A Ranking-Based Approach to Discover Semantic Associations Between Linked Data

María-Esther Vidal[1] and Louiqa Rashid[2] and Luis Ibáñez[1] and Jean Carlo Rivera[1] and Héctor Rodríguez[1] and Edna Ruckhaus[1]

[1] Universidad Simón Bolívar
Caracas, Venezuela
{mvidal,libanez,jrivera,hector,ruckhaus}@ldc.usb.ve
[2] University of Maryland
louiqa@umiacs.umd.edu

Abstract. Under the umbrella of the Semantic Web, Linked Data projects have the potential to discover links between datasets and make available a large number of semantically inter-connected data. Particularly, Health Care and Life Sciences have taken advantage of this research area, and publicly hyper-connected data about disorders and disease genes, drugs and clinical trials, are accessible on the Web. In addition, existing health care domain ontologies are usually comprised of large sets of facts, which have been used to annotate scientific data. For instance, annotations of controlled vocabularies such as MeSH or UMLS, describe the topics treated in PubMed publications, and these annotations have been successfully used to discover associations between drugs and diseases in the context of the Literature-Based Discovery area. However, given the size of the linked datasets, users have to spend uncountable hours or days, to traverse the links before identifying a new discovery. In this paper we provide an authority-flow based ranking technique that is able to assign high scores to terms that correspond to potential novel discoveries, and to efficiently identify these highly scored terms. We propose a graph-sampling method that models linked data as a Bayesian network and implements a Direct Sampling reasoning algorithm to approximate the ranking scores of the network. An initial experimental study reveals that our ranking techniques are able to reproduce state-of-the-art discoveries; additionally, the sampling-based approach is able to reduce the exact solution evaluation time.

1 Introduction

During the last decade, emerging technologies such as the Semantic Web, the Semantic Grid, Linked Data projects, and affordable computation and network access, have made available a great number of publicly inter-connected data sources. Life science is a good example of this phenomenon. This domain constantly evolves, and has generated publicly available information resources and services whose number and size, have dramatically increased during the last years. For example, the amount of gene expression data has grown exponentially, and most of the biomedical sources that publish this information have been gaining data at a rate of 300 % per year. The same trend is observed in biomedical literature where the two largest interconnected bibliographic databases in biomedicine, PubMed and BIOISIS, illustrate the extremely large size of

the scientific literature today. PubMed publishes at least 16 million references to journal articles, while BIOSIS makes available more than 18 million abstracts.

On the other hand, a great number of ontologies and controlled vocabularies have become available under the umbrella of the Semantic Web. Ontologies are specified in different standard languages, such as XML, OWL or RDF, and regular requirements are expressed using query languages such as SPARQL. Ontologies play an important role and provide the basis for the definition of concepts and relationships that make global interoperability among available Web resources possible. In the Health Care and Life Sciences domains, large ontologies have been defined; for example, we can mention MesH [15], Disease [1], Galen [16], EHR_RM [2], RxNorm [20], and GO [5]. Ontologies are commonly applied in these domains to annotate publications, documents, and images; also ontologies can be used to distinguish similar concepts, to generalize and specialize concepts, and to derive new properties. To fully take advantage from the linked data sources and their ontology annotations, and to be able to recognize novel discoveries, scientists have to navigate through the inter-connected sources, and compare, correlate and mine some of these annotated data. Nevertheless, because the size and number of available sources and the set of possible annotations are very large, users may have to spend countless hours or days before recognizing relevant findings.

In order to facilitate the specification of scientist's semantic connection needs, we present a ranking technique able to assign high scores to potential novel associations. Furthermore, given the size of the search space and to reduce the effect of the number of available linked data sources and ontology annotations on the performance, we also propose an approximate solution named graph-sampling. This approximate ranking technique samples events in a Bayesian network that models the topology of the data connections; it also estimates ranking scores that measure how important and relevant are the associations between two terms. In addition, the approximate technique exploits information about the topology of the hyperlinks and their ontology annotations, to guide the ranking process into the space of relevant and important terms.

In this paper we describe our ranking techniques and show their effectiveness and efficiency. The paper is composed of five additional sections. In Section 2, we compare existing approaches. Section 3 illustrates techniques proposed in the area of Literature Based Discovery (LBD) by showing the discovery reported in [21] where curcumin longa was associated with retinal diseases. Section 4 describes our proposed sampling technique. Section 5 reports our experimental results. Finally, we give our conclusions and future work in Section 6.

2 Related Work

Under the umbrella of the Semantic Web, Linked Data projects have proposed algorithms to discover links between datasets. Particularly, the Linking Open Drug Data (LODD) task has connected a list of datasets that includes disorders and disease genes [6], clinical trials [9] and drug banks [26]. Some of these link discovery or generation tools apply similarity metrics to detect potential similar concepts and their relationships [25]. However, none of the existing link discovery techniques make use of information about the link structure to identify potential novel associations. Also, the ontology voID [24]

has been proposed to describe interlinked datasets and enable their discovery and usage, and provides the basis for our proposed approach.

The discovery of associations between data entries implies descriptive and predictive inference tasks based on the link structure [4] and on semantics suggested by relevant ontologies. In general the idea is to perform random walks in the space of possible associations and discover those that satisfy a particular pattern; correspondences between the discovered patterns are measured in terms of similarity functions. In [7], heuristics are used to discover relevant subgraphs within RDF graphs; relationships among the metadata describing nodes is used to discover relevant relationships among entities. To decide if two objects are semantically similar, Jeh et. al. [11] propose a measure that reflects when two objects are similar based on the relationships that they hold with similar objects. Yan et al. [8] propose strategies to efficiently search subgraphs that are similar to a given query graph. Finally, Hu et al. [10] and Kuramochi and Karypis [12] describe efficient algorithms to discover subgraphs (patterns) that occur in graphs and to aggregate them.

Sampling techniques have been successfully applied to a variety of approximation techniques. For example, in the context of query optimization, different sampling-based algorithms have been proposed to estimate the cardinality of a query efficiently [13, 14, 18]. The challenge of these methods is to reach estimates that satisfy the required confidence levels while the size of the sample remains small. A key decision involves when to stop sampling the population and this is determined by the mean and variance of the sample in comparison to the target population. In this paper we propose a technique that samples paths in an acyclic directed graph that models a dataset of linked data. Paths are sampled based on the joint probability which is computed as the multiplication of the authority transfer flow value of the edges that comprise the path. Similarly, we define the stop condition of the sampling, based on an estimate of the metric score mean. Related to the problem of estimating authority flow metrics, Fogaras et. al. [3] implement a Monte-Carlo based method to approximate personalized PageRank scores. They sample paths whose length is determined by a geometric distribution. Paths are sampled from a Web graph based on a probability that represents whether objects in the paths can be visited by a random surfer. This approach may provide a solution to PageRank; however, it is not applicable to our proposed approach because the length of the paths is determined by the number of layers in the results graph and cannot be randomly chosen. In contrast, graph-sampling samples objects layer by layer, until the last layer in the result graph is visited. Objects with higher probability to be visited by a random surfer and links between these objects, will have greater chance to be chosen during the sampling process. Thus, graph-sampling may be able to only traverse relevant paths that correspond to relevant discoveries.

3 Motivating Example

Consider the area of Literature-Based Discovery (LBD) where by traversing scientific literature annotated with the controlled vocabularies like MesH, drugs have been associated with diseases [21, 22]. LBD can perform *Open* or *Closed* discoveries, where a scientific problem is represented by a set of articles that discuss an input problem

(*Topic A*), and the goal is to prove the significance of the associations between *A* and some other *C* topics discussed in the set of publications reachable from the initial set of publications relevant to *A*. Srinivasan et al. [21] followed this idea and improved the *Open* and *Closed* techniques by recognizing that articles in PubMed have been curated and heavily annotated with controlled vocabulary terms from the MeSH (Medical Subject Heading) ontology. Relationships between publications and terms are annotated with weights or scores that represent the relevance of the term in the document. MeSH term weights are a slight modification of the commonly used TF/IDF scores. Figure 1 illustrates a directed graph that represents the terms and publications visited during the evaluation of an *Open* discovery. Topic *A* is used to search on the PubMed site and retrieve relevant publications, named Pub_A. Then, MeSH term annotations are extracted from publications in Pub_A, and filtered by using a given set of semantic types of the ontology Unified Medical Language System (UMLS) [3]; this new set of MeSH terms is named *B* and is used to repeat the search on the PubMed site. Similarly, sets Pub_B, *C* and Pub_C are built.

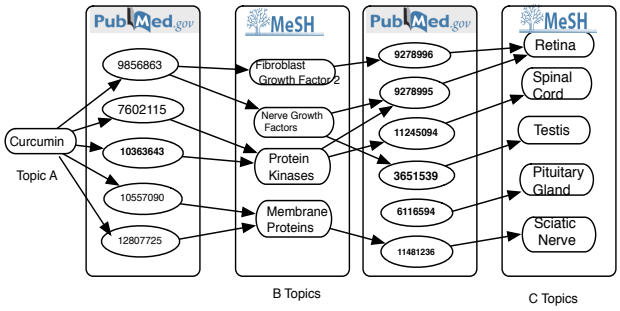

Fig. 1. Open Discovery Graph LBD

The Srinivasan's algorithm considerably reduces the space of intermediate results while identifying novel relationships; however, it still requires human intervention to create the intermediate datasets as well as to rank the terms that may not conduce to potential novel discoveries. We propose a sampling-based ranking technique that is able to estimate which are the nodes that will conduce to novel discoveries, and thus, reduce the discovery evaluation time. We illustrate the usage of this technique in the context of Literature-based Discovery. However, we hypothesize that this technique can be used to efficiently discover associations between the data published in the Cloud of Linked Data.

[3] http://www.nlm.nih.gov/research/umls/

4 A Ranking-based Solution to Discover Semantic Associations

We propose ranking-based solutions to the problem of the semantic association discovery. The proposed techniques take advantage of existing links between data published on the Cloud of Linked Data, or make use of annotations with controlled vocabularies such as MeSH, GO, PO, etc. We present an exact solution, and an approximate technique; both methods have been implemented in BioNav [23].

4.1 An Exact Ranking Technique

The exact ranking technique extends existing authority-flow based metrics like PageRank, ObjectRank or any of their extensions [17]. This ranking approach assumes that the linked data comprise a layered graph, named layered Discovery Graph, where nodes represent published data and edges correspond to hyperlinks.

Formally, a layered Discovery Graph, $lgDG$=(V_{lg}, E_{lg}) is a layered directed acyclic graph, comprised of k layers, $L_1, \ldots, L_k$. Layers are composed of data entries which point to data entries in the next layer of the graph. Data entries in the k-th layer or last layer of the graph, are called target objects. Authority-flow based metrics are used to rank the target objects, and we use these scores to identify relevant associations between objects in the first layer and target objects.

Figure 2 illustrates an example of a layered Discovery Graph that models the Open Discovery Graph in Figure 1. In this example, odd layers are composed of MeSH terms while even layers are sets of publications. Also, an edge from a term b to a publication p indicates that p is retrieved by the PubMed search engine when b is the search term. Finally, an edge from a publication p to a term b represents that p is annotated with b. Each edge $e = (b, p)$(resp., $e = (p, b)$) between the layers l_i and l_{i+1} is annotated with the TF/IDF score; this value either represents how relevant is a term b in the collection of documents in l_{i+1}, or a document relevance regarding to a set of terms. The path of thick edges connects Topic A with $C3$; the value 0.729 corresponds to the authority-flow score and represents the relevance of the association between Topic A and $C3$.

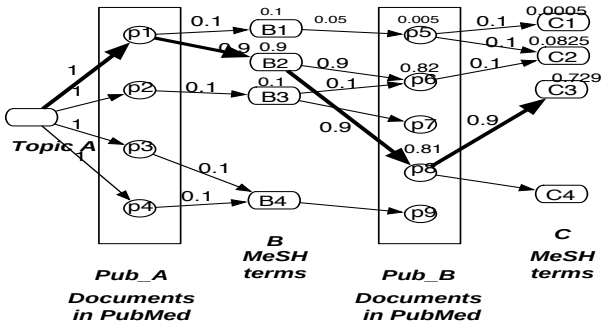

Fig. 2. A Layered Discovery Graph

Given a layered Discovery Graph $lgDG$=(V_{lg}, E_{lg}) of k layers, the authority-flow scores of the target objects are formally defined as a ranking vector R:

$$R = M^{k-1} R_{ini} = (\prod_{l=1}^{k-1} M^l)\, R_{ini}$$

where, M is a transition matrix and R_{ini} is a vector with the scores of the objects in the first layer of the graph. An entry $M[u, v]$ in the transition matrix M, where u and v are two data objects in $lgDG$, corresponds to $\alpha(u, v)$ or is 0.0. The value $\alpha(u, v)$ is calculated according to the metric used to compute the ranking score of the data.

$$M[u, v] = \begin{cases} \alpha(u, v) \text{ if } (u, v) \in E_{lg}, \\ 0.0 \quad \text{otherwise.} \end{cases}$$

For instance, the *layered graph Weighted Path Count* (lgWP) is an extension of ObjectRank and Path Count and the value of $\alpha(u, v)$ corresponds to the TF/IDF score that denotes how relevant is the object u with respect to the object v. Nodes with high lgWP scores are linked by many nodes or linked by highly scored nodes; for example, in Figure 2, $C3$ is pointed by relevant nodes. In the context of LBD, we use this metric to discover novel associations between a topic A and MeSH terms in the last layer of the $lgDG$, and we have been able to discover the associations identified by Srinivasan et al. [21].

4.2 A Sampling-based Ranking Solution

Although the ranking induced by an authority-flow based metric is able to distinguish relevant associations, the computation of this ranking may be costly. Thus, to speed up this task, we propose a sampling-based technique that traverses only nodes in the layered graph that may conduce to highly ranked target objects.

Given a layered Discovery Graph $lgDG = (V_{lg}, E_{lg})$, the computation of highly ranked target objects is reduced to estimating a subgraph $lgDG'$ of $lgDG$, so that with high confidence (at least δ), the relative error of the distance between the approximate highly ranked target objects in $lgDG'$ and the exact highly ranked target objects, is at least ϵ.

A set SS=$\{lgDG_1, ..., lgDG_m\}$ of independent and identically distributed (i.i.d.) subgraphs of $lgDG$ is generated. Then, $lgDG'$ is computed as the union of the m subgraphs. Each subgraph $lgDG_i$ is generated using a *graph-sampling* technique. This sampling approach is based on a Direct Sampling method for a Bayesian network [19]. This network represents all the navigational information encoded in $lgDG$ and in the transition matrix M of the authority-flow metric. The Direct Sampling technique generates events from a Bayesian network [19].

A Bayesian network $BN = (VB, EB)$ for a layered Discovery Graph $lgDG$, is built as follows:

- BN and $lgDG$ are homomorphically equivalent, i.e., there is a mapping $f : VB \rightarrow V_{lg}$, such that, $(f(u), f(v)) \in E_{lg}$ iff $(u, v) \in EB$.

- Nodes in VB correspond to discrete random variables that represent if a node is visited or not during the discovery process, i.e., $VB = \{X \mid X$ takes the value 1 (true) if the node X is visited and 0 (false), otherwise$\}$.
- Each node X in VB has a conditional probability distribution:

$$Pr(X \mid Parents(X)) = \sum_{j=1}^{n} \alpha(f(Y_j), f(X))$$

where, Y_j is the value of the random variable that represents the j-th parent of the node X in the previous layer of the Bayesian network and n corresponds to the number of parents of X. The value $\alpha(f(Y_j), f(X))$ represents the weight or score of the edge $(f(Y_j), f(X))$ in the layered Discovery Graph and corresponds to an entry in the transition matrix M; it is seen as the probability to move from Y_j to X in the Bayesian network. Furthermore, the conditional probability distribution of a node X represents the collective probability that X is visited by a random surfer starting from the objects in the first layer of the layered Discovery Graph. Finally, the probability of the nodes in the first layer of the Bayesian network corresponds to a score that indicates the relevance of these objects with respect to the discovery process; these values are represented in the R_{ini} vector of the ranking metric.

Given a Bayesian network generated from the layered Discovery Graph $lgDG$, the Direct Sampling generates each subgraph $lgDG_i$. Direct Sampling selects nodes in $lgDG_i$ by sampling the variables from the Bayesian network based on the conditional probability of each random variable or node. Algorithm 1 describes the Direct Sampling algorithm.

Algorithm 1 The Direct Sampling Algorithm

Input: $BN = (VB, EB)$ *A Bayesian network for a layered discovery graph*
Output: A subgraph $lgDG_i$

```
TP ⇐ topologicalOrder(BN);
for X ∈ TP do
    Pr(X | Parents(X)) ⇐ Σ_{j=1}^{n} α(f(Y_j), f(X));
    if (randomNumber >= Pr(X | Parents(X))) then
        X_i ⇐ 1;
    else
        X_i ⇐ 0;
    end if
end for
```

Variables are sampled in turn following a topological order starting from the variables in the first layer of the Bayesian network; this process is repeated until variables in the last layer are reached. The values assigned to the parents of a variable define the probability distribution from which the variable is sampled. The conditional probability of each node in the last layer of $lgDG_i$ corresponds to the approximate value of the implemented metric.

Figure 3 illustrates the behavior of the graph-sampling technique; unmarked nodes correspond to visited notes and comprise a subgraph $lgDG_i$. Direct Sampling is performed as follows: initially, all the nodes in the first layer have the same probability to be visited and all of them are considered. All their children or nodes in the second layer are also visited and the conditional probability is computed; nodes with the highest scores survive, i.e., $n5$ and $n7$. Then, the children of these selected nodes are also visited, and the process is repeated until nodes in the last layer are reached. Note that nodes $n9$ and $n11$ are the target objects with the highest values of the lgWP metric and with the highest conditional probability. These nodes are pointed by nodes with high lgWP scores or pointed by many nodes; thus, they are very likely to be visited when the Direct Sampling algorithm is performed.

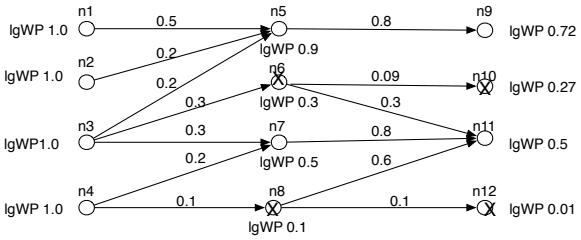

Fig. 3. Graph Sampling

Once an iteration i of the Direct Sampling is finalized, the sampled layered Discovery Graph $lgDG_i = (V_i, E_i)$ is created. Nodes in V_i correspond to the variables sampled during the Direct Sampling process that are connected to a visited variable in the last layer of the Bayesian network. Additionally, for each edge (u, v) in the Bayesian network that connects nodes $f(u)$ and $f(v)$ in V_i, an edge $(f(u), f(v))$ is added to E_i. The conditional probabilities of the target objects of each subgraph $lgDG_i$ correspond to the approximate values of the ranking metric. After all the subgraphs $lgDG_1, \ldots, lgDG_m$ are computed, an estimate $lgDG'$ is obtained as the union of these m subgraphs. The approximation of the ranking metric in the graph $lgDG'$ is computed as the average of the approximate ranking metric values of target objects in the subgraphs $lgDG_1, \ldots, lgDG_m$. A bound of the number of iterations or sampled subgraphs is defined in terms of the Chernoff-Hoeffdings bound.

Theorem: Let $lgDG$ be an exact layered Discovery Graph and $lgDG_i$ be one of the m sampled subgraphs. Let T be a list of the target objects in $lgDG$ ranked with respect to exact values of the ranking metric RM. Let T_i be a list of the target objects in $lgDG_i$ ranked with respect to the approximation of RM. Let $J(lgDG_1, lgDG, \beta), \ldots, J(lgDG_m, lgDG, \beta)$ be independent identically distributed (i.i.d.) random variables with values in the set {0,1}. Each random variable $J(lgDG_i, lgDG, \beta)$ has value=1 if a distance metric value between the ranking list T_i and the list T is at least β; otherwise, value=0. Let S denote the average of these variables, i.e., $X = \frac{1}{m} \sum_{i=1}^{m} J(lgDG_i, lgDG, \beta)$ and $E(S)$ the expectation of S. Then, the size m of the sample has to satisfy the fol-

lowing formula to ensure that the relative error of $E(S)$ is greater than ϵ with some probability:

$$P(|S - E(S)| \geq \epsilon) \leq 2\exp(-2m\epsilon^2)\,.$$

5 Experimental Results

In this section we show the quality of our proposed discovery techniques. First, we compare the results obtained by our ranking technique with respect to the results obtained by the Manjal system [21]. Then, we show the behavior of this technique in the DBLP dataset. Experiments were executed on a Sun Fire V440 equipped with two UltraSPARC IIIi processors running at 1.593 GHZ with 16 GB RAM. The ranking and sampling techniques were implemented in Java 1.6.1.

To conduct the first experiment, we have created a catalog populated with the PubMed publications from the NCBI source[4], all the MeSH terms, and all the links between Mesh terms and PubMed publications. We stored the downloaded data in two tables, *Pub-MeSH* and *MeSH-Pub*. Table *Pub-MeSH* relates a publication p with all the MeSH terms that correspond to annotations of p in PubMed; these annotations are manually done by experts at the National Library of Medicine site. Table *MeSH-Pub* relates a MeSH term m with all publications that are retrieved when the term m is used to search on PubMed. Both tables have an attribute *score* that represents the relevance of the relationships represented in the table. Suppose there is a tuple (p, m, s) in table *Pub-MeSH*, then the score $s = A \times T \times C$, where:

- A: is the augmented document frequency of the publication p, i.e., $A = 0.5 + 0.5 + \frac{tf}{tf_{max}}$, where, tf is the frequency of p in table *Pub-MeSH*, and tf_{max} is the maximum document frequency of any publication in *Pub-MeSH*.
- T: inverse term frequency $log_2(\frac{N}{Np})$, where N is the number of collected MeSH terms, i.e., 20,652, and N_p corresponds to the number of MeSH terms associated with the publication p in the table *Pub-MeSH*.
- C: is a cosine normalization factor.

Similarly, scores in table *MeSH-Pub* were computed. To reproduce the results reported by Srinivasan et al. in [21], we ran the metric lgWP on a layered Discovery Graph *lgDG* comprised of 5 layers, 3,107,901 nodes and 10,261,791 edges. Sets Pub_A, B, Pub_B and C and were built following the criteria proposed by Srinivasan et al., and by selecting data from tables *Pub-MeSH* and *MeSH-Pub*. We ranked the target objects in the graph, and we could observe that our ranking technique was able to produce 4 of the top-5 semantic associations identified by Srinivasan et al. [21]. Table 1 compares the top-5 target objects discovered by [21] and the ones discovered by our ranking technique, i.e., our ranking technique exhibits a precision and recall of 80%.

We have also studied the benefits of performing the graph-sampling technique, and we ran the sampling process for 5 iterations, i.e., 5 sampled subgraphs were computed. Table 2 reports on the top-10 MeSH terms identified by graph-sampling. We can observe that 4 of the top-5 MeSH terms identified by the Srinivasan's algorithm [21],

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

k	Srinivasan's Ranking [21]	lgWP
1	Retina	Testis
2	Spinal Cord	Retina
3	Testis	Spinal Cord
4	Pituitary Gland	Obesity
5	Sciatic Nerve	Pituitary Gland

Table 1. Top-5 MeSH terms

are also identified. We note that iterations do not improve the quality of the discovery process.

k	i=1	i=2	i=3	i=4	i=5
1	Spinal Cord	Spinal Cord	Spinal Cord	Spinal Cord	Spinal Cord
2	Pituitary Gland	Pituitary Gland	Pituitary Gland	Pituitary Gland	Pituitary Gland
3	Celiac Disease	Celiac Disease	Celiac Disease	Celiac Disease	Disease
4	Hepatic Encepha.	Hepatic Encepha.	Hepatic Encepha.	Hepatic Encepha.	Hepatic Encepha.
5	Uremia	Uremia	Uremia	Uremia	Uremia
6	Retina	Anemia	Anemia	Anemia	Anemia
7	Obesity	Retina	Retina	Retina	Retina
8	Testis	Obesity	Phenylketonurias	Phenylketonurias	Phenylketonurias
9	Hypothalamus	Testis	Obesity	Obesity	Obesity
10	Osteoporosis	Hypothalamus	Testis	Testis	Testis

Table 2. Effectiveness of Graph Sampling Techniques

Finally, we report on the number of target MeSH terms produced by the Srinivasan's algorithm and the ones produced during each iteration of graph-sampling (Table 3). We can observe that graph-sampling is able to discover 80% of the top novel MeSH terms, while the number of target terms is reduced by up to one order of magnitude.

# Srinivasan's target MeSH Terms [21]	i=1	i=2	i=3	i=4	i=5
570	24	38	49	61	71

Table 3. Performance of Graph-Sampling Techniques

In the second experiment, we downloaded the DBLP file in a relational database. We ran the graph-sampling technique to discover associations between a given author and the most relevant conferences where this author has published at least one paper. We ran 3 sets of 30 queries and compared the ranking produced by the exact solution and the one produced by graph-sampling; layered Discovery Graphs were comprised

of 5 layers and at most 876,110 nodes and 4,166,626 edges. Author's names with high, medium and low selectivity were considered, where high selectivity means that the author has few publications while low selectivity represents that the author is very productive. The top-5 conferences associated with each author were computed by using the exact ranking and the approximation produced by graph-sampling during 6 iterations. Table 4 reports the average precision of the approximate top-5 conferences with respect to the exact top-5. We can observe that graph-sampling is able to identify almost 65% of the top-5 conferences after iteration 3. The time required to execute the graph-sampling technique was reduced at least by half. These results suggest that the proposed discovery techniques provide an effective and efficient solution to the problem of identifying associations between terms.

Author's Name Selectivity	i=1	i=2	i=3	i=4	i=5	i=6
high	0.390	0.4874	0.635	0.813	0.823	0.871
medium	0.341	0.562	0.681	0.724	0.872	0.890
low	0.64	0.660	0.749	0.803	0.806	0.815

Table 4. Effectiveness of Graph Sampling Techniques DBLP- Average Precision

6 Conclusions and Future Work

In this paper we have presented a sampling-based technique that supports the discovery of semantic associations between linked data. We have reported the results of an empirical study where we have observed that our proposed techniques are able to efficiently reproduce the behavior of existing LBD techniques. This observed property of our discovery technique may be particularly important in the context of large datasets as the ones published in the Cloud of Linked Data. In the future we plan to extend this study to identify potential associations between other sources of the Cloud of Linked Data.

References

1. Disease Ontology. http://diseaseontology.sourceforge.net.
2. EHR Ontology. http://trajano.us.es./ isabel/EHR/EHRRM.owl.
3. D. Fogaras, B. Rácz, K. Csalogány, and T. Sarlós. Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. *Internet Mathematics*, 2(3), 2005.
4. L. Getoor and C. P. Diehl. Introduction to the special issue on link mining. *SIGKDD Explorations*, 7(2), 2005.
5. The Gene Ontology. http://www.geneontology.org/.
6. K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabasi. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104:8685–8690, 2007.
7. C. Halaschek-Wiener, B. Aleman-Meza, I. B. Arpinar, and A. P. Sheth. Discovering and ranking semantic associations over a large rdf metabase. In *VLDB*, pages 1317–1320, 2004.

8. J. Han, X. Yan, and P. S. Yu. Mining, indexing, and similarity search in graphs and complex structures. In *ICDE*, page 106, 2006.
9. O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller, and M. Wang. Linkedct: A linked data space for clinical trials. In *Proceedings of the WWW2009 workshop on Linked Data on the Web (LDOW2009)*, 2009.
10. H. Hu, X. Yan, Y. H. 0003, J. Han, and X. J. Zhou. Mining coherent dense subgraphs across massive biological networks for functional discovery. In *ISMB (Supplement of Bioinformatics)*, pages 213–221, 2005.
11. G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *KDD*, pages 538–543, 2002.
12. M. Kuramochi and G. Karypis. Finding frequent patterns in a large sparse graph*. *Data Min. Knowl. Discov.*, 11(3):243–271, 2005.
13. Y. Ling and W. Sun. A supplement to sampling-based methods for query size estimation in a database system. SIGMOD Record, 21(4):12–15, 1992.
14. R. Lipton and J. Naughton. Query size estimation by adaptive sampling (extended abstract). In PODS '90: Proceedings of the ninth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, pages 40–46. New York, NY, USA: ACM Press, 1990.
15. Medical Subject Heading (MeSH). http://www.nlm.nih/gov/mesh.
16. O. C. Organization. GALEN common reference model.
17. L. Raschid, Y. Wu, W. Lee, M. Vidal, P.Tsaparas, P. Srinivasan, and A. Sehgal. Ranking target objects of navigational queries. In *WIDM*, pages 27–34, 2006.
18. E. Ruckhaus, E. Ruiz, and M. Vidal. Query optimization in the semantic web. In Theory and Practice of Logic Programming. Special issue on Logic Programming and the Web, 2008.
19. S. Russell and P.Norvig. *Artificial Intelligence: A Modern Approach. Second Edition*. Princeton Hall, 2003.
20. An Overview to RxNorm. http://www.nlm.nih.gov/research/umls/rxnorm/.
21. P. Srinivasan, b. Libbus, and A. Kumar. Mining medline: Postulating a beneficial role for curcumin longa in retinal diseases. In L. Hirschman and J. Pustejovsky, editors, *LT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*, pages 33–40, 2004.
22. D. Swanson. Migraine and magnesium: Eleven neglected connections. In *Perspective in Biology and Medicine*, 1988.
23. M.-E. Vidal, E. Ruckhaus, and N. Marquez. BioNav: A System to Discover Semantic Web Associations in the Life Sciences. In *ESWC 09-Poster Session*, 2009.
24. voiD Guide - Using the Vocabulary of Interlinked Datasets. http://rdfs.org/ns/void-guide.
25. J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data. In *International Semantic Web Conference (ISWC)*, 2009.
26. D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34, 2006.