Mining Biomedical Ontologies and Data with Unified RDF Hypergraphs

Haishan Liu
University of Oregon
Eugene, OR, 97403, USA
ahoyleo@cs.uoregon.edu

Dejing Dou
University of Oregon
Eugene, OR, 97403, USA
dou@cs.uoregon.edu

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

Paea LePendu
Stanford University
Stanford, CA, 94305, USA
plependu@stanford.edu

Nigam Shah
Stanford University
Stanford, CA, 94305, USA
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 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 in *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 use random walk with restart to efficiently calculate the similarity between concepts as a way to measure associations. 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.

Keywords

Semantic Data Mining, RDF Hypergraphs, Semantic Associations, Biomedical Ontologies

1. INTRODUCTION

Researchers around the world are linking more and more data to ontologies which are formal specification of concepts and relationships in various domains. Formal ontologies have been extensively developed and harnessed in scientific research, particularly in biomedical research. Knowledge evolves rapidly in biomedicine and has promoted the creation and use of ontologies to advance scientific progress. Besides the size of data increases exponentially, an increasing number of "large" biomedical ontologies are developed. Prominent examples of this effort include the Gene Ontology (GO) [9] and the Unified Medical Language System (UMLS) [16]. Over 300 ontologies have been loaded into the National Center of Biomedical Ontology (NCBO) BioPortal library at Stanford [19], 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. Previous research has not been sufficient to address these challenges. For example, some approaches have utilized ontologies in data mining, but usually only a small portion of ontologies is used (typically the subsumption relationship) on very few tasks (most often concept aggregation). On the other hand, limited attempts have been made to check errors in large ontologies. Traditional approaches to sanitize knowledge bases by using an inference engine for logic reasoning and consistency checking is hard to scale.

With the increasing amount of ontology-annotated data, new possibilities are opened up for both data mining and ontology development. Therefore an emerging research direction, which we call *semantic data mining*, focuses 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 pattern discovered in a noisy environment. It also benefits the 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 [1] 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 [11]. 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 [10]. These graphs may represent both homogeneous networks and heterogeneous networks.

Hence, our semantic data mining approach is inspired by a combination of graph representation [6], hypergraphs [29], and random walks [8, 30]. This paper extends our previous work that implements a hypergraph-based approach to learn associations from interlinked data (without ontologies) [17]. The RDF hypergraph representation proposed by Hayes et al. [11] is a key innovation which connects data and ontologies. We adopt this representation in our approach because properties in RDF triples can be represented as first class objects among all interrelated objects, enabling us to embed both the ontology and the data together and to serialize it into a bipartite graph representation for scalable processing.

In RDF hypergraphs, we use random walk with restart to efficiently calculate the similarity between concepts to determine their associations from large set of data and ontologies. Traditional association mining relies on co-frequencies of items (concepts) within transactions [3]. We look one step further to find indirectly associated items (concepts). This extension has far reaching applications in biomedicine. For instance, consider a simple scenario illustrated by Swanson [27] years ago while studying Raynauld's syndrome. He noticed that the literature discussed Raynauld's syndrome (Z), a peripheral circulatory disease, together with certain changes of blood in human body (Y); and, separately, the consumption of dietary fish oil (X) was also linked in the literature 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., $X \rightsquigarrow Y \rightsquigarrow Z$. We call such indirect associations, (X, Z), *semantic associations*.

We evaluate the effectiveness of the results on large real-world biomedical data and ontologies. We show the proposed method is indeed capable of capturing semantic (indirect) associations while seamlessly incorporate domain knowledge defined in formal ontologies. We also show that our data mining methods can discover misinformation in biomedical ontologies. Our work makes the following main contributions: First, we employ a RDF hypergraph representation to capture both semantics of ontologies and data. We can weight each hyperedge so that certain links (such as *is_a* or *may_treat* relationships) can carry appropriate strength. Next, we serialize the hypergraph and weighted hyperedges into a bipartite representation for efficient processing. Then, we implement highly efficient and scalable random walks with restart over the bipartite graph to generate semantic associations, including associations that may not necessarily be co-frequent. The discovered semantic associations can be used to detect potential errors in biomedical ontologies.

2. RELATED WORK

2.1 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 [9]). These publicly available GO annotation databases make enrichment analysis possible, which enables researchers to functionally profile sets of interesting genes identified by microarray experiments [13]. At Stanford, the National Center for Biomedical Ontology (NCBO) annotates large volumes of biomedical text for search and mining [12] and has been used, for instance, to profile disease research [18]. Finally, millions of patient electronic health records are being annotated using medical ontologies like SNOMED-CT in efforts to advance patient healthcare [23].

In general domains, Staab and Hotho [22] 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. [2] enable cluster visualization for gene expression data by navigating various levels of Gene Ontology hierarchy. Wen et al. [28] take into consideration the ontology hierarchy to offset biases toward overly-general terms in text mining.

2.2 Graphs in Mining RDF and Ontologies

RDF data and OWL ontologies can be represented as graphs for data mining. Lin et al. [15] treat the RDF triple store as a datasource during mining and develop Relational Bayesian Classifiers (RBCs) that aggregate SPARQL queries. Kiefer et al. [14] extend the SPARQL query language to enable creating and working with the data mining model. Similarly, Bicer et al. [5] 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. [25, 26]. HIN leverages semantics of various types of nodes and links in a network for the graph and network mining tasks. Since ontologies and annotated data can be treated as a large graph of concepts and entities linked with different type of relationships, there should be a way to represent certain relationships and semantics (e.g., class subsumption) of ontologies and data in HIN. However, considering many biomedical ontologies are large, manually representing them in heterogeneous information networks is not practical. It is not easy to build an automatic translator from OWL ontologies to HIN either. We prefer RDF Hypergraphs because the RDF syntax of OWL ontologies and data makes it much easier to automatically transform existing biomedical ontologies and their annotated data into RDF hypergraphs.

3. METHOD

3.1 Graph Representation for Biomedical Ontologies

The Web Ontology Language (OWL [1]) 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):

$$\boxed{\text{subject}} \xrightarrow{predicate} \boxed{\text{object}}\,.$$

One disadvantage of DG is that it makes an artificial distinction between resources and properties, which leads to incongruous representations. Consider the RDF statements in the following example:

⟨*Lipitor has_ingredient Active_Ingredient*⟩,
⟨*Active_Ingredient has_chemical Atorvastatin_Calcium*⟩,
⟨*has_chemical subPropertyOf has_ingredient*⟩.

From the first two statements, *Lipitor* (a drug made by Pfizer), *Active_Ingredient* and *Atorvastatin_Calcium* can be represented in a DG as nodes connected by edges *has_ingredient* and *has_chemical* respectively. However, from the last statement, to express the relationship between *has_chemical* and *has_ingredient*, these concepts have to be represented as nodes themselves. Representing this set of statements in DG inevitably separates the information into two inconsistent subgraphs, making it difficult for graph-based methods to utilize the information in a holistic and systematic manner.

To overcome the inconsistency, Hayes et al. [11] proposed to model RDF as a *hypergraph.* A hypergraph [4] is a generalization of a traditional 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 meta-data and data level statements can be 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 an *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 have many desirable properties for developing intuitive mining algorithms because they turn hypergraphs into a simple form so that many algorithms designed on simple graphs can be readily applied. Therefore, we propose to use bipartite graphs to represent domain knowledge and data expressed in RDF.

3.2 Graph representation for Ontology-Annotated Data

There already exist methods for transforming data, such as those in relation databases, into RDF [21]. 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 [12]. 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 (Table 1). 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.

	f_1	$\cdots$	f_n
r_1 :	0	$\cdots$	1
$\vdots$	$\vdots$	$\ddots$	$\vdots$
r_m :	1	$\cdots$	0

(A)

s	p	o
<r_1	`mentions`	f_n>
<r_m	`mentions`	f_1>

(B)

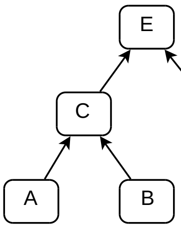
Table 1: Ontology-annotated data (A) is a feature matrix attributing classes from ontologies, f_i, to entities such as documents, r_j, that is easily represented in sparse matrix form using RDF triples (B).

3.3 Unified Graph Representation for Biomedical Ontologies and Data

In order to facilitate the synergy between data and domain knowledge, information from both sources needs to be first combined. This is achieved by the process called *semantic annotation.* Semantic annotation aims at assigning formal semantic descriptions to the basic element of data, and it is crucial in realizing semantic data mining by bridging formal semantics in domain knowledge with data. If data is annotated, a unified graph incorporating information from both data and ontologies can be created. The following example demonstrates the combination of an ontology graph and a data graph.

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, 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



(A)

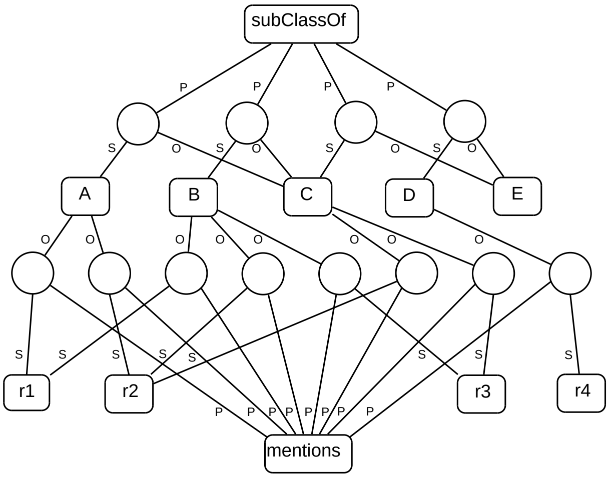
	A	B	C	D	E
$A:$	0	0	1	0	0
$B:$	0	0	1	0	0
$C:$	0	0	0	0	1
$D:$	0	0	0	0	1
$E:$	0	0	0	0	0

(B)

	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

(C)

Figure 1: Five concepts (*A*–*E*) are represented visually as a hierarchy (A) and also as a 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*.



(A)

	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>`

(B)

Figure 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.

(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 properties.

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 Equation 1. The internal block structure of the concatenated biadjacency matrix $\mathbf{M}$ is shown in Equation 2.

$$\mathbf{M} = \begin{bmatrix} w_d\mathbf{M}_d & w_{o_1}\mathbf{M}_{o_1} & w_{o_2}\mathbf{M}_{o_2} & \ldots \end{bmatrix} \tag{1}$$

$$\mathbf{M} = \begin{array}{c} \\ r \\ a \end{array} \begin{array}{c|c|c|c} ds & os_1 & os_2 & \ldots \\ \mathbf{M}_{dr} & \mathbf{0} & \mathbf{0} & \ldots \\ \mathbf{M}_{da} & \mathbf{O}_1 & \mathbf{O}_2 & \ldots \end{array} \tag{2}$$

With the RDF bipartite graph and its biadjacency matrix

$$\overbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}^{V_s}$$

$$\mathbf{M}_d = \begin{array}{c|ccccc|cc}
 & \multicolumn{5}{c}{\overbrace{\qquad\qquad\qquad\qquad}^{V_d}} & & \\
 & s_5 & s_6 & s_7 & \ldots & s_{12} & & \\
r_1 & 1 & 1 & 0 & & 0 & & \\
r_2 & 0 & 0 & 1 & \cdots & 0 & \} V_r & \\
r_3 & 0 & 0 & 0 & & 0 & & \\
r_4 & 0 & 0 & 0 & & 1 & & \} V_v \\
A & 1 & 0 & 1 & & 0 & & \\
B & 0 & 1 & 0 & & 0 & & \\
C & 0 & 0 & 0 & \cdots & 0 & \} V_a & \\
D & 0 & 0 & 0 & & 1 & & \\
E & 0 & 0 & 0 & & 0 & &
\end{array}
\qquad
\mathbf{M}_o = \begin{array}{c|cccc|cc}
 & \multicolumn{4}{c}{\overbrace{\qquad\qquad\qquad}^{V_o}} & & \\
 & s_1 & s_2 & s_3 & s_4 & & \\
r_1 & 0 & 0 & 0 & 0 & & \\
r_2 & 0 & 0 & 0 & 0 & \} V_r & \\
r_3 & 0 & 0 & 0 & 0 & & \\
r_4 & 0 & 0 & 0 & 0 & & \} V_v \\
A & 1 & 0 & 0 & 0 & & \\
B & 0 & 1 & 0 & 0 & & \\
C & 1 & 1 & 1 & 0 & \} V_a & \\
D & 0 & 0 & 0 & 1 & & \\
E & 0 & 0 & 1 & 1 & &
\end{array}$$

Figure 3: An example RDF bipartite graph and a detailed anatomy of its biadjacency matrix.

defined, in the following, we move on to describe our method to mine semantic associations based on traversing the graph using random walk with restart.

3.4 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 $b(\neq a) \in G_d \cap G_o$. The result is a one-column vector containing all similarity scores with respect to a [7]. 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. 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.

In particular, RWR on a bipartite graph works as follows: assume we have a random walker that starts from node a. For each step, the walker chooses randomly among the available edges from the current node. After each iteration, with probability c, it resets its position back to node a. The final steady-state probability that the random walker reaches node b is the similarity score of b with respect to a. We choose the random walk approach to compute the relevance score because it gives node b high ranking if b and a are connected by many nodes; this is due to the random walker having more paths to reach b from a. The purpose of the periodic restart is to raise the chance that close related nodes are visited more often than other nodes.

The RWR score on bipartite graphs has a desired property that it is easier to compute when the numbers of nodes in the two parts are highly unbalanced. The unified RDF bipartite graph of ontologies and data satisfies this condition because there are generally many 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. The algorithm can be used in situations where, for example, users are interested in knowing products that are usually bought together in the same transactions by different customers, or common side effects of the same drugs prescribed to different patients, etc.

Given a biadjacency matrix $\mathbf{M}$ in Equation 1 for the unified RDF bipartite graph G, we can construct the adjacency matrix $\mathbf{A}$ of G as following:

$$\mathbf{A} = \left[\begin{array}{cc} \mathbf{0} & \mathbf{M} \\ \mathbf{M}^T & \mathbf{0} \end{array} \right].$$

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 need to compute a $(k+n)\times 1$ steady-state probability vector $\mathbf{u}_a$ over all nodes in 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 iterative equation.

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

$$\mathbf{u}_a = (1-c)\mathbf{P}_A\mathbf{u}_a + c\mathbf{q}_a \ . \tag{3}$$

Algorithm 1 Calculate Semantic Association

```
Input: query attribute $a$, bipartite matrix $M$, restarting
  probability $c$, tolerant threshold $\epsilon$
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 $\mathbf{u}_a(1:k)$
```

The iterative update of $\mathbf{u}_a$ can be performed as shown in Algorithm 1. The while loop is modified from Equation 3 to avoid materializing $\mathbf{A}$ and $\mathbf{P}$ for scalability.

4. EXPERIMENT

In this section, we evaluate the method of random walk with restart on the unified RDF bipartite graph for discovering semantic associations and detecting misinformation in biomedical ontologies. We conducted a series of experiments to demonstrate the effect of the incorporating the ontologies in the mining task. We evaluated our methods on an *electronic health records* dataset to highlight its scalability and applicability for problems in the biomedical domain.

4.1 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.

One strength of the Annotator is the highly comprehensive and interlinked lexicon that it uses. It can incorporate the entire NCBO BioPortal ontology library of over 250 ontologies to identify biomedical concepts from text using a dictionary of terms generated from those ontologies. Terms from these ontologies are linked together via mappings. For this study, we specifically configured the workflow to use a subset of those ontologies that are most relevant to clinical domains, including Unified Medical Language System (UMLS) terminologies such as SNOMED-CT, the National Drug File (NDFRT) and RxNORM, as well as ontologies like the Human Disease Ontology. The resulting set of ontologies contains 1 million subsumption statements.

From this set of 1.6 million patients with annotated records, we vectorize texts and turned them into a huge 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.

We applied our algorithms to all previous records in the patient's timeline, looking at just the set of drugs and their semantically related diseases. Therefore, at a very simplistic level, the experiment result shows that strong semantic associatoins in this context could possibly represent sets of drugs that could lead toward certain diseases. To summarize, the size of the dataset in terms of numbers of RDF statements in the bipartite graph is shown in Table 2.

# data stmts	# is_a stmts	# may_treat stmts
148,690,056	1,048,604	43,780

Table 2: Numbers of RDF statements in the unified bipartite graph extracted from the electronic health dataset.

4.2 Results

4.2.1 Discovering Semantic Associations

Before studying the drug-disease association, we first carried out a study on the drug-drug association. To this purpose, we combine the subsumption hierarchy in the ontology graph with the data graph. Table 3 demonstrates semantic association with 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 non-steroidal anti-inflammatory drugs part of the COX-2 inhibitor class of drugs.

Table 3 shows that, with only the ontology graph, the algorithm successfully picks up almost all other active ingredients part 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, although rofecoxib is a subclass of COX-2 inhibitor drugs, it is also a derived class from a much broader parent called "Drug

Products by Generic Ingredient Combinations," whose subclasses are organized by descendants' initial alphabets. In other words, rofecoxib is 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 never be able to achieve the same meaningful ranking as our algorithm does in this case.

Without any preprocessing and prior knowledge about how the clinical notes are prescribed, the results with data graph alone do not seem to have a strong pattern because of the appearance of too many 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 that have been shown in previous studies. Applying general information extraction techniques, such as pruning out general terms ("medical history, " "today," etc.) should be able to improve the performance with data graph alone.

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.

To verify the drug-disease association and study the impact of different semantic relationships on finding such association, we carry out the following experiment. Table 4 illustrates the 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 pair 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. For example, arthritis is ranked as the 527th semantic association to rofecoxib according to similarity ranking based only on data graph. All these item pairs are actually gold standard associations backed by known drug-disease relationships, we know the strength of associations between them should be strong.

We observe that the ranking based on data graph alone is fairly high already, consider there are approximately 1 million concepts of interest. However, the results based on the combination of data and subsumption ("is_a") graph are worse. It is because the subsumption hierarchies for drugs and diseases are largely separate structures. Therefore the subsumption relationships can only boost the association within the drug and disease hierarchies respectively, but obfuscate the cross-hierarchy associations that we aim to find between drugs and diseases. 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.

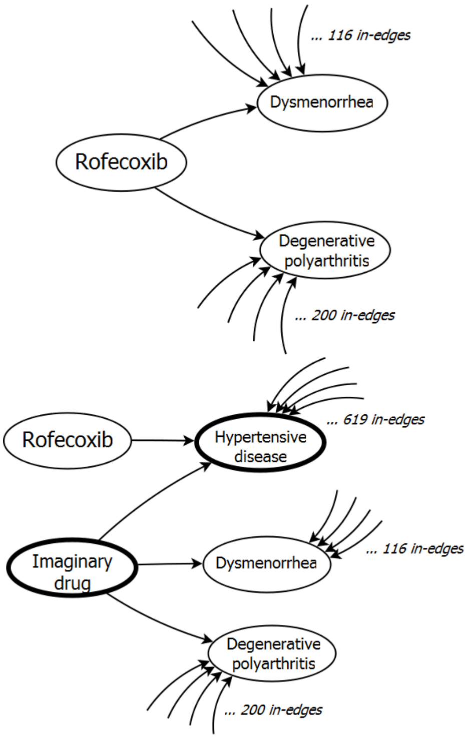


Figure 4: The upper part of the figure shows the ground-truth may_treat relationships between the drug rofecoxib and two diseases. The lower part shows the same subgraph with deliberate falsehoods.

4.2.2 Detecting Misinformation in Ontologies

Conversely, we are also interested in learning whether the data graph can help discover misinformation in ontologies. Figure 4 (upper) 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 (hence the in-degrees of the nodes). To simulate an imperfect ontology, we alter the ground truth graph by introducing some deliberate misinformation and falsehoods, as shown in Figure 4 (lower). In more details, we specify 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 another imaginary drug to treat degenerative polyarthritis, dysmenorrhea, and hypertensive disease. In this way, the original immediate connections between rofecoxb and degenerative polyarthritis and dysmenorrhea become erroneously indirect and are obfuscated by the noise of high-degree nodes along the path. With this setup, we hope to learn if the incorporation of data graph can help correct the ontology.

Table 5 shows the result of ranks of the associations between rofecoxib and degenerative polyarthritis and dysmenorrhea respectively. The ranks of the associations drastically drop to the 555th and 246th respectively on the noisy graph from the top two on the original ground truth graph. This is mainly due to the presence of a large node, hypertensive disease, in the middle of the connections. However, with

rank	w/ data only	w/ is_a only	w/ both data and is_a
1	reflux	valdecoxib	reflux
2	medical history	meloxicam	obstruction
3	history of previous events	celecoxib	injury
4	diagnosis	parecoxib	valdecoxib
5	pharmaceutical preparations	etoricoxib	medical history
6	blood and lymphatic system disorders	deracoxib	foreign body sensation
7	disease	lumiracoxib	history of previous events
8	infantile neuroaxonal dystrophy	firocoxib	adverse effects
9	today	nabumetone	celecoxib
10	hypersensitivity	macrolides	actual hypothermia

Table 3: Results of items ranked by the strength of semantic association with the term "rofecoxib."

	w/ data only		w/ data and "is_a"		w/ data and "may_treat"	
	p(%)	rank	p(%)	rank	p(%)	rank
$\langle \mathit{rofecoxib}, \mathit{degenerative\ polyarthritis}\rangle$	0.006	527	0.004	632	0.51	13
$\langle \mathit{valdecoxib}, \mathit{degenerative\ polyarthritis}\rangle$	0.007	613	0.005	695	0.63	17
$\langle \mathit{troglitazone}, \mathit{diabetes}\rangle$	0.006	478	0.005	514	0.44	11

Table 4: Rankings of three semantic associations under different settings.

the unified data and may_treat graph, we notice that the rank of rofecoxib and degenerative polyarthritis increases to 263rd, while the rank of rofecoxib and dysmenorrhea decreases to 1703rd. This shows that the data graph endorses more strongly the association between rofecoxib and degenerative polyarthritis. Indeed, although rofecoxib are known to treat both degenerative polyarthritis and dysmenorrhea, the former is a much more popular usage. A search on the National Library of Medicine's 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 in a sense, it also gives a clue of how prior beliefs fit with reality.

5. CONCLUSION AND FUTURE WORK

We mine biomedical ontologies and data using a unified RDF hypergraph representation. Our method uses random walks with restart over a bipartite graph serialization to discover semantic associations that cannot be found by methods that rely on co-frequencies. We are one of the first research groups to consider mining biomedical ontologies and data together without using a separate system for pre-processing or pre-computing ontologies. 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 to both the size of data and the size of 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 in mining the unified RDF bipartite graph.

Developing scalable semantic data mining algorithms is critically important. The electronic health dataset in our study has grown beyond 100 billion triples and the size of the ontologies also tremendous (SNOMED-CT has nearly 400,000 classes). The RWR method works well for query-based node similarity, but it is not applicable to generate full pair-wise node similarities at such scale. While the size of practical problem is bound to increase, the graph-based formalism of our method makes it possible to leverage decades of work on graph mining. A promising direction going forward is to employ approximation and develop parallelizable algorithms.

The weighted hyperedges provide a great deal of flexibility to users who may prefer domain knowledge over data (or vice versa) and opens up new research questions on how to optimally configure learning algorithms. In reality, the appropriate ratio for the edge weights is not only dependent on the size of graphs but also the graph configuration (depth, average degree, etc). Moreover, allowing users to specify the ratio of prior knowledge in ontologies versus inductive evidence from data enables us to discover empirically optimal configurations. We would explore a few permutations on these hyperedge weights going forward.

The RDF bipartite graph representation has limited expressivity compared to OWL itself. For example, domain, range and cardinality constraints are not straightforward to model. One possible approach is to model domain constraints by explicitly describing the desired or acceptable walk (traversal sequence) in the RDF hypergraph. In this case, the recently proposed *regular traversal expression* [20] technique may apply. However, their fast power-iteration approach for computing the stationary probability may not be applicable any more due to the label sequence constraint, but the Monte-Carlo simulation of the random walk may help to approximate the similarity measure.

We will continue to focus on healthcare datasets mainly because large number of biomedical ontologies have been developed and used to annotate the real-life data, but also because of the enormous potential in advancing the state-of-the-art in clinical informatics and improving the quality of care for millions of patients. This research also contributes to an emerging research direction of semantic data mining, in which formal semantics that exist in data and knowledge can be represented and incorporated into the data mining process in a seamless way.

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

	w/ noisy may_treat only		w/ data and noisy may_treat	
	p(%)	rank	p(%)	rank
$\langle rofecoxib, degenerative\ polyarthritis\rangle$	3.60e-3	555	8.14e-3	263
$\langle rofecoxib, dysmenorrhea\rangle$	1.54e-2	246	1.26e-3	1703

Table 5: Rankings of associations on the noisy may_treat graph (Figure 4 right) between Rofecoxib and two diseases derived with and without data.

6. REFERENCES

[1] OWL Web Ontology Language. `http://www.w3.org/TR/owl-ref/`.

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

[3] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *VLDB*, pages 487–499, 1994.

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

[5] V. Bicer, T. Tran, and A. Gossen. Relational Kernel Machines for Learning from Graph-Structured RDF Data. In *ESWC*, pages 47–62, 2011.

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

[7] J. Chen, O. R. Zaiane, R. Goebel, and P. S. Yu. Tuplerank: Ranking relational databases using random walks on extended k-partite graphs. Technical report, Department of Computer Science, University of Alberta, 2009.

[8] 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*, 19(3):355 –369, 2007.

[9] T. Gene Ontology Consortium. Creating the Gene Ontology Resource: Design and Implementation. *Genome Research*, 11(8):1425–1433, 2001.

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

[11] J. Hayes and C. Gutierrez. Bipartite Graphs as Intermediate Model for RDF. In *ISWC*, pages 47–61, 2004.

[12] 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*, 9(3):316–324, 2011.

[13] P. Khatri and S. Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, 2005.

[14] C. Kiefer, A. Bernstein, and A. Locher. Adding data mining support to SPARQL via statistical relational learning methods. In *ESWC*, pages 478–492, 2008.

[15] H. T. Lin, N. Koul, and V. Honavar. Learning relational bayesian classifiers from RDF data. In *ISWC*, pages 389–404, 2011.

[16] D. Lindberg, B. Humphries, and A. McCray. The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291, 1993.

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

[18] 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*, pages 47–56, 2012.

[19] 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.

[20] M. A. Rodriguez and P. Neubauer. The graph traversal pattern. *CoRR*, abs/1004.1001, 2010.

[21] 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. Technical report, W3C, 2009.

[22] S. Staab and A. Hotho. Ontology-based text document clustering. In *IIPWM*, pages 451–452, 2003.

[23] P. E. Stang, P. B. Ryan, J. A. Racoosin, J. M. Overhage, A. G. Hartzema, C. Reich, E. Welebob, T. Scarnecchia, and J. Woodcock. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Annals of Internal Medicine*, 153(9):600–606, 2010.

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

[25] 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*, 4(11):992–1003, 2011.

[26] 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*, pages 1348–1356, 2012.

[27] D. R. Swanson. Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science*, 38(4):228–233, Jan. 1999.

[28] J. Wen, Z. Li, and X. Hu. Ontology Based Clustering for Improving Genomic IR. In *IEEE CBMS*, pages 225–230, 2007.

[29] D. Zhou, J. Huang, and B. Scholkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *NIPS*, pages 1601–1608, 2007.

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