2013 12th International Conference on Machine Learning and Applications

Mining Biomedical Ontologies and Data Using RDF Hypergraphs

Haishan Liu, Dejing Dou
University of Oregon
Eugene, OR, 97403, USA
{ahoyleo, dou}@cs.uoregon.edu

Ruoming Jin
Kent State University
Kent, OH, 44242, USA
jin@cs.kent.edu

Paea LePendu, Nigam Shah
Stanford University
Stanford, CA, 94305, USA
{plependu, nigam}@stanford.edu

***Abstract*—As researchers analyze huge amounts of data that are annotated by large biomedical ontologies, one of the major challenges for data mining and machine learning is to leverage both ontologies and data together in a systematic and scalable way. In this paper, we address two interesting and related problems for mining biomedical ontologies and data: i) how to discover *semantic associations* with the help of formal ontologies; ii) how to identify potential errors in the ontologies with the help of data. By representing both ontologies and data using *RDF hypergraphs*, and subsequently transforming the hypergraphs to corresponding bipartite forms, we provide a generalized data mining method that scales beyond what existing ontology-based approaches can provide. We show the proposed method is indeed capable of capturing semantic associations while seamlessly incorporate domain knowledge in ontologies by performing evaluations on real-world electronic health dataset and NCBO ontologies. We also show that our data mining methods can discover and suggest corrections for misinformation in biomedical ontologies.**

I. Introduction

Researchers around the world are linking more and more data to ontologies that are formal specification of concepts and relationships in various domains. Ontologies have been extensively harnessed in an array of research fields, particularly in biomedicine where knowledge evolves rapidly and has promoted the creation and use of ontologies to advance scientific progress. For example, over 300 ontologies have been loaded into the National Center of Biomedical Ontology (NCBO) BioPortal library at Stanford [1], specifying more than 5.6 million terms in the biomedical domain.

There are two major challenges facing researchers when it comes to mining large sets of biomedical ontologies and data. The first is to leverage both ontologies and data in a systematic and scalable way. The second is to deal with errors in both ontologies and data since neither of them is perfect in reality. With the increasing amount of ontology-annotated data, a new research direction emerges, which we call *semantic data mining*, focusing on drawing insights from both domain knowledge and data in a systematic way. It aims at bringing domain knowledge seamlessly into the data mining process, and helps improve the quality of patterns discovered in a noisy environment. It also aims at improving ontologies by utilizing empirical substantiation from data to either bolster a priori ontological assertions, or detect potential errors therein.

Semantic data mining leverages links between entities defined by ontologies—via annotations—to the mining algorithms explicitly in a unified model. This requires traversing links across the ontologies to infer implicit inter-connections among the data. Graph techniques fit this research nicely because both domain knowledge and semantically rich datasets can be represented as graphs. For example, OWL [2] is the standard ontology language built on RDF. Inheriting the graph nature of RDF, any collection of OWL ontologies or RDF data is an RDF Graph [3]. In fact, many semantically rich datasets of interest today, such as DBpedia, are best described as a linked collection, or a graph, of interrelated objects [4].

Hence, our semantic data mining approach is inspired by a combination of graph representation [3], [5], [6], and mining techniques [7], [8]. This paper extends our previous work that implements a hypergraph-based approach to learn associations from interlinked data (without ontologies) [9]. We adopt the RDF hypergraph representation proposed by Hayes et al. [3] to connect data and ontologies. Under such representation, properties in RDF triples can be represented as first class objects among all interrelated objects, enabling us to combine both ontologies and data in a consistent manner.

The graph-centric root of our approach makes it possible to leverage decades of work on graph mining. In the present study, we use random walk with restart to derive similarity between concepts and tackle the problem of discovering associations on a large scale. Traditional association mining relies on co-frequencies of items (concepts) within transactions [10]. We look one step further to find indirectly associated items. This extension has far reaching applications in biomedicine. For instance, consider a simple scenario illustrated by Swanson [11] years ago while studying Raynauld's syndrome. He noticed that Raynauld's syndrome (Z) had been linked with certain changes of blood in human body (Y) in the literature; and, separately, the consumption of dietary fish oil (X) was also linked to similar blood changes. But fish oil and Raynauld's syndrome were never linked directly in any previous publications. Swanson reasoned (correctly) that fish oil could potentially be used to treat Raynauld's syndrome, i.e., $X \sim Z$ from $X \sim Y$ and $Y \sim Z$. We term such indirect connection between X and Z the *semantic association*.

Our work makes the following main contributions: First, we employ a RDF hypergraph representation to capture information from both ontologies and data. Next, we transform the hypergraph and weighted hyperedges into a bipartite form for efficient processing. Then, we use the random walks with restart over the bipartite graph to generate semantic associations. Finally, the discovered semantic associations can be used to detect potential errors in biomedical ontologies.

978-0-7695-5144-9/13 $31.00 © 2013 IEEE
DOI 10.1109/ICMLA.2013.31

II. Related Work

A. Ontologies and Data Mining

Using formal ontologies to annotate data has become increasingly popular in biomedical domains. For instance, in genetics, researchers curate literature to generate ontology-annotated data for different species of model organisms by linking specific proteins to various classes in the Gene Ontology (GO [12]). At Stanford, the National Center for Biomedical Ontology (NCBO) annotates large volumes of biomedical text for search and mining [13] and has been used, for instance, to profile disease research [14].

In general domains, Staab and Hotho [15] were one of the earliest to utilize the idea of mapping terms in text to classes in an ontology and they essentially use the ontology to aggregate data and thus reduce feature dimensionality during clustering. Adryan et al. [16] enable cluster visualization for gene expression data by navigating various levels of the Gene Ontology hierarchy. Wen et al. [17] take into consideration the ontology hierarchy to offset biases toward overly-general terms in text mining.

B. Graphs in Mining RDF and Ontologies

RDF data and OWL ontologies can be represented as graphs for data mining. Lin et al. [18] treat the RDF triple store as a data source during mining and develop Relational Bayesian Classifiers (RBCs) that aggregate SPARQL queries. Bicer et al. [19] define kernel machines over RDF data where features are constructed by ILP-based dynamic propositionalization. In each case, RDF is merely a data model, paying little attention to the use of domain-specific knowledge in related ontologies. Recently, a promising graph-based approach to represent and mine ontologies and data together is Heterogeneous Information Networks (HIN) developed by Sun et al. [20], [21]. HIN leverages semantics of various types of nodes and links in a network for the graph and network mining tasks. One advantage of the proposed method to HIN is the convenience of RDF bipartite graphs to which existing ontologies and their annotated data can be easily transformed.

III. Method

A. Graph Representation for Biomedical Ontologies

The Web Ontology Language (OWL [2]) is the W3C's standard for representing Semantic Web ontologies and has been adopted by most biomedical ontology development efforts. OWL ontologies can be used along with RDF data because OWL uses the RDF syntax. RDF's abstract triple syntax has a graph nature. The RDF graph is defined as a set of triples and can be viewed as a *directed labeled graph* (DG). One disadvantage of DG is that it makes an artificial distinction between resources and properties, which leads to incongruous representations (e.g., properties are usually represented as arcs, but in meta-statement about the properties themselves, the properties have to be represented as nodes).

To overcome the inconsistency, Hayes et al. [3] proposed to model RDF as a *hypergraph*. A hypergraph [22] is a generalization of a simple graph where edges, called hyperedges, can connect more than two vertices. If each edge in a hypergraph covers the same number of nodes, it is called r-uniform hypergraph, r being the number of nodes on each edge. Any RDF graph can be represented by a simple ordered 3-uniform hypergraph, in which an RDF triple corresponds to a hyperedge, with incident nodes being the subject, predicate and object from the triple. In this way, both ontology and data statements are integrated in a consistent graph representation.

Formally, a hypergraph $HG = (V, E)$, is a pair in which V is the vertex set and E is the hyperedge set where each $e \in E$ is a subset of V. A weighted hypergraph is a hypergraph that has a positive number $w(e)$ associated with each hyperedge; called the weight of hyperedge. A weighted hypergraph can be denoted by $G = (V, E, W)$. Furthermore, A hypergraph $HG = (V, E)$ can be transformed to a *bipartite graph* BG as follows: let the node sets V and E be the two parts of BG. Then (v_1, e_1) is connected with an edge if and only if vertex v_1 is contained in the hyperedge e_1 of HG. In other words, the incidence matrix of HG can be viewed as the node adjacency (biadjacency) matrix of the bipartite graph. BG turns HG to a simple form, where algorithms designed on simple graphs can be readily applied. Therefore, we use RDF bipartite graphs as the combined representation of domain knowledge and data.

B. Graph representation for Ontology-Annotated Data

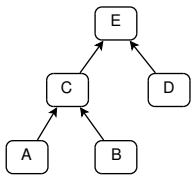
There already exist methods for transforming data, such as those in relation databases, into RDF [23]. An ontology-annotation, as we see it, is a binary value representing whether some ontological concept (or class) is associated with some entity. Often, this means that some concept appears in some document and thus the ontology serves to index the document with related concepts [13]. Thus, we can think of ontology-annotations as a table, with each row representing an entity (e.g., a document), and each column is a class from some ontology. Cells having a "1" denotes that the document *mentions* the term defined by the class. RDF can be seen as a sparse matrix representation of this data. This idea can be easily extended to nominal-valued tables as well, or with other relationships besides *mentions* as we illustrate when discussing unified bipartite graphs in the next section.

C. Unified Graph Representation for Biomedical Ontologies and Data

Given the ontology-annotated data, a unified graph incorporating information from both the ontology and data can be created, as demonstrated in the following example.

Figure 1 (A) shows a simple ontology with only subsumption relationships defined for five entities (A–E) representing, for example, concepts in the biomedical domain. Figure 1 (B) is a binary-valued RDB table in the same domain A–E being column headers (features). We use the same concept labels in the ontology and the RDB table because we assume the mapping between the ontology nodes and the table features are pre-assigned manually or established by automatic annotation. Figure 2 (B) shows the RDF statements derived from both the ontology and the RDB table. Figure 2 (A) demonstrates the unified RDF bipartite graph.

Formally, the RDF bipartite graph as a unified representation for both data and ontologies is defined as $G = \langle V_v \cup V_s, E \rangle$, where V_v denotes *value nodes* corresponding to components of RDF statements (i.e., subject, predicate,

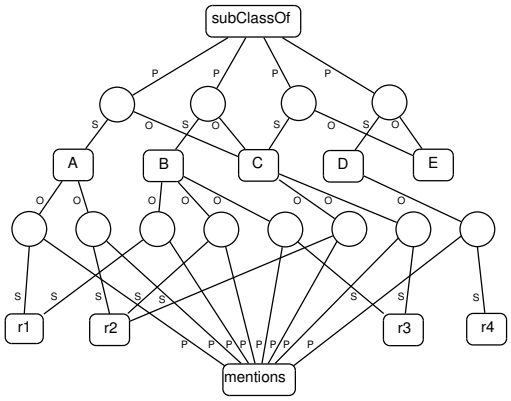


	A	B	C	D	E
A :	0	0	1	0	0
B :	0	0	1	0	0
C :	0	0	0	0	1
D :	0	0	0	0	1
E :	0	0	0	0	0

	A	B	C	D	E
r_1 :	1	1	0	0	0
r_2 :	1	1	1	0	0
r_3 :	0	1	1	0	0
r_3 :	0	0	0	1	0

(A) (B) (C)

Fig. 1. Five concepts (*A*–*E*) are represented visually as a hierarchy (A) and also as a hypergraph using the binary feature matrix (B), where a "1" denotes *rdfs:subClassOf*, which is similar to the ontology-annotated data (C), where "1" denotes *mentions*.



	s	*p*	*o*
s_1:	<A>	<subClassOf>	<C>
s_2:	<B>	<subClassOf>	<C>
s_3:	<C>	<subClassOf>	<E>
s_4:	<D>	<subClassOf>	<E>
s_5:	<r1>	<mentions>	<A>
s_6:	<r1>	<mentions>	<B>
s_7:	<r2>	<mentions>	<A>
s_8:	<r2>	<mentions>	<B>
s_9:	<r2>	<mentions>	<C>
s_{10}:	<r3>	<mentions>	<B>
s_{11}:	<r3>	<mentions>	<C>
s_{12}:	<r4>	<mentions>	<D>

(A) (B)

Fig. 2. The RDF bipartite graph representation (A) easily combines both the ontology-annotated data with the ontological relationships (B) based on the information described in Figure 1.

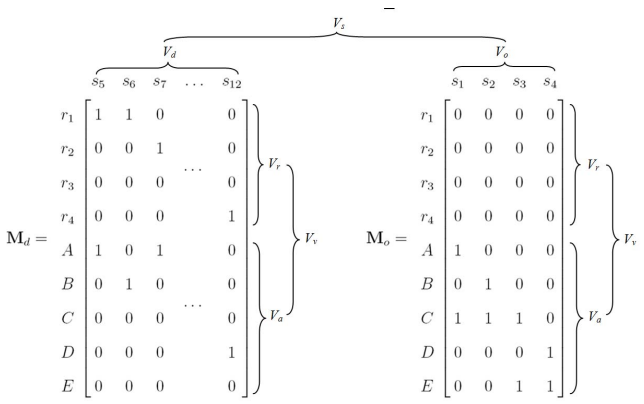


Fig. 3. A detailed anatomy of the biadjacency matrix for the RDF bipartite graph in Figure 2 (A).

or object), and V_s denotes *statement nodes* corresponding to RDF statements. More specifically, statement nodes can be further divided according to whether they are from data or ontology, i.e., $V_s = V_d \cup V_o$; Value nodes can be divided according to whether they represent rows (records) or columns (attributes) in data, i.e., $V_d = V_r \cup V_a$. The graph G can be represented in a biadjacency matrix $\mathbf{M}$, where $\mathbf{M}(i, j)$ is non-zero if there is an edge between $\langle V_{v_i}, V_{s_j} \rangle$. For an unweighted graph, the value can be 0/1, and for a weighted graph, any non-negative value. Weights assigned to different paths in the graph are used to distinguish various semantic types or relationships (properties) from the ontology and data, such as class subsumption, "part_of", and other general or domain-specific relationships.

For example, Figure 3 shows the biadjacency matrices $\mathbf{M}_d$ and $\mathbf{M}_o$ for the RDF bipartite graph shown in Figure 2(A). $\mathbf{M}_d$ and $\mathbf{M}_o$ correspond to the data and ontology part of the RDF bipartite graph respectively. We can see that rows of $\mathbf{M}_d$ and $\mathbf{M}_o$ correspond to *value nodes*, (V_v), which can be further divided into row nodes V_r and attribute nodes V_a. On the other hand, columns of $\mathbf{M}_d$ are nodes that correspond to RDF statements about data (V_d), and columns of $\mathbf{M}_o$ correspond to the ontology (V_o). The union of V_d and V_o constitutes the whole set of statement nodes V_s (all circle nodes in Figure 2(A), i.e., s_1–s_{12} in Figure 2(B)).

From the above example we notice that the biadjacency matrix $\mathbf{M}$ can be split into vertical stripes by statement nodes V_s. To obtain the biadjacency matrix $\mathbf{M}$ of the unified RDF bipartite graph in Figure 3(A), we can simply concatenate $\mathbf{M}_d$ and $\mathbf{M}_o$ horizontally: $\mathbf{M} = [\mathbf{M}_d\ \mathbf{M}_o]$. This gives us a way to construct the matrix modularly from its independent components. In general, if there are k different semantic relationships in ontologies, $\mathbf{M}_o$ can be divided into more vertical stripes $\{\mathbf{M}_{o_i}, i = 1 \ldots k\}$, where $\mathbf{M}_{o_i}$ may represent, for example, the "part_of" lattice. Each $\mathbf{M}_{o_i}$ can be distinguished from others by different weights assigned to it. In short, $\mathbf{M}$ is the horizontal concatenation of all weighted vertical stripes as

shown in the first equality of Equation 1. The internal block structure of the concatenated biadjacency matrix $\mathbf{M}$ is shown in the second equality of Equation 1.

$$\mathbf{M} = \begin{bmatrix} w_d\mathbf{M}_d & w_{o_1}\mathbf{M}_{o_1} & w_{o_2}\mathbf{M}_{o_2} & \dots \end{bmatrix} = \begin{array}{c} \\ r \\ a \end{array} \begin{array}{c} \begin{array}{cccc} ds & os_1 & os_2 & \dots \end{array} \\ \left[\begin{array}{c|c|c|c} \mathbf{M}_{dr} & \mathbf{0} & \mathbf{0} & \dots \\ \hline \mathbf{M}_{da} & \mathbf{O}_1 & \mathbf{O}_2 & \dots \end{array}\right] \end{array} \tag{1}$$

D. Mining Unified RDF Bipartite Graphs

In this section, we present our method for discovering semantic associations based on the unified RDF bipartite graph of both the ontology and data. Similar to the relevance score [24], we believe that two items have a strong semantic association if they are related to many similar objects. We denote the similarity score between entities e_1 and e_2 by $s(e_1, e_2)$, where $s(e_1, e_2) \in [0, 1]$ and $s(e_1, e_2) = 1$ if $e_1 = e_2$. Now the problem of ranking semantic associations in the unified graph can be described as follows.

Given an attribute node a in the unified graph $G = G_d \cup G_o$ and $a \in G_d \cap G_o$ we want to compute a similarity score $s(a, b)$ for all nodes $b(\neq a) \in G_d \cap G_o$. We choose to apply random walks with restart (RWR) from the given node a, and use the steady-state probability of each other node at convergence as the similarity measure [25]. In other words, the similarity score of node b is defined as the probability of visiting b via a random walk which starts from a and goes back to a with a probability c. The RWR score on the unified RDF bipartite graph is efficient to compute since the graph is skewed (generally there are more statement nodes than value nodes on large graphs). In the following, we describe how to algorithmically calculate the RWR-based similarity on the RDF bipartite graph.

Given a $(n \times m)$ biadjacency matrix $\mathbf{M}$ in Equation 1 for G, we can construct the adjacency matrix $\mathbf{A}$ as follows: $\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{M} \\ \mathbf{M}^T & \mathbf{0} \end{bmatrix}$. The probability of a random walker taking a particular edge $\langle a, b \rangle$ from a node a while traversing the graph is proportional to the edge weight over the total weight of all outgoing edges from a, i.e., $\mathbf{P}(a, b) = \mathbf{A}(a, b)/\Sigma_{i=1}^{m+n}\mathbf{A}(a, i)$. Therefore, the Markov transition matrix $\mathbf{P}$ of G is constructed as: $\mathbf{P} = normc(\mathbf{A})$, where $normc(\mathbf{A})$ normalizes $\mathbf{A}$ such that every column sum up to 1.

Given the transition matrix $\mathbf{P}$, we can calculate the similarity scores using the following steps. First, we transform the input attribute node a into a $(k+n) \times 1$ query vector $\mathbf{q}_a$ with 1 in the a-th row and 0 otherwise. Second, we to compute a $(k + n) \times 1$ steady-state probability vector $\mathbf{u}_a$ over G. Last we extract only the steady-state probabilities of row nodes in $\mathbf{M}$ (corresponding to value nodes in the RDF bipartite graph) as the output similarity score vector. Notice that $\mathbf{u}_a$ can be computed by an iterated method from the following equation.

Let c be the probability of restarting random-walk from the node a. Then the steady-state probability vector $\mathbf{u}_a$ satisfies

$$\mathbf{u}_a = (1 - c)\mathbf{P}_A\mathbf{u}_a + c\mathbf{q}_a \ . \tag{2}$$

The iterative update of $\mathbf{u}_a$ can be performed as shown in Algorithm 1. The while loop is modified from Equation 2 to avoid materializing $\mathbf{A}$ and $\mathbf{P}$ for scalability.

Algorithm 1 Calculate Semantic Association

Input: query attribute a, bipartite matrix M, restarting probability c, tolerant threshold ϵ
Output: similarity vector $\mathbf{u}_a(1 : k)$
  $\mathbf{q}_a \Leftarrow \mathbf{0}$
  $\mathbf{q}_a(a) = 1$ (set a-th element of $\mathbf{q}_a$ to 1)
  while $|\Delta\mathbf{u}_a| > \epsilon$ **do**

$$\mathbf{u}_a = (1 - c) \begin{bmatrix} normc(\mathbf{M})\mathbf{u}_a(k + 1 : k + n); \\ normc(\mathbf{M}^T)\mathbf{u}_a(1 : k) \end{bmatrix} + c\mathbf{q}_a$$

  end while
  return $\mathbf{u}_a(1 : k)$

IV. Experiment

In this section, we evaluate the proposed method for discovering semantic associations and detecting errors in biomedical ontologies on an *electronic health records* dataset. We first describe the dataset and then present the evaluation results.

A. Dataset

In this evaluation, we analyze the electronic health records of real patients. The clinical note data are from Stanford Hospital's Clinical Data Warehouse (STRIDE). These records archive over 17-years worth of data comprising of 1.6 million patients, 15 million encounters, 25 million coded ICD9 diagnoses, and a combination of pathology, radiology, and transcription reports totaling over 9 million clinical notes (i.e., unstructured text). We obtained the set of drugs and diseases for each patient's clinical note by using a new tool, the *Annotator Workflow*, developed at the National Center for Biomedical Ontology (NCBO), which annotates clinical text from electronic health record systems and extracts disease and drug mentions from the electronic health records. For this study, we specifically configured the workflow to use a subset of the NCBO ontology library that are most relevant to clinical domains. The resulting set of ontologies contains 1 million subsumption ("is_a") statements.

From this set of 1.6 million patients with annotated records, we vectorize texts and turned them into a bag-of-word representation, from which an RDF bipartite graph is constructed, including 148 million RDF statements for the data.

To highlight the capability of our method for incorporating multiple types of relationships, we also explore the "may_treat" relationship between drugs and diseases defined in the NDFRT ontology, for example, Thiabendazole "may_treat" Larva Migrans. In the experiment, we extracted 43,780 "may_treat" statements from the ontology. Since we are interested in learning the interaction between drugs and diseases, "may_treat" is naturally a better indicator relationship to include while mining semantic associations than the subsumption relationship. Our results below illustrate this point.

To summarize, in terms of the size of the electronic health dataset, the derived unified bipartite graph contains 148,690,056 statements from data, 1,048,604 statements from the *is_a* ontology subgraph, and 43,780 statements from the *may_treat* ontology subgraph.

B. Results

1) Discovering Semantic Associations: We first apply our method to study the drug-drug association by combining the subsumption hierarchy in the ontology graph with the data graph. Table I demonstrates semantic associations for the term *rofecoxib* given different configurations of the unified graphs. Rofecoxib is the active ingredient of the drug *Vioxx*, which was recalled in 2005 because it was causing an increased risk of heart attacks. Vioxx is one of several COX-2 inhibitor nonsteroidal anti-inflammatory drugs.

Table I shows that, with only the ontology graph, the algorithm successfully picks up almost all other active ingredients of the COX-2 inhibitor class of drugs (valdecoxib, celecoxib, etc.). Drugs of the COX inhibitor (the parent of COX-2) class also appear in the top results (meloxicam, nabumetone, etc.). These are indeed semantic associations since the top items are related to rofecoxib indirectly through parent classes. It is worth noticing that, not only is rofecoxib a subclass of COX-2 inhibitor drugs, it is also a descendent of a much broader parent called "Drug Products by Generic Ingredient Combinations," whose subclasses are organized by descendants' initial alphabets. In other words, rofecoxib is also a direct child of a class that contains all drug ingredients starting with the letter R. The fact that our algorithm selects the neighboring class of rofecoxib in the COX-2/COX family instead in the R-initialed family demonstrates its capability of discovering interesting and meaningful semantic associations. An ontology inference engine that is able to derive sibling classes would hardly achieve the same meaningful ranking.

Without any preprocessing or prior knowledge about how the clinical notes are prescribed, the results with data graph alone do not show a strong pattern because of the frequent appearance of general terms. However, the noteworthy inclusion of "reflux" and "infantile" may be due to the causal relationships between rofecoxib and acid reflux and infantile gastroenteritis respectively, which have been discussed in the literature. On the other hand, adding the *is_a* graph to the data graph can be also seen as a mean for denoising and enhancement of the data. In the results with both data and *is_a* graphs, valdecoxib and celecoxib are promoted to the top results. This suggests that the evidences from both data and ontology conforms with previous studies in which celecoxib, valdecoxib are shown to be, similar to rofecoxib, also associated with increased risk of cardiovascular pathologies.

w/ data only	w/ *is_a* only	w/ both data and *is_a*
reflux	valdecoxib	reflux
medical history	meloxicam	obstruction
history of previous events	celecoxib	injury
diagnosis	parecoxib	valdecoxib
pharmaceutical preparations	etoricoxib	medical history
blood and lymphatic disorders	deracoxib	foreign body sensation
disease	lumiracoxib	history of previous events
infantile neuroaxonal dystrophy	firocoxib	adverse effects
today	nabumetone	celecoxib
hypersensitivity	macrolides	actual hypothermia

TABLE I. RESULTS OF ITEMS RANKED BY THE STRENGTH OF SEMANTIC ASSOCIATION WITH THE TERM "ROFECOXIB."

To verify the drug-disease association and study the effect of different semantic relationships, we carry out the following experiment. Table II illustrates rankings of three associations (one per row) under different settings (data alone, data plus *is_a*, and data plus *may_treat*, respectively). The first element in the association is the query item, which are all active ingredients of some prescription drugs, and the ranking shown in the table is for the second item, which are diseases. E.g., arthritis is ranked as the 527th semantic association to rofecoxib based on the similarity from the data graph alone. All these item pairs are known gold standard drug-disease relationships.

	data only	w/ *is_a*	w/ *may_treat*
$\langle rofecoxib, deg.\ polyart. \rangle$	527	632	13
$\langle valdecoxib, deg.\ polyart. \rangle$	613	695	17
$\langle troglitazone, diabetes \rangle$	478	514	11

TABLE II. RANKINGS OF THREE SEMANTIC ASSOCIATIONS UNDER DIFFERENT SETTINGS.

We observe that the ranking based on data graph alone is fairly high already, considering there are approximately 1 million concepts of interest. However, the result based on the combination of data and *is_a* graph is worse. It is because the subsumption hierarchies for drugs and diseases are largely separate structures. Therefore the "is_a" relationships can only boost the association within the hierarchies, but obfuscate the cross-hierarchy associations that we aim to find. On the other hand, however, the association between these pairs can be exactly captured by the NDFRT "may_treat" relationship (e.g., NDFRT explicitly defines that rofecoxib "may_treat" arthritis). When the *may_treat* graph is incorporated into the mining process, the ranking for the association is greatly boosted.

2) Detecting Misinformation in Ontologies: Conversely, we are also interested in learning whether the data graph can help discover misinformation in ontologies. Figure 4 (left) shows a subgraph of the NDFRT "may_treat" relationship. According to the ontology, rofecoxib can treat two diseases, namely, dysmenorrhea and degenerative polyarthritis. There are also 116 and 200 other drugs known to treat dysmenorrhea and degenerative polyarthritis respectively. To simulate an imperfect ontology, we alter the ground truth graph by introducing some deliberate misinformation, as shown in Figure 4 (right). Specifically, we assert that rofecoxib may treat hypertensive disease, which in fact can be treated by the most number of drugs (619 in total) according to the NDFRT ontology. Then we add an imaginary drug to treat degenerative polyarthritis, dysmenorrhea, and hypertensive disease. In this way, the original immediate connections between rofecoxib and degenerative polyarthritis and dysmenorrhea are broken.

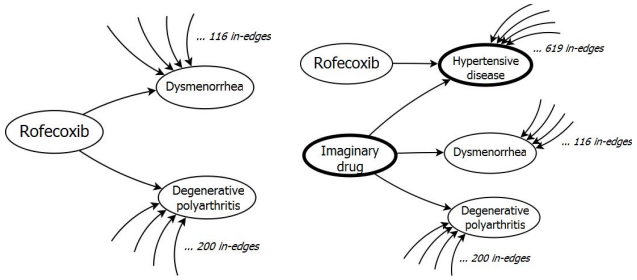


Fig. 4. The left part of the figure shows the ground-truth may_treat relationships between the drug rofecoxib and two diseases. The right part shows the same subgraph with deliberate falsehoods.

Table III shows the result of ranks of associations between rofecoxib and degenerative polyarthritis and dysmenorrhea

	noisy *may_treat*	noisy *may_treat* w/ data
$\langle rofecoxib, deg.\ polyart. \rangle$	555	263
$\langle rofecoxib, dysmenorrhea \rangle$	246	1703

TABLE III. RANKINGS OF ASSOCIATIONS ON THE NOISY *may_treat* GRAPH (FIGURE 4 RIGHT) DERIVED WITH AND WITHOUT DATA.

respectively. The ranks of the associations drastically drop on the noisy graph. This is mainly due to the presence of a large node, hypertensive disease, in the middle of the connections. However, with the incorporation of the *may_treat* graph, we notice that the rank involving degenerative polyarthritis increases, while the other involving dysmenorrhea drops even further, which implies the data graph endorses more strongly the former association. Indeed, although rofecoxib is known to treat both degenerative polyarthritis and dysmenorrhea, the former is a much more popular usage. A search on the PubMed database[1] for "rofecoxib and polyarthritis" returns 518 results, while "rofecoxib and dysmenorrhea" only returns 29. This result shows that the data graph can help correct misinformation in ontologies to some extent, and can also give a clue on how prior beliefs fit with reality.

V. CONCLUSION AND FUTURE WORK

We propose to mine biomedical ontologies and data using a unified RDF hypergraph representation. We use random walks with restart on the unified graph to discover semantic associations that cannot be found by only co-frequencies. We allow users to customize the weight of each semantic component, providing flexibility to express how strongly the role of the ontology plays over the data, or vice-versa. Our evaluations show that the method discovers semantic associations and that it scales to size of both data and ontologies. Moreover, we also show that our methods can discover and suggest corrections for misinformation in biomedical ontologies.

In the following we discuss some future research directions. First, the scalability of semantic data mining algorithms is of critical importantance. While the size of practical problem is bound to increase, the graph-centric root of our method makes it possible to leverage decades of work on graph mining. A possible direction is to develop parallelizable algorithms. Second, the appropriate ratio for the edge weights is not only dependent on a variety of factors. We plan to explore automatic ways to determine optimal hyperedge weights going forward. Third, to enable the RDF bipartite to incorporate more complicated semantics, we will study the possible approach that models some domain constraints by explicitly describing the desired or acceptable walk (traversal sequence) in the RDF hypergraph.

VI. ACKNOWLEDGMENT

This work is partially supported by the NIH/NIBIB with Grant No. R01EB007684 and NIH/NIGMS with Grant No. R01GM103309. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the supporting institutions.

[1] http://www.ncbi.nlm.nih.gov/

REFERENCES

[1] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M. A. Storey, C. G. Chute, and M. A. Musen, "Bioportal: ontologies and integrated data resources at the click of a mouse," *Nucleic Acids Research*, 2009.

[2] "Web Ontology Language," `http://www.w3.org/TR/owl-ref/`.

[3] J. Hayes and C. Gutierrez, "Bipartite Graphs as Intermediate Model for RDF," in *ISWC*, 2004, pp. 47–61.

[4] L. Getoor and C. P. Diehl, "Link Mining: A Survey," *SIGKDD Explor. Newsl.*, vol. 7, pp. 3–12, December 2005.

[5] M. Chein and M.-L. Mugnier, *Graph-based Knowledge Representation: Computational Foundations of Conceptual Graphs*. Springer, 2008.

[6] D. Zhou, J. Huang, and B. Scholkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *NIPS*, 2007, p. 1601.

[7] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens, "Random-Walk Computation Of Similarities Between Nodes Of A Graph, With Application To Collaborative Recommendation," *TKDE*, vol. 19, no. 3, pp. 355 –369, 2007.

[8] Y. Zhou, H. Cheng, and J. X. Yu, "Graph Clustering Based On Structural/Attribute Similarities," *PVLDB*, vol. 2, pp. 718–729, 2009.

[9] H. Liu, P. LePendu, R. Jin, and D. Dou, "A Hypergraph-based Method for Discovering Semantically Associated Itemsets," in *ICDM*, 2011, pp. 398–406.

[10] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," in *VLDB*, 1994, pp. 487–499.

[11] D. R. Swanson, "Two medical literatures that are logically but not bibliographically connected," *Journal of the American Society for Information Science*, vol. 38, no. 4, pp. 228–233, Jan. 1999.

[12] T. Gene Ontology Consortium, "Creating the Gene Ontology Resource: Design and Implementation," *Genome Research*, vol. 11, no. 8, pp. 1425–1433, 2001.

[13] C. Jonquet, P. LePendu, S. Falconer, A. Coulet, N. F. Noy, M. A. Musen, and N. H. Shah, "Ncbo resource index:ontology-based search and mining of biomedical resources," *Web Semantics*, vol. 9, no. 3, pp. 316–324, 2011.

[14] Y. Liu, P. LePendu, S. Iyer, M. Udell, and S. N. H., "Using temporal patterns in medical records to discern adverse drug events from indications," in *AMIA Summit on Clinical Research Informatics*, 2012, pp. 47–56.

[15] S. Staab and A. Hotho, "Ontology-based text document clustering." in *IIPWM*, 2003, pp. 451–452.

[16] B. Adryan and R. Schuh, "Gene-Ontology-based clustering of gene expression data," *Bioinformatics*, vol. 20, pp. 2851–2852, 2004.

[17] J. Wen, Z. Li, and X. Hu, "Ontology Based Clustering for Improving Genomic IR," in *IEEE CBMS*, 2007, pp. 225–230.

[18] H. T. Lin, N. Koul, and V. Honavar, "Learning relational bayesian classifiers from RDF data," in *ISWC*, 2011, pp. 389–404.

[19] V. Bicer, T. Tran, and A. Gossen, "Relational Kernel Machines for Learning from Graph-Structured RDF Data." in *ESWC*, 2011, p. 47.

[20] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," *PVLDB*, vol. 4, no. 11, pp. 992–1003, 2011.

[21] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu, "Integrating meta-path selection with user-guided object clustering in heterogeneous information networks," in *KDD*, 2012, pp. 1348–1356.

[22] C. Berge, "Hypergraphs," *Bull. Symbolic Logic*, 1989.

[23] S. S. Sahoo, W. Halb, S. Hellmann, K. Idehen, T. T. Jr, S. Auer, J. Sequeda, and A. Ezzat, "A Survey of Current Approaches for Mapping of Relational Databases to RDF," W3C, Tech. Rep., 2009.

[24] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos, "Neighborhood Formation and Anomaly Detection in Bipartite Graphs," in *ICDM*, 2005, pp. 418–425.

[25] J. Chen, O. R. Zaiane, R. Goebel, and P. S. Yu, "Tuplerank: Ranking relational databases using random walks on extended k-partite graphs," University of Alberta, Tech. Rep., 2009.