm, to the literature-based discovery domain. Our application method extends the closed discovery model of Swanson to the discovery of potentially useful knowledge in the forms of relational frequent patterns from two complementary but disjoint (CBD) sets of literatures after the indirect connections are found. The experimental results show that our approach can be used to provide information additional to standard literature-based discovery, and that it can be used as an exploratory method in studies. In addition, the results have shown that Inductive Logic Programming is well suited for text mining because it provides flexibility in defining the textual patterns to be discovered and has good expressive power of the outputs.

Several issues need to be considered. For two different topics, the number of returned literatures may be unbalanced for a number of reasons. With unbalanced data sets, patterns in the small literature set may never be included in the output because they do not pass the threshold. In addition, the WARMR setting file is critical. It allows users to provide

constraints to reduce the search space, and to specify queries to be refined.

In the future, our approach can be improved in several ways. First, biomedical concepts can be used instead of words or phrases. Using concepts not only gives semantics to the discovered knowledge, but also helps reduce the search space. Secondly, other types of predicates may be added in addition to the predicate, verbs may be used as a predicate name instead of one of the arguments. Finally, background knowledge can be added to provide additional information or used as constraint of the search space such as information about variations of gene symbols and names.

REFERENCES

[1] D. R. Swanson, "Fish oil, raynaud's syndrome, and undiscovered public knowledge." *Perspectives in Biology and Medicine*, vol. 30, no. 1, pp. 7–18, 1986.

[2] R. K. Lindsay and M. D. Gordon, "Literature-based discovery by lexical statistics," *Journal of American Society for Information Science (JASIST)*, vol. 50, no. 7, pp. 574–587, 1999.

[3] W. Pratt and M. Yetisgen-Yildiz, "Litlinker: capturing connections across the biomedical literature," in *K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture*. New York, NY, USA: ACM, 2003, pp. 105–112.

[4] P. Srinivasan, "Text mining: generating hypotheses from medline," *Journal of the American Society for Information Science and Technology*, vol. 55, no. 5, pp. 396–413, 2004.

[5] W. Jin and R. K. Srihari, "Knowledge discovery across documents through concept chain queries," in *ICDMW '06: Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 448–452.

[6] D. Hristovski, B. Peterlin, J. A. Mitchell, and S. M. Humphrey, "Improving literature based discovery support by genetic knowledge integration." *Stud Health Technol Inform.*, vol. 95, pp. 68–73, 2003.

[7] J. Stegmann and G. Grohmann, "Hypothesis generation guided by co-word clustering," *Scientometrics*, vol. 56, no. 1, pp. 111–135, 2003.

[8] J. D. Wren, "Extending the mutual information measure to rank inferred literature relationships," *BMC Bioinformatics*, vol. 5, no. 145, 2004.

[9] S. Muggleton and L. De Raedt, "Inductive logic programming: Theory and methods," *Journal of Logic Programming*, vol. 19/20, pp. 629–679, 1994.

[10] S. Muggleton, "Inverse entailment and progol," *New Generation Computing, Special issue on Inductive Logic Programming*, vol. 13, no. 3-4, pp. 245–286, 1995.

[11] J. R. Quinlan, "Learning logical definitions from relations," *Mach. Learn.*, vol. 5, no. 3, pp. 239–266, 1990.

[12] L. Dehaspe and H. Toivonen, "Discovery of frequent datalog patterns," *Data Min. Knowl. Discov.*, vol. 3, no. 1, pp. 7–36, 1999.

[13] L. Dehaspe and H. T. T. Toivonen, "Discovery of relational association rules," in *Relational data mining*. Springer-Verlag, 2001, pp. 189–212.

[14] M. Liakata and S. Pulman, "Learning theories from text," in *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2004, p. 183.

[15] J. Blaťák, "First-order frequent patterns in text mining," *Artificial intelligence, 2005. epia 2005. portuguese conference on*, pp. 344–350, Dec. 2005.

[16] R. D. King, A. Srinivasan, and L. Dehaspe, "Warmr: a data mining tool for chemical data," *Journal of Computer-Aided Molecular Design*, vol. 15, no. 2, pp. 173–181, 2001.

[17] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases." in *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 1993, pp. 207–216.

[18] D. R. Swanson and N. R. Smalheiser, "An interactive system for finding complementary literatures: a stimulus to scientific discovery," *Artificial Intelligence*, vol. 91, no. 2, pp. 183–203, 1997.

[19] D. R. Swanson, "Complementary structures in disjoint science literatures," in *SIGIR '91: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 1991, pp. 280–289.

[20] Y. Tsuruoka, Y. Tateishi, J.-D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii, "Developing a robust part-of-speech tagger for biomedical text," in *Advances in Informatics -10th Panhellenic Conference on Informatics*, 2005, pp. 382–392.

[21] Y. Tsuruoka and J. Tsujii, "Bidirectional inference with the easiest-first strategy for tagging sequence data," in *Proceedings of HLT/EMNLP*, 2005, pp. 467–474.

[22] J. D. Wren, R. Bekeredjian, J. A. Stewart, R. V. Shohet, and H. R. Garner, "Knowledge discovery by automated identification and ranking of implicit relationships," *Bioinformatics*, vol. 20, no. 3, pp. 389–398, 2004, bibtex is from ACM.

[23] H. Blockeel, L. Dehaspe, B. Demoen, G. Janssens, J. Ramon, and H. Vandecasteele, "Improving the efficiency of inductive logic programming through the use of query packs," *Journal of Artificial Intelligence Research*, vol. 16, pp. 135–166, 2002.