International Journal of Distributed Systems and Technologies, 5(1), 22-39, 2014

Discovering Complex Relationships of Drugs over Distributed Knowledgebases

Juan Li, North Dakota State University, USA*
Ranjana Sharma, North Dakota State University, USA
Yan Bai, University of Washington Tacoma, USA

ABSTRACT

Drug discovery is a lengthy, expensive and difficult process. Indentifying and understanding the hidden relationships among drugs, genes, proteins, and diseases will expedite the process of drug discovery. In this paper, we propose an effective methodology to discover drug-related semantic relationships over large-scale distributed web data in medicine, pharmacology and biotechnology. We utilize semantic web and distributed system technologies, and developed a novel hierarchical knowledge abstraction and an efficient discovery protocol. Our approach effectively facilitates the realization of the full potential of harnessing the collective power and utilization of the drug-related knowledge scattered over the Internet.

Keywords - drug discovery; drug repositioning; distributed systems; relation, semantic web

INTRODUCTION

Drug discovery is a process of discovering and designing drugs. It is generally related to the fields of medicine, pharmacology and biotechnology. Despite the advances in chemical synthesis techniques as well as combinatorial and cheminformatics and understanding of biological systems, drug discovery is still a lengthy, expensive, and difficult process with a low rate of new therapeutic discovery. The process of drug development consists of drug compounds proposal, pre-clinical, clinical trial, and FDA review and approval (Anyanwu & Sheth, 2003). Usually, 5000 to 10,000 compounds are proposed for a potential drug (DiMasi et al., 2003; DiMasi et al., 1991; Masia 2009; Collier, 2009) . An extensive research of the proposed compounds is conducted to select 2.5% to 5% for preclinical trials to be tested on animals. Among the selected compounds for preclinical trial, only 2% may get approved for clinical trial. Finally, only 1 compound becomes an approved drug for the treatment of diseases and use on humans (Collier, 2009). Studies also show that it takes about 15 years from compound proposal to FDA approval of a new drug to treat a disease, and the total cost is between 0.8 to 1 billion US dollars (DiMasi, 2003).

To overcome the aforementioned problems of drug discovery, drug repositioning (Sleigh & Barton, 2010) has been proposed. Drug repositioning is the application of the existing drugs to new indications or new diseases. An existing drug has passed significant pre-clinical and clinical tests. Its toxicity and other effects are already known. Hence, the cost of using it for some other diseases will be much less as compared to developing a drug from scratch (Hopkins, 2008; Boran & Iyengar, 2010; Druker et al., 1996). Conventional drug design follows the principle of “one gene, one disease, one drug”. One drug is targeted for the treatment of one disease caused by one gene (Sleigh & Barton 2010). Unlike conventional drug design, drug repositioning studies interactions of drugs with multiple targets Hopkins, 2008; Boran & Iyengar, 2010). One drug can be used with multiple diseases. Some repositioned drugs have been approved with new uses that are different from original uses (Chong & Sullivan, 2007; Verma et al., 2005). Here are some examples.

Thalidomide normally as a drug for sedation, nausea and insomnia is being used in the treatment of multiple myloma (Durk, 2006). Acetylsalicylic acid, as a drug for reducing aches and pains and fever is being used in cardiology to prevent heart attacks, strokes and blood clotting due to its antiplatelet activity (Krumholz et al., 1995). Also Miltefosine for the treatment of Cancer has the new indication as Visceral leishmania (Sundar et al., 1998). Another very common example is of Sildenafil which was earlier used for hypertensionis now being used for male erectile dysfunction in the name of Viagra (Boolell et al., 1996).

Presently, most of repositioned drugs are developed by observing the side-effects of other drugs (Aronson, 2007; Ashburn &Thor, 2004). However, it will be more effective to study the polypharmacological action of drugs and examine proteins, genes, pathway and other important factors, rather than discovering the effects merely by observation. The hidden relationship between drugs and diseases could not be observed and identified incipiently, so does the information about the drug and its new applications.

With the development of Semantic Web technologies, more and more Semantic Web data including data related to proteins, genes, drugs, disease are generated. For example, the Bio2RDF project generates a network of the life science data. Different databases across life sciences platform have been linked using open-source Semantic Web technologies to provide support biological knowledge discovery (Nolin et al., 2008). The Linking Open Drug Data project brings the data sources about drugs, Chinese medicine, clinical trials, diseases and pharmaceutical companies together onto the Web of Linked Data and facilitates the integration of about 8.4 million data (Samwald et al., 2011). To effectively utilize the large amount of semantic data, efficient search mechanisms for Semantic Web data have been proposed for both humans and software agents. For instance, the Semantic Search scans objects to capture instances in a given data set (Guha et al., 2003). By utilizing keywords, the Swoogle search engine retrieves semantic entities as Uniform Resource Identifiers (URI) (Ding et al., 2004). To support complex queries over Resource Description Framework (RDF) bases, query languages, such as SPARQL, have been used to express various restrictions on semantic entities and relationships (http://www.w3.org/TR/rdf-sparql-query/).
These technologies effectively assist human and software agents to locate desirable information from large amounts of semantic data on the Web. However, they may not reveal hidden knowledge large semantic datasets. For instance, knowledge of determining the complex relationships between multiple semantic entities, i.e., “how drugs X and Y are related.”

In Semantic Web, a sequence of complex relationship between the semantic entities forms a semantic association. The entities could be from disparate sources. More than one path may exist between the entities or they may be associated in more than one way. An entity may be related to another entity directly or through one or more intermediate entities. A semantic association where an entity is related to another entity through intermediate entities is extremely important in drug discovery. Let’s look at how drugs work in the human body: The chemicals in the body may be broken down to make simpler chemicals or form other chemicals used in the functioning of the body. It involves various enzymes, proteins, and genes. When one molecule is transformed into another molecule through a series of steps, the process is called as pathway. The pathway is catalyzed by enzymes at various steps. Diseases affect the body physically and may affect the internal processes of the body and various pathways. In order to cure the diseases, drugs are used as foreign molecules. These foreign molecules are converted into other molecules and help in curing the disease. This could involve inhibiting the pathway of various organisms including bacteria causing the disease. Therefore, the drugs can be related to the diseases via intermediate protein targets. The drugs and the diseases may also be related via genes and pathway they inhibit. All these complex relationships among drugs, diseases, enzymes, proteins, and genes can be effectively captured by semantic links and associations. Currently, information about the various

molecules in the body and their properties, functions and their mutual interacting processes is widely available in the public knowledgebases such as DrugBank (Knox et al. 2011), KEGG (Kanehisa et al. 2010; Kanehisa et al. 2006; Kanehisa et al. 2004), OMIM (Hamosh et al. 2004).

Although there have been literature on discovering semantic links and relations (Meza et al. 2006, Sheth et al. 2005; Heim et al. 2010; Heim et al. 2010; Lehmann et al. 2007; Anyanwu & Sheth 2003), they require a centralized knowledgebase, where all of the entities and relationships are available for analysis. However, this centralized knowledgebase is normally not available for drug discovery. Drug-related knowledge is dispersed in heterogeneous domains such as chemical and biological domains, which are owned by different organizations and distributed in different locations. It is impractical for the transmittal of terabytes of datasets over long distances in order to merge these data sources to a central location. Furthermore, some repositories may not be allowed to be merged either for legal reasons, for the risk of revealing business secrets, or for posing other social challenges. On the other hand, analyzing a local knowledgebase can only obtain limited knowledge that is constricted by spatial and temporal constraints. Apparently, existing centralized search technologies cannot be used in the drug discovery directly.

In this paper, we propose a discovery scheme that breaks the traditional barriers of the centralized scheme into the realm of decentralized and distributed strategies. It supports automatic discovery of semantic relationships between drug-related entities, such as drugs, proteins, genes, and diseases, over geographically distributed knowledgebases on an unprecedented scale. Our discovery scheme is fully decentralized and scalable. It not only efficiently addresses the issues of drug relation discovery, but also improves the traditional search and discovery of drug-related semantic knowledge, and thus making general drug-related knowledge sharing more effectives and efficient.

The rest of the paper is organized as follows. In the following section, we detail the background knowledge and related work. Then we introduce the concept of semantic relationship. Thereafter, we describe the design of the discovery system. We evaluate the proposed methods and show their effectiveness with a comprehensive set of simulations and case study. Concluding remarks are provided at last.

BACKGROUND AND RELATED WORK

Drug Discovery

In order to develop an effective drug, it is necessary to understand how disease and infection are controlled at the molecular and physiological levels. Basically, both the disease and its underlying cause need to be understood as well as possible. To understand the disease completely it is essential to understand as to which genes are affected by the disease as they in turn affect the proteins that they encode.

Since the proteins interact with each other in living cells, and affect the tissue in the areas in which the cells are located, and consequently, affect the patient on the whole. On gaining a complete understanding of the disease a target protein or a target gene needs to be identified. Once a proposed compound for the potential drug is identified, it is characterized and screened for its efficiency to deal with target protein or gene. In practice, for every 5000-10000 proposed compounds, only one gets approved as a drug for the treatment of diseases and use on humans.

Currently, natural products play a very significant role in designing a drug. The drugs could be plant derived, such as Belladonna, or from the microbes, such as streptomyces or marine invertebrates.

Despite advances in understanding of biological systems and in chemical synthesis techniques and combinatorial and cheminformatics, there is no increase in the number of new drugs. Currently, it still takes about 15 years for a new drug to appear on the market. The research and development cost of

each new molecular entity is approximately 1.8 billion US dollars.

Polypharmacology and Drug Repositioning

In order to reduce the time and cost in drug discovery, polypharmacology is receiving increasing attention. Polypharmacology is the study of interaction of drugs with multiple protein and gene targets (Hopkins 2008; Boran & Iyengar 2010).

When a drug have a polypharmacological action on multiple targets that may fall in the same pathway (chain of reactions associated with a particular metabolic process), it is able to interrupt the pathway at multiple points and exhibits high efficacy.

Drug repositioning is based on polypharmacology. It is the application of the existing drugs to new indications or new diseases (Sleigh & Barton 2010) since approved drugs could affect more than one protein target due to polypharmacological action, thereby affecting more than one disease. An example of drug repositioning is the cancer drug, such as Gleevac, can bind to multiple kinases (Druker et al. 1996). Other examples are Propiomazine (Largon) and Promazine (Sparine) which have 14 targets each (Yıldırım et al. 2007).

There are also certain diseases where a single gene or single protein may not be responsible for it. Hence, one drug is insufficient to treat the diseases. The drug-target pair is thus crucial for polypharmacology. There could be many-to-many relationships between the drug-target pairs.

Linked Data Projects Related to Drugs

Linked Data project uses the Web to connect related data that wasn't previously linked, or lowers the barriers to linking data currently linked using different methods. More specifically, Wikipedia defines Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF."

Many drug-related Linked Data projects exist. DailyMed publishes Linked Data of marketed drugs along with general background on the chemical structure of the compound and its therapeutic purpose, details on the compound's clinical pharmacology, indication and usage, contraindications, warnings, precautions, adverse reactions, overdosage, and patient counseling (http://dailymed.nlm.nih.gov/). DrugBank publishes Linked Data of almost 5000 FDA-approved small molecule and biotech drugs (Knox et al. 2011). It contains detailed information about drugs including chemical, pharmacological and pharmaceutical data; along with comprehensive drug target data such as sequence, structure, and pathway information. LinkedCT.org (Linked Data Source of Clinical Trials) contains roughly 25 million triples (the underlying structure of any expression in RDF, each consisting of a subject, a predicate and an object) as of April 2011, about 106,000 clinical trials, with more than 167,000 links to external sources such as DBpedia (http://dbpedia.org), DailyMed (http://dailymed.nlm. nih.gov/), DrugBank (Knox et al. 2011), and Bio2RDF.org's PubMed (http://www.ncbi.nlm. nih.gov/pubmed). The readers can refer to live stats page for up-to-date statistics about the entities and external links. SIDER publishes Linked Data of almost 1,000 marketed drugs and their adverse effects (http://sideeffects.embl.de/). The information is extracted from public documents and package inserts. Linking Open Drug Data (LODD) project brings the data sources together onto the Web of Linked Data and facilitates the integration of data (Samwald et al, 2011). All the data and the datasets used from different sources have been strongly linked and also linked to the other Linked data. LODD contains 8.4 million RDF triples and the data from the data sources about drugs, Chinese medicine, clinical trials, diseases and pharmaceutical companies have been linked together.

Our work focuson discovering hidden relationships between drugs and protein and gene targets over these linked datasets. We further extend the concept of linked data to a larger scale and more distributed ad hoc data

sets. Our goal is to discover useful information from dispersed knowledgebases to expedite drug discovery process based on polypharmacological action.

Discovery of Hidden Relationships

There have been many studies on semantic relation discovery using data mining approaches, such as association rules and clustering (Nagano et al. 2010; Ruiz-Casado et al. 2005, Syed et al. 2005; Jiang et al. 2007). However, their discovery mechanisms are normally based on the co-occurrence of the entities in documents, which is significantly different from our work. Our work focuses on semantic web datasets in which semantic entries are linked through real semantic links. The relationships are located by discovering the real semantic links but not co-occurring. Therefore, in our work, we can not only identify if entities are related, but also explain how these entities are connected.

The query supporting RDF semantic relationships were first proposed by K. Anyanwu and A. Sheth (Anyanwu & Sheth 2003). They define a semantic association as a complex relationship between two resources, and introduce a set of operators for querying semantic associations. Based on their work, several applications have appeared that use semantic relations (Meza et al. 2006; Sheth et al. 2005; Heim et al. 2010; Heim et al. 2010; Lehmann et al. 2007). Most of these applications assume a centralized dataset. Yet, researchers in (Perry et al. 2005) propose a method for computing semantic associations over a P2P network. The authors use a super-peer based query planning algorithm for ρ-path quires. In their proposed system, knowledgebases are stored at the peer level, while indexes are stored at the super-peer level. Each super-peer is responsible for a group of peers. A super-peer knows about all of the other super-peers in the network and can query them to determine the semantic paths. It is an effective approach, but the scalability is still an unsolved issue. A couple of questions still remain: (1) how to organize the peer group to reflect the semantic proximity, and (2) how super peers efficiently communicate. In (Li et al, in press), we proposed a fully decentralized approach to discover semantic relationships over large-scale networks to address these questions. In this work, we adopt the idea of searching in different abstraction and extend it with a specific discovery scheme for drug discovery.

REPRESENTING SEMANTIC RELATIONS

The RDF is a World Wide Web Consortium (W3C) recommendation for describing Web resources. The RDF provides a basic data model, such as the entity-relationship model for writing simple statements about Web objects. It can make statements about resources in the form of subject-predicate-object expressions, termed triples in the RDF terminology. The subject denotes the resource that has a URI. The predicate denotes traits or aspects of the resource and expresses a relationship between the subject and object. It is also identified by URIs. The object is the actual value that can either be a resource or a literal.

The RDF can also represent statements about resources as a directed labeled graph with typed edges and nodes. In this model, a directed edge labeled with a property name connects the subject to the object. For instance, the group of statements, “There is a drug identified by http://www.drugbank.ca/drugs/proguanil. It has a protein target http://www.uniprot.org/ DihydrofolateReductase, which is associated with a gene identified with

http://www.genecards.org/DHFR. The gene is related with disease http:// www.ncbi.nlm.nih.gov/ omim/Malaria” could be represented as the RDF graph that is depicted in Figure 1.

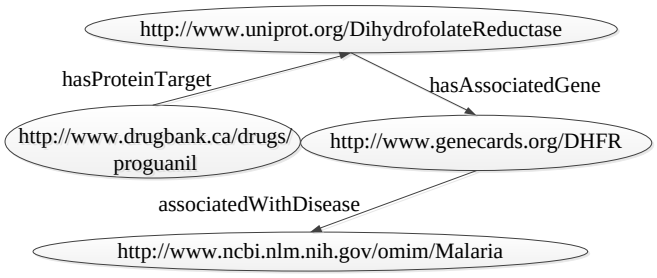


Figure 1. Example of Part of a RDF Graph.

As depicted in Figure 1, paths of the RDF graph represent semantic relationships among the participating resources (entities), explicitly or implicitly. We adopted the ρ-path query reported in (Anyanwu& Sheth, 2003) as a way of expressing semantic associations between entities in the RDF graph. A path $\rho=e_1,p_1,e_2,p_2,e_3,\ldots,e_{n-1},p_{n-1},e_n$ is a sequence of RDF statements, where each triple e_i, p_i, e_{i+1}, represents a single statement in which p_i is the predicate and one of e_i or e_{i+1} is the subject and the other is the object. While a ρ-path has been defined as a directed path in (Anyanwu& Sheth, 2003) , we treat paths as undirected in our paper because as long as two entities are connected by undirected paths, they are semantically associated, even though there is no directed path connecting them. For example, in Figure 1, drug proguanil and gene DHFR are related although there are no directed paths between them. We can consider directions of the edges after undirected paths have been located. Therefore, in this paper, two resources *x* and *y* are said to be ρ-path associated if there exists an undirected path ρ of length $n>0$ between them.

DRUG RELATION DISCOVERY OVER DISPERSED KNOWLEDGEBASES

As mentioned, drug discovery involves data over many diverse domains such as chemical and biological domains. Data is created and stored by different organizations or individuals that are geographically distributed. To discover complex relationships over the dispersed data sites, new challenges have emerged. First, it is difficult to achieve global optimum of semantic association discovery with dispersed local operations because of the lack of a global view or unified understanding of the distributed semantic data; second, an efficient discovery protocol is required to forward search requests between knowledgebases and later to gather the search results because the relationships between two entities may span over multiple distributed knowledgebases; and third, it would be difficult to achieve scalability and low latency in large-scale distributed systems due to the complexity of semantic-relation queries. We propose a novel hierarchical knowledge abstraction and an efficient discovery protocol to address these challenges. Our goal is to provide a flexible and effective drug relation discovery framework. In this work, we focus on Semantic Web data represented in RDF format. For data represented in other format, there are existing technologies to convert data to RDF format (Li & Su, 2009). We present our idea of hierarchical knowledge abstraction and efficient discovery protocol in the following subsections.

Overview of Hierarchical Knowledge Abstraction

As presented above, drug discovery can be converted to a path discovery problem in a linked RDF graph. However, path discovery is much more difficult than entity discovery, because it needs to locate not only the entities but also all paths connecting them. Our solution is inspired by the strategy of Inter-Domain Routing in the Internet. Considering Internet routing that is scalable with millions of nodes, our semantic graph is very similar to Internet in that both are large-scale including millions of nodes and edges, and both are distributed without a global view at any individual node. Therefore, we believe that we can adopt a similar abstraction strategy for our semantic path-finding.

Given enormous links of the Internet, it is too expensive, if not impossible at all, to compute the route between all the nodes. Instead of working at such a low level of details, Internet routing is planned at the Autonomous System (AS) level. Autonomous System corresponds to an administrative domain. Once the path reaches an AS border, the best route is computed from AS to AS. The Border Gateway Protocol (BGP) is the core routing protocol being used for AS level routing. We adapt a similar idea of BGP for semantic relationship discovery.

As shown in Figure 2, instead of starting from millions of semantic entities and relationships at the lower level, we consider each knowledgebase containing multiple entities and relations as an abstract unit (as an AS in the Internet). Each site hosts an individual knowledgebase. Based on our previous work on semantics-based topology adaptation (Li, 2010),

we can create links between distributed knowledgebases. Next, through URL links and ontology-mappings, we connect knowledgebases to form a graph. In the graph, we treat knowledgebases as black boxes and ignore the detailed semantic entities and their relations. The graph, called the semantic graph, acts as the blueprint of our search graph. Based on the semantic graph, the semantic path discovery problem is analogous to the route discovery of Internet. This abstraction dramatically reduces the size of a potentially huge search space.

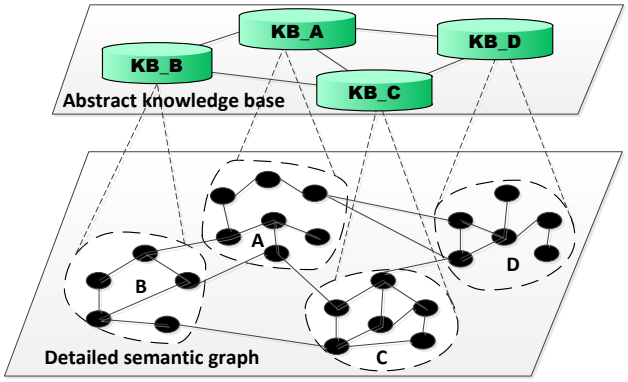


Figure 2. Hierarchical Structure of Semantic Path Discovery of Drug Relation Discovery

Preparation

To link the dispersed knowledgebases to form a connected semantic graph, we proposed a semantics-based topology adaptation scheme to connect the knowledgebases containing similar or related semantic properties and facilitate the establishment of semantic mappings or links. The foundation of this scheme is a metric that measures knowledgebase's semantic similarities. We have done work on measuring the semantic similarities between ontologies (Li & Khan, 2009) and support multiple ontologies and improve the accuracy by integrating factors, such as the depth of a node in the ontology hierarchy and the type of links. After semantically related knowledgebases have been located, mapping or linking can be established between these knowledgebases. In some scenarios related knowledgebases can be merged. The afore-mentioned process can be achieved by P2P-based neighbor discovery and overlay formation. We encourage the readers to find more details about our proposed topology adaptation scheme by referring to our previous studies in (Li, 2010; Li & Vuong, 2008). After crating links/mappings between distributed knowledgebases, we can create a distributed semantic graph.

As a semantic graph is constructed, the path finding problem is reduced to two steps: firstly, locate the source and destination semantic entities, and secondly, search for paths from the source knowledgebase containing the source entity to the knowledgebase containing the goal entity, at the knowledge-base level It turns out a much faster search. To efficiently locate the source and destination semantic entities, we adopt a distributed hash table (DHT)-based overlay (Rowstron & Druschel 2001; Stoica et al. 2001; Ratnasamy et al. 2001) to index the semantic graph, with which semantic entities can be efficiently located.

As mentioned, entities are subjects and objects in RDF triples. Triples in distributed knowledgebase that share common entities (i.e., the same subjects and/or objects) should be indexed together in one of the distributed knowledgebases, where they can be located later. The challenge in this scenario lies in assigning "index rendezvous points" for entities. To avoid the centralized bottleneck, we use a DHT overlay to provide decentralized and scalable rendezvous for RDF triple entities. Each triple is sent to two rendezvous knowledgebases for its subject and object respectively, which ensures that triples with common subjects and/or objects will be co-located. Unlike RDFPeer's data indexing (Cai & Frank 2004), we do not index predicates (i.e., edges of the semantic graph), because normally we only need to locate entities of the semantic graph not the edges. We store each triple twice by applying a hash function to its subject and object. The DHT indexing guarantees the entities can be located within log(N) hops, where N is number of knowledgebases in the Semantic Web.

Semantic Relation Discovery

When the semantic graph is created, and both source and goal entities are located, the next step is to locate paths between the source and goal

entities. We explain the details in finding the path in this section.

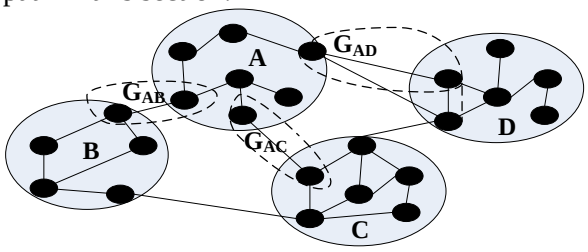


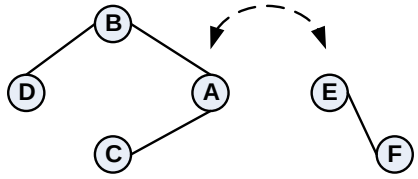
Figure 3. An Example of Connection of Knowledgebases and Their Gateways.

To find a *k*-hop limited semantic path, we need to count the path length. Therefore, each knowledgebase records a set of distances (in terms of semantic edges in the semantic graph) before they can be treated as a black box. The distance that matters is the set of shortest semantic hops between knowledgebases. We illustrated our path location idea in Figure 3.

As shown in Figure 3, Knowledgebase *A* is linked to knowledgebases *B*, *C*, and *D* though ontology mappings. The entities in *A* that are linked or mapped to other knowledgebases are called *gateway* nodes of *A*. For example, G_{AB}, G_{AC}, G_{AD} are knowledgebase *A*'s gateways to neighboring knowledgebases. If two knowledgebases *A* and *B* have more than one gateway nodes, we pick the one that contributes to the shortest semantic path. *A* records the shortest distances between all of its gateways. , we can see that from *A*'s local knowledgebase *dist(*G_{AB}, G_{AC} *)=2*, (i.e., the path cost of from *B* to *C* via *A* is *2*), *dist (*G_{AB}, G_{AD}*)=3, dist (*G_{AC}, G_{AD}*)=5*.

To locate paths between knowledgebases, we propose the Semantic Border Gateway Protocol (SBGP). This routing protocol was inspired by the BGP routing, but the protocol itself is different from BGP: BGP only needs to locate one shortest path from source to destination, while our SBGP has to locate multiple paths to discover multiple relationships between entities. The cost computation of SBGP is also different from BGP's. Unlike BGP, SBGP does not consider the cost of an edge; instead, it considers the cost of the paths between gateway nodes.

The SBGP is a path-vector protocol. In particular, for each knowledgebase *j*, knowledgebase *i* stores the knowledgebase paths of the lowest costs (maybe more than one) from *i* to *j*; in this vector, knowledgebases are identified by their knowledgebase ID. SBGP's route computation is similar to all path-vector routing protocols. Each knowledgebase sends its routing table to its neighbors, and each knowledgebase can then, based on this information, compute its own routing information.



Routing table format

Direct Neighbors: B, C		
Destination	Paths	Cost

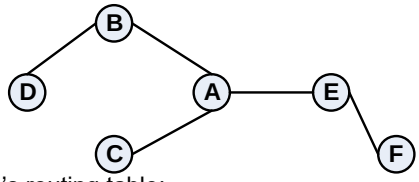
A's routing table:

Direct Neighbors: B, C		
D	B	dist($G_{(A, B)}$, $G_{(B, D)}$)

E's routing table:

Direct Neighbors: F

(a) Initial Routing Tables



A's routing table:

Direct Neighbors: B, C, E		
D	B	dist($G_{(A, B)}$, $G_{(B, D)}$)
F	E	dist($G_{(A, E)}$, $G_{(E, F)}$)

E's routing table:

Direct Neighbors: A, F		
B	A	dist($G_{(E, A)}$, $G_{(A, B)}$)
C	A	dist($G_{(E,A)}$, $G_{(A, C)}$)
D	A, B	dist($G_{(A,B)}$,$G_{(B,D)}$)+dist($G_{(B,A)}$+$G_{(A, E)}$)

(b)Updated Routing Tables.

Figure 4. Routing table Updating after Establishing a New Link

Figure 4 illustrates the SBGP updating process when a new path between node *A* and *E* is found. Figure 4 (a) shows the routing tables of

A and *E* before the connection is established. The routing tables record the path (nodes traversed to reach a destination node) and cost information to other nodes. The cost information is the distance between gateways. For example, in *A*'s routing table, the distance between *A* and *D* is the distance between two gateways: G_{BA} and G_{BD} . It is reported by *A*'s neighbor *B*. *A* does not record the cost to direct neighbors, since no gateway is used. *A* and *E* then exchange their routing tables, and compute their updated routing tables. As shown in Figure 4 (b), *E*'s new routing table now includes all nodes in *A*'s routing table and *A*'s direct neighbors. The cost is updated by adding the distance of gateway G_{AE} to *A*'s other boundaries in the path. For example, in *A*'s routing table, destination *D* can be reached with a cost of $dist(G_{BA}, G_{BD})$. Then *E* can reach *D* through *A*'s boundaries G_{AE} and G_{AB}. Therefore, in *E*'s routing table the cost is $dist(G_{AB},G_{BD})+dist(G_{AE},G_{A,B})$. If the cost is greater than the predefined *k*-hop limit, the path is ignored. In such a way, nodes can construct and update their routing tables. According to the routing table, a query looking for a destination node can be forwarded between knowledgebases in the routing path.

Semantic Relation Retrieval

Given a query of discovering the relationships between entity *A* and entity *B*, the system first locates the two knowledgebases, say, P_A, P_B, in charge of these two entities. Since there may exist multiple such knowledgebases, we consider each individual possible combination of them.

With SBGP, the system can find the path connecting P_A and P_B. To retrieve semantic relations, the system has to go back to the detailed semantic graph (at the lower level shown in Figure 2) in two steps: (1) find the semantic path from entity *A* to one gateway node in P_A which is on the path to P_B. Similarly, find the semantic path from entity B to one gateway node in P_B which is on the path to P_A, and (2) retrieve the path from P_A to P_B. This is a process of finding paths between all boundaries in the path, which can be achieved by SBGP. In such a way, drug-related semantic relations and associations can be obtained.

In sum, with hierarchical knowledge abstraction and drug relation discovery protocol, one can find relevant drug information and determine their importance in drug discovery therefore expediting drug discovery process tremendously. In later sections, we further illustrate the significance of our work via examples.

EVALUATION AND CASE STUDIES

Setting up Simulation Experiments

We tested the performance of the proposed mechanism with publicly available data sources on the Internet for genes, genetic information, pathways, proteins, drugs and diseases. The information contained on these data sources may be similar and overlapping for the same entity. Linking these datasets produces a network of linked drugs, proteins, genes and diseases.

In our experiments, drug data was obtained from the most widely used drugs database, Drugbank (http://hwww.drugbank.ca). Protein data was obtained from UniProt (http://www.uniprot.org) which is a widely used database for proteins and protein sequences. The proteins acted upon by the drugs in DrugBank are identified by the UniProt ID, which links them to the protein database. While Gene data was obtained from GeneCards (http://www.genecards.org) which contains information about all known and predicted human genes and information on diseases that was linked in OMIM database (http://www.omim.org). All the data were parsed to the RDF triple format.

Our goal here is to discover complex semantic relationships between drugs and diseases, as well as the target proteins and the genes that are affected by those proteins. If the relationship exists through intermediate entities it can be expressed as a chained triple. Every element in a RDF triple is an URI. When two resources have the same URI they are said to be identical and all data for identical resources is merged.

To simulate a distributed environment where dispersed knowledge can be shared, we created a network simulator with 1024

computers (nodes). Thereafter, we divided the parsed RDF triples into smaller parts (sub-knowledgebases), and deployed each sub-knowledgebase on one of the nodes within the simulated network. In order to model the inter-knowledgebase ontology mappings and links, we used the pre-existing links connecting entities that were located in different sub-knowledgebases after knowledgebase decomposition. Therefore, we converted the knowledgebases to a set of smaller knowledgebases and distributed them in the network.

We used BRITE (Medina et al. 2001) to generate network topologies. In particular, our simulator used a parser to parse the output file exported by BRITE and create the targeted topology. Since Power-law distributions have been observed in the Internet and also Semantic Web (Ding & Finin 2006; Theoharis et al. 2008), we incorporate power-law in the topology generation by using Waxman and BarábasiAlbert models. The knowledgebase distribution also follows the power-law distribution. In particular, we used a Zipf distribution (Zipf, 1949) to model the distribution of knowledgebases.

Queries were generated by providing two semantic entities, the source and the destination. The source was picked randomly from a dataset's knowledgebase. We also randomly picked a semantic path starting from the source entity with a path of length limit (1-4), which led to a semantic entity that is labeled as the destination. The path may also cross multiple data sets (i.e., knowledgebases). Therefore, the query was to find all of the paths of length limited to k, between these two semantic entities. Each experiment was repeated ten times with different random seeds.

Performance Evaluation

First, we evaluated the completeness of the proposed search algorithm. Table 1 illustrates the recall rate of the proposed distributed discovery scheme. The recall rate is defined as the fraction of the successfully retrieved semantic relations that are relevant to the query. The results illustrate that our approach is complete, i.e., it can find all of the related relationships. As it is difficult to determine which relationships are more important for non-professionals, our work focuses on locating as many relationships as possible without considering the precision.

TABLE I. RECALL OF THE DISCOVERY APPRAOCH

Path-length	1	2	3	4
Recall Rate	100%	100%	100%	100%

Next, we studied the effectiveness of the hierarchy in the discovery scheme. We believe the hierarchy improves the scalability of the system. To test this hypothesis, we compare the number of semantic entities traversed when searching at different levels of hierarchy. Figure 5 demonstrates the effectiveness of the hierarchy strategy. As shown in Figure 5, our discovery protocol significantly reduces semantic entities traversed. As a result, the network overhead and computation overhead related with traversing can be reduced. This efficiency is achieved because using two levels of hierarchy and applying SBGP routing technologies can reduce the search space and consequently improving the scalability.

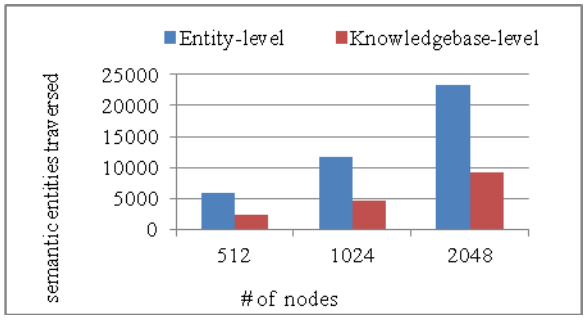


Figure 5. Discovery at Different Levels of Hierarchy

To further evaluate the scalability and efficiency of the proposed discovery protocol SBGP, we compared its performance with breadth-first search (BFS)-based discovery scheme (Heim et al. 2010) in terms of average bandwidth and the number of semantic entities traversed. BFS-based discovery (DiMasi et al. 2003, Samwald et al. 2011) was used as the benchmark scheme because of its simplicity and popularity (Krishna et al. 2011). In the simulations, a random set of nodes periodically

issue discovery queries. The queries were to find the relationships between a local entity of the querying node and another entity randomly selected from a node in the network. In each simulation time slice (1 second), the query probability for a given node was set to 0.1%. The length limit of the relations k was set to 4. The results are shown in Figures 6 and 7. As can be seen from Figure 6 and Figure 7, SBGP significantly outperforms BFS-based discovery in terms of the bandwidth consumed and the number of semantic entities traversed. (Note that Figure 7 was plotted on a logarithmic scale to better illustrate the significance of SBGP.)

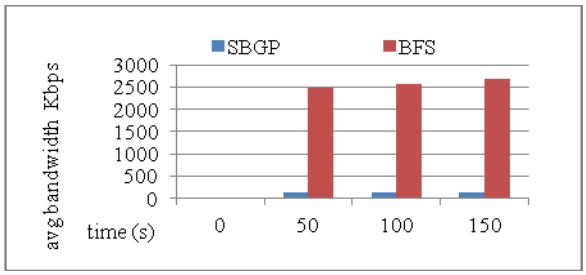


Figure 6. Comparison of SBGP and BFS in bandwidth comsumption.

Figure 7. Comparison of SBGP and BFS in Semantic Entities Traversed.

Case Study

In this section, we present three examples to illustrate the significance of the discovery scheme.

Example 1: *Different relations between drug and disease.*

Figure 8 shows different paths from anti-malarial drugs and malaria. For example, Chloroquine, a commonly used treatment of malaria, is related the decease through a chain of protein target, gene, and related disease. While Proguanil relates malaria with a different path through gene DHFR, which treats the disease by acting on the metabolism of the parasite, Plasmodium falciparum and Plasmodium vivax, known to cause malaria. Some other antimalarial drugs, such as Chloroquine, Halofantrine, Mefloquine, Primaquine, Amodiaquine, and Quinine, contain Ferriprotoporphyrin IX. This compound is known to form cytotoxic complexes with the antimalarial drugs that cause plasmodial membrane damage.

Example 2: *Finding relations to demonstrate Polypharmacology of drugs.*

Polypharmacological properties of the drugs are demonstrated by different paths originating from one disease to a drug and vice-versa with intermediate genes and protein targets. If two drugs have at least two same targets, they will show the polypharmacological properties. Alzheimer's disease is generally treated by inhibiting the enzymes Acetylcholinesterase and Cholinesterase. The gene that codes Acetylcholinesterase is ACHE which is a known factor in Alzheimer's disease and the gene that codes Cholinesterase is BCHE. There are a number of approved drugs in drugbank that are known to act as the inhibitors of Acetylcholinesterase as well as Cholinesterase. Of the 16 approved drugs in drugbank, 4 drugs, namely, Tacrine, Rivastigmine, Galantamine and Choline contained both Acetylcholinesterase as well as Cholinesterase and one Donepezil contained Acetylcholinesterase. Tacrine is given as an example in Figure 9.

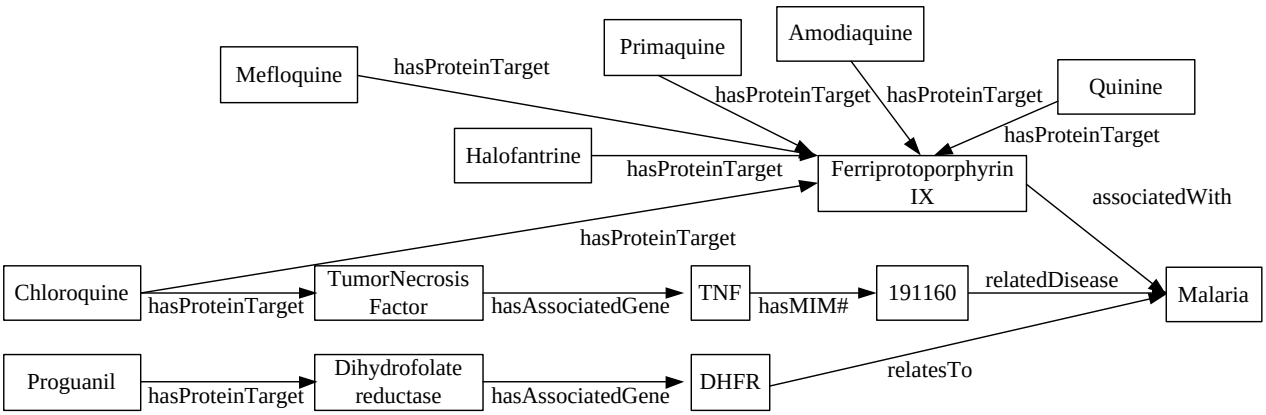


Figure 8. Example Relations between Anti-malarial Drugs and Malaria.

The experimental drugs such as Methylphosphinic Acid, 2-(N-Morpholino)-Ethanesulfonic Acid or Fucose, could be the potential new drugs and may show the polypharmacological properties. Methylphosphinic Acid has both Acetylcholinesterase and Cholinesterase as the principal components. Hence, it could be used for the treatment of Alzheimer's Disease whereas the 2-(N-Morpholino)-Ethanesulfonic Acid or Fucose act on as many as 30 proteins. They should be studied for more than one indication as they are being developed for treating more than one disease including Alzheimer's Disease, Myasthenia Gravis and Glaucoma.

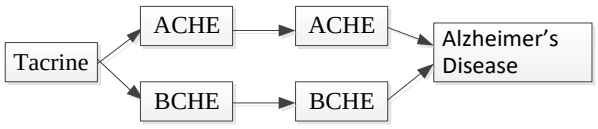


Figure 9. Multiple-paths from Tacrine to Alzheimer's disease

***Example 3**: Finding complex relationships among drug, gene, and disease.*

Since at least one gene is involved in all the diseases, we attempted to find the relationship between drug, gene and the disease. To understand the relationship, we give a brief overview of protein and gene.

Protein is a chain of polypeptides linked together. A polypeptide is a linear chain of amino acids linked together by means of a peptide (- H – N – CO -) bond. Enzymes are proteins that catalyze a chemical reaction in the human body. A polypeptide chain in a protein is coded by a particular gene. A gene is made up of a genetic code. The sequence of a genetic code results in the sequence of amino acids in a polypeptide. In other words each gene codes a polypeptide which is contained in a protein. The same polypeptide may be contained in more than one protein. Different polypeptides contained in one protein may be synthesized by different genes. Hence, a many-to-many relationship exists between proteins and genes where as a one-to-one relationship exists between a polypeptide and a gene. A gene may be involved in causing a disease by synthesizing a protein/enzyme. A drug acts as an inhibitor for the action of a protein to cure a disease.

We assumed that if a gene causes a disease and a drug cures the disease, the same drug can be used to treat another disease that are involved the same gene. Based on this assumption, we studied the drug-gene-disease relationship for the gene ACHE which is a known factor in Alzheimer's disease. There are a number of approved drugs in DrugBank that are known to act as the inhibitors of Acetylcholinesterase. In our study we considered three drugs. They are Tacrine used primarily for the treatment of Alzheimer's disease, Pyridostigmine used for the treatment of Myasthenia Gravis, and Demecarium used for the treatment of Glaucoma. As shown in Figure 10, many other drugs could be studied for their potential use for the treatment of Alzheimer's disease.

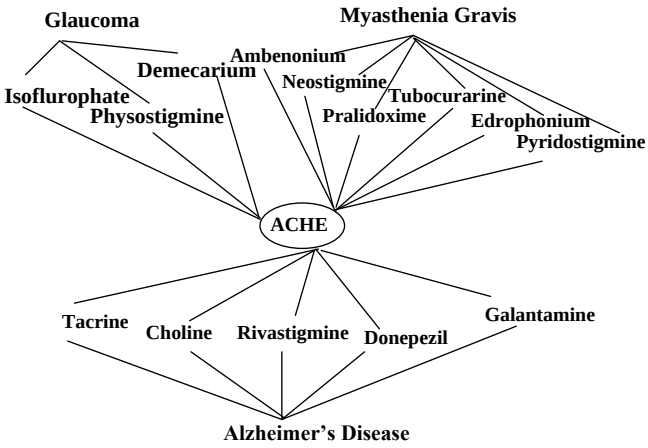


Figure 10.The drug-disease-gene relationship for gene ACHE

CONCLUSIONS

A large number of diseases have been known to affect humans every day. Computational methods have been proposed to find the disease genes and protein interaction. However, few methods have been proposed to facilitate drug development. On the other hand, large amount of biological and chemical data is present on the Internet in public databases. It will be impossible for humans to digest all of the content on the web without the assistance of effective knowledge discovery tools. As the Internet expands to contain more and more data, information definition and searching is becoming increasingly important.

In this paper, we take advantage of the recent Semantic Web technologies and integrate them with distributed routing algorithms to address drug discovery problem. We developed a novel framework consisting of a hierarchical knowledge abstraction and an efficient discovery protocol to speed up drug discovery. By extracting drug-related semantic metadata of web resources, our discovery scheme can capture the semantic association of drugs. In our discovery framework, complex semantic relations, identified by the chaining of the ontological triples in the metadata will allow us to identify more complex relationships among distributed drug-related data sources; these relationships were not known to the public previously. We have evaluated the proposed discovery framework with real web data. The experimental results revealed the scalability and efficiency of our proposed discovery framework.

REFERENCES

Aronson, J. K., "Old drugs – new uses", Br J Clin Pharmacol. 64(5), 563–565, 2007.

Ashburn, T. T. and Thor, K. B., "Drug repositioning: Identifying and developing new uses for existing drugs", Nature Review: Drug Discovery, 3, 673-683, 2004.

Anyanwu, K., Sheth, A. "ρ-Queries: Enabling Querying for Semantic Associations on the Semantic Web." The Twelfth International World Wide Web Conference, Budapest, Hungary. 2003.

Boolell, M., Allen, M.J., Ballard, S.A., Gepi-Attee, S., Muirhead, G.J., Naylor, A.M., Osterloh, I.H., Gingell, C., "Sildenafil: an orally active type 5 cyclic GMP-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction". Int J Impot Res 8 (2), 47–52, 1996.

Boran, A.D., Iyengar, W R., "Systems approaches to polypharmacology and drug discovery ", Curr Opin Drug Discov Devel., 13(3): 297–309, 2010.

Cai, M., Frank, M., "RDFPeers: A scalable distributed RDF repository based on a structured peer-to-peer network", in proc of WWW conference, May 2004.

Chong, C.R., Sullivan, D.J., "New uses for old drugs". Nature. 448,645–6, 2007.

Collier, R., "Drug development cost estimates hard to swallow", CMAJ, 180(3), 279–280, 2009.

Chan, A.T., Ogino, S., Fuchs, C.S., "Aspirin and the risk of colorectal cancer in relation to the expression of COX-2". N Engl J Med.,356,2131–42, 2007.

Clive, S., Gardiner, J., Leonard, R.C., "Miltefosine as a topical treatment for cutaneous metastases in breast carcinoma",Cancer Chemother Pharmacol., 44(Suppl), S29–30, 1999.

Ding, L. and Finin, T., "Characterizing the Semantic Web on the web". In Proceedings of the International Semantic Web Conference (ISWC). 2006.

Durk, H.A.,"Maintenance therapy for multiple myeloma with particular emphasis on thalidomide". Onkologie,.29,582–90, 2006.

Druker, B.J., Tamura, S., Buchdunger, E., Ohno, S., Segal, G.M., Fanning, S., Zimmermann, J., Lydon, N.B., "Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells", Nat Med. 2(5):561-6, 1996.

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., Sachs, J., "Swoogle: a search and metadata engine for the semantic web". Proc. 13th ACM Conf. on Information and Knowledge Management. 2004.

DiMasi, J.A., Hansen, R.W., Grabowski, H.G., "The Price of Innovation: New Estimates of Drug Development Costs," Journal of Health Economics, 22, 151-185, 2003.

DiMasi, J.A., Hansen, R.W., Grabowski, H. G., Lasagna, L., "Cost of innovation in the pharmaceutical industry", Journal of Health Economics, 10, 107-142, 1991.

McCool, R.R., Miller, E., "Semantic search". In: Proc. of the 12th Intl. Conf. on World Wide Web. 2003.

Hamosh, A., Scott, A., Amberger, J., Bocchini, C., McKusick, V., "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders", Nucleic Acids Research: Database issue", 33, D514–D517, 2004.

Heim, P., Lohmann, S., Stegemann, T., "Interactive Relationship Discovery via the Semantic Web" Proceedings of the 7th Extended Semantic Web Conference (ESWC2010), Part I, LNCS 6088, Berlin, Heidelberg: Springer; pp. 303–317, 2010.

Jiang, T., Tan, A.H., Wang, K., "Mining Generalized Associations of Semantic Relations from Textual Web Content", IEEE Transactions on Knowledge and Data Engineering. Volume 19 Issue 2, February 2007.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., & Hirakawa, M., "KEGG for representation and analysis of molecular networks involving diseases and drugs", Nucleic Acids Res., 38, D355-D360, 2010.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M., "From genomics to chemical genomics: new developments in KEGG", Nucleic Acids Res. 34, D354-357, 2006.

Kanehisa, M. & Goto, S., "Kyoto Encyclopedia of Genes and Genomes", Nucleic Acids Res. 28, 27-30, 2000.

Knox, C., Law,V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C. & Wishart, D. S. "DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs", Nucleic Acids Research, Database issue, 39, D1035–D1041, 2011

Krishna, V., Ranga Suri, N. N. R. and Athithan, G., "A Comparative Survey of Algorithms for Frequent Subgraph Discovery", Current Science, Vol. 100, No. 2, 2011.

Heim, P., Lohmann, S., Stegemann, T., "Interactive Relationship Discovery via the Semantic Web" Proceedings of the 7th Extended Semantic Web Conference (ESWC2010), Part I, LNCS 6088, Berlin, Heidelberg: Springer; pp. 303–317, 2010.

Heim, P., Lohmann, S., Stegemann, T., "Interactive Relationship Discovery via the Semantic Web". In Proceedings of the 7th Extended Semantic Web Conference (ESWC2010), Heraklion, Greece. 30 May - 03 Jun, 2010.

Hopkins, A.L.,"Network Pharmacology: The Next Paradigm in Drug Discovery", Nature Chemical Biology, 4,(11), 2008.

Jeffreys , D., "Aspirin The Remarkable Story of a Wonder Drug", New York & London, Bloomsbury, Diane Publishing Company, 2004.

Krumholz, H.M., Radford, M.J., Ellerbeck, E.F., Hennen, J., Meehan, T.P., Petrillo, M., Wang, Y., Kresowik, T.F., Jencks, S.F., "Aspirin in the

Treatment of Acute Myocardial Infarction in Elderly Medicare Beneficiaries : Patterns of Use and Outcomes". Circulation 92 (10): 2841–2847, 1995.

Lehmann, J., Schuppel, J. and Auer, S., Discovering Unknown Connections – the DBpedia Relationship Finder. In Proc. of CSSW, 2007.

Li, J. Wang, H., Khan, S., Li, Q., Zomaya, A.J., "A Fully Distributed Scheme for Discovery of Semantic Relationships," IEEE Transactions on Services Computing.(Forthcoming)

Li, J. and Su, Y., "A Framework for Efficient Query Answering on Semantically Heterogeneous Grids", the International Journal of Computational Science, 2009; 3(4).

Li, J., "Grid Resource Discovery Based on Semantically Linked Virtual Organizations", Journal of Future Generation Computer Systems - Elsevier, Volume 26, issue 3, 2010.

Li, J. and Khan, S. U., "MobiSN: Semantics-based Mobile Ad Hoc Social Network Framework," in IEEE Global Communications Conference (Globecom 2009), Honolulu, HI, USA, December 2009.

Li, J. and Vuong, S., "SOON: A Scalable Self-Organized Overlay Network for Distributed Information Retrieval", in 19th IFIP/IEEE International Workshop on Distributed Systems (DSOM), pp.1-13, 2008.

Masia, N., "Cost of Developing a new Drug". , US Dept of state publication – Focus on Intellectual Property rights, <http://www.america.gov/st/econ-english/2008/April/20080429230904myleen 0.5233981.html> , 2008.

Meza, B.A., Nagarajan, M., Ramakrishnan, C., Ding, L., Kolari, P., Sheth, A. P., Arpinar, I. B., Joshi, A., Finin, T., "Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection", Proceedings of the 15th international conference on World Wide Web, May 23-26, 2006.

Medina, A., Lakhina, A., Matta, I., Byers, J., "BRITE: An Approach to Universal Topology Generation", Proceedings of the Ninth International Symposium in Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS'01), p.346, August 15-18, 2001.

Nagano, S., Inaba, M., and Kawamura, T., "Extracting Semantic Relations for Mining of Social Data", The 3rd international workshop Social Data on the Web, 2010.

Nolin, M.A., Ansell, P., Belleau, F., Idehen, K., Rigault, P., Tourigny, N., Roe,P., Hogan, J.M., Dumontier, M., "Bio2RDF network of linked data", Semantic Web Challenge; International Semantic Web Conference(ISWC 2008) Karlsruhe, Germany, 2008.

Perry, M., Janik, M., Ramakrishnan, C., Arpinar, B. and Sheth, A., "Peer-to-Peer Discovery of Semantic Associations." In Workshop on P2P Knowledge Management, pages 1–12, 2005.

Pedersen, T., Patwardhan, S., Michelizzi, J., "WordNet: Similarity-Measuring the Relatedness of Concepts," in 19th National Conference on Artifical Intelligence (AAAI), 2004.

Rada, R., Mili, H., Bicknell, E., Blettner,M., "Development and Application of a Metric on Semantic Nets," IEEE Transaction on Systems, Man, and Cybernetics, vol. 19, no. 1, pp. 17-30, 1989.

Ratnasamy, S., Francis, P., Handley, M., Karp, R., and Shenker, S., "A Scalable Content-Addressable Network," In Proceedings of the ACM SIGCOMM, August, pp. 161-172. 2001.

Rowstron, A. and Druschel, P., "Pastry: Scalable, Distributed Object Location and Routing for Large-Scale Peer-to-Peer Systems," In Proceedings of the IFIP/ACM International Conference on Distributed Systems Platforms, Middleware, 2001.

Ruiz-Casado, M., Alfonseca, E., and Castells, P., "Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia". In Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems, NLDB'05, pp. 67–79, Alicante, Spain, June 15–17, 2005.

Samwald, M., Jentzsch, A., Bouton, C., Kallesøe, C.S., Willighagen, E., Janos Hajagos, J., Marshall, M.S., Prud'hommeaux, E., Hassenzadeh, O., Pichler, E.& Stephens, S "Linked open drug data for pharmaceutical research and development", J Cheminform., 3(19), 2011.

Sheth, A., Aleman-Meza, B., Arpinar, I. B., Halaschek, C., Ramakrishnan, C., Bertram, C., Warke, Y., Avant, D., Arpinar, F. S., Anyanwu, K., Kochut, K., "Semantic Association Identification and Knowledge Discovery for National Security Applications", Journal of Database Management, 16 (1), pp. 33--53, Jan-Mar 2005.

Stoica, I., Morrism R., Karger, D., Kaashoek, M. F., and Balakrishnan, H,. "Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications," In Proceedings of the ACM SIGCOMM, pp. 149-160. August 2001.

Sleigh, S.H., Barton, C.L., "Repurposing Strategies for Therapeutics", Pharm Med, 24 (3), 151-159, 2010.

Syed, Z., Viegas, E., and Parastatidis, S., "Automatic Discovery of Semantic Relations using MindNet" Proceedings of the Seventh International Conference on Language Resources and Evaluation, 2010.

Sundar, S., Rosenkaimer, F., Makharia, M.K., Goyal, A.K., Mandal, A.K., Voss, A., Hilgard, P., Murray, H.W., "Trial of oral miltefosine for visceral leishmaniasis". Lancet.,352, 1821–3, 1998.

Theoharis, Y. et al., "On Graph Features of Semantic Web Schemas," IEEE Trans. Knowl. Data Eng., vol. 20, no. 5, pp. 692–702, 2008.

Verma, U., Sharma, R., Gupta, P., Kapoor, B., Bano, G., Sawhney, V., "New uses for old drugs: Novel therapeutic options", Indian J Pharmacol, 37(5), 279-287, 2005.

Webb, D.J., Freestone, S., Allen, M.J., Muirhead, G.J., (March 4,). "Sildenafil citrate and blood-pressure-lowering drugs: results of drug interaction studies with an organic nitrate and a calcium antagonist". Am. J. Cardiol. 83 (5A), 21C–28C, 1999.

Yıldırım, M.A., Goh, K.I., Cusick,M.E., Barabási, A.L. & Vidal,M. "Drug–Target Network", Nature Biotechnology, 25, (10), 1119, 2007.

Zipf, G. K., "Human Behaviour and the Principle of Least-Effort", Addison-Wesley, Cambridge MA, 1949.