On The Discovery of Relational Patterns in Semantically Similar Annotated Linked Data

Guillermo Palma

Universidad Simón Bolívar, Venezuela
gpalma@ldc.usb.ve

Advisor: Maria-Esther Vidal (mvidal@ldc.usb.ve)

Abstract. A wide variety of publicly linked datasets have been annotated with domain-specific ontologies. Annotations can be naturally represented with graphs, and the knowledge encoded in these annotations can be mined to discover potential novel relationships. We propose novel mining techniques that exploit semantics represented by these graphs to discover relational patterns. Initial experimental results suggest that our approach can be effectively applied in different biomedical domains, and exhibit performance comparable to state-of-the-art solutions.

Keywords: #eswcphd2014Palma, Mining Patterns; Linking Data; Annotated Graph.

1 Introduction and Motivation

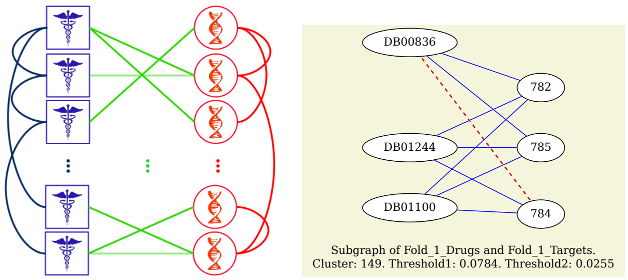
The number of highly connected datasets has exploded during the last years. The Linked Open Data (LOD) cloud has more than 50 billion facts from many different domains, e.g., media, biology, chemistry, economy and energy [1]. The LOD cloud can be naturally represented as graphs, specifically, with heterogeneous information networks. Heterogeneous information networks are graphs with multiple typed nodes and links that represent different relationships. Examples of heterogeneous information networks include social networks, the World Wide Web, research publication networks [10], biological networks, knowledge networks, among other networks. Due to the diverse meanings in heterogeneous information networks, mining patterns is difficult without considering the semantic of the typed concepts and relationships. A heterogeneous information network can be built upon highly structured data in the form of a graph, representing different types of nodes and edges. As many of these approaches rely on graph-based tasks, several efficient algorithms have been proposed not only to consume, but also to mine Linked Data. For example, Saha et al. [20] and Thor et al. [26] have defined densest subgraphs and graph summarization techniques to identify patterns between linked datasets of genes.

Furthermore, ontologies are developed by domain experts to capture knowledge specific to some domain. They have been extensively developed and widely adopted in the last decade. Simultaneously, Linked Open Data initiatives have made available a diversity of collections that have been annotated with domain-specific ontologies. These annotations describe *properties* of these concepts. For example, the biomedical community has taken the lead in such activities; every model organism database has genes and proteins that are widely annotated using the Gene Ontology (GO). These annotated datasets have created many opportunities for large scale Linked Data mining. Annotations induce an annotated graph where nodes correspond to concepts or ontology terms, and edges represent relationships between concepts. Our research aims as defining novel methodologies to exploit and mining annotated graph datasets and the semantic knowledge captured within ontologies to discover complex patterns of semantically related concepts in the Linked Data. The methodologies are based on various mining tasks in the Linked Data, including clustering, classification, similarity metrics between concepts, relationship prediction and structural learning. We tackle these mining tasks, their principles and methodologies.

Motivating Example: we motivate our work with the link prediction problem presented by Fakhraei et al.[8]. The development of new drugs is a time-consuming and costly procedure, and one possibility is repurposing already approved drugs for new diseases. Repurposing existing drugs using computational methods has the benefits of shorter timelines to bring a drug to the market and reduce its cost. Drugs are molecules that participate in some biomolecular reaction associated with a disease target. There may be multiple relationships between drugs and targets. With the goal of predicting drug-target interactions, we can build a bipartite graph between drugs and targets, where edges are interactions known by the scientific community. We can augment the bipartite graph with drug-drug and target-target similarities. The similarities between drugs and between targets have different semantics. For example, drugs can have similarities based on chemical structure or shared side-effects, while gene targets may share sequence based or gene annotation based similarity [18]. Figure 1(a) shows a drug-target interaction network. The challenge is that the drug-target interaction graph, with multiple types of similarities, expresses a multi-relational graph structured knowledge. This drug-target graph, combined with knowledge in ontologies and additional LOD resources, will be used for discovery potentially new drug-target interaction.

2 State of the art

Graph data mining [6] covers a broad range of methods dealing with the identification of structures and patterns in graphs. Popular techniques include graph clustering [5], community detection [9] and cliques [16]. Clustering, classification and ranking are basic mining functions for information networks. Spectral graph clustering [27] is state-of-the-art method to do clustering on homogeneous networks. For heterogeneous networks RankClus [25] is proposed and generates clusters integrated with ranking. A ranking-based classification of multiple types of objects, denoted by GNetMine, is proposed by Ming et al. [12]. Link prediction has been extensively studied in the recent years [8, 26]. The problem of a 1-to-1 weighted maximal bipartite match has been applied to many problems, e.g., semantic equivalence between two sentences and measuring similarity be-


DB00836
782
DB01244
785
DB01100
784
Subgraph of Fold_1_Drugs and Fold_1_Targets.
Cluster: 149. Threshold1: 0.0784. Threshold2: 0.0255

(a) Drug-Target network (b) Our approach prediction.

Fig. 1. (a) Drug-Target interaction network from [8]. Blue lines expressing drug-drug similarities, red lines target-target similarities and Green lines are known drug target interactions. (b) Cluster obtained by our approach on Dataset 1: network of drugs and targets and their interactions. The red line is a predicted interaction.

tween shapes for object recognition[22]. A key element in finding patterns is identifying related concepts and similarity metrics can be used to measure ontological relatedness. A class of metrics are path-similarity metrics based on the paths that connect the concepts in a graph. Nodes in the paths can be of the same type (e.g., PathSim [24]) or they can be heterogeneous (e.g., HeteSim [23]). Furthermore, semantic similarity metrics can be classified into two categories: *i*) structure-based metrics that exploit ontology hierarchy structure to compute the semantic similarity between terms [14, 2], *ii*) information content (IC) based metrics that use IC of concepts derived from corpus statistics to measure the semantic similarity between terms [19].

Loza et al.[15] apply data mining techniques to estimate the number of bidders in public contracts represented as semantically annotated Linked Data. The proposed techniques rely on existing machine learning algorithms which are applied to a relational representation of the linked data. Our proposed approach also copes this problem but it exploits knowledge encoded in ontologies to uncover hidden relational patterns.

3 Problem statement

How much effectively are data mining techniques to discover relational patterns in annotated Linked Data?. Our research addresses the challenge of mining large annotated graphs, and exploiting knowledge from ontologies to discover patterns that uncover hidden relationships between semantically similar data.

Our first research goal is to propose a novel similarity metric based on annotations, called *AnnSim*, that is able to measure relatedness between concepts in an annotated graph, based on the similarity of the sets of their annotations with respect to one or more ontologies. This is the necessary first step to discover complex patterns in annotated graph datasets. A practical example is identifying the relatedness or similarity of (drug, drug) pairs, based on the annotation evidence of conditions or diseases from domain-specific ontologies as the NCI Thesaurus (NCIt). NCIt Home Page: `http://ncit.nci.nih.gov/`.

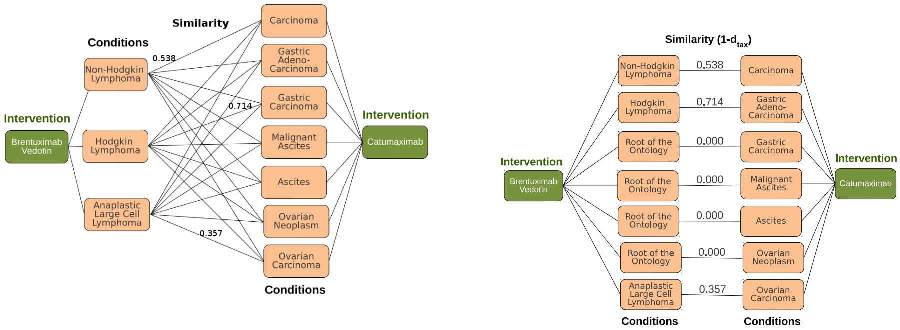
Our second research goal is to discover complex relational patterns, thus we define an *annotation signature* between a pair of concepts, e.g., a pair of drugs or a pair of genes. The annotation signature builds upon the shared annotations or shared ontology terms between the pair of concepts. The signature further makes use of knowledge in the ontology to determine the ontological relatedness of the shared terms. The annotation signature is represented by N groups (clusters) of ontologically related shared terms. For example, the annotation signature for a (drug, drug) pair will be a set of N clusters, where each cluster includes a group of ontologically related disease terms from NCIt. Given a pair of concepts, and their sets of annotations, A_i and A_j from ontology O, elements $a_i \in A_i$ and $a_j \in A_j$ form the nodes of a bipartite graph (BG). Between nodes a_i and a_j there may be an edge or a path through O; an edge is the special case where a_i and a_j are identical terms from O. There may be a choice of paths between a_i and a_j depending on the the ontology structure and relationship types captured within O. One can use a variety of similarity metrics, applied to the edges and paths through the ontology O, to induce a weighted edge between a_i and a_j in BG; the weight represents the (ontologically related) similarity score in the range $[0, 1]$ between a_i and a_j. Our objective is to determine an annotation signature based on the BG. There are many alternatives to create the signature. One could partition the edges of BG with possible overlap of the nodes. Another solution is to cluster the nodes and edges of BG. One may also consider a one-to-one bipartite match. The clusters obtained may identify multiple communities (subgraphs) of ontologically related shared terms, as well as potentially overlapping communities. Our research on annotation similarity will explore such patterns that can be used for link prediction or to rank the graph concepts. Thus, we can exploit ontologically related communities identified within a data mining framework.

Our third research goal is the development of machine learning frameworks to identify interesting relational patterns involving ontologically related concepts.

4 Proposed approach and Methodology

We can use the weight of annotation evidence, represented by a set of annotations, to define a metric to compare a pair of concepts. An annotated graph *G=(V,E)* is a particular graph comprised of two type of nodes in V: scientific concepts and ontology terms. Given two concepts c_1 and c_2 from an annotated graph *G=(V,E)*, we define an annotation similarity metric, *AnnSim*, based on their sets of annotations, A_1 and A_2, respectively. We assume that we have a pair-wise similarity between elements of A_1 and A_2, i.e., $sim(a_1, a_2) \in [0, 1]$ for all $a_1 \in A_1$ and $a_2 \in A_2$. The value of $sim(a_1, a_2)$ is determined by the previous task of ontological similarity. These relationships between terms in A_1 and A_2 can be represented as a weighted bipartite graph $WBG{=}(A_1 \cup A_2,\ WE)$, see Figure 2(a). An edge between $a_1 \in A_1$ and $a_2 \in A_2$ has a weight $sim(a_1, a_2)$, where $sim(a_1, a_2)$

is computed using a distance metric. The computation of *AnnSim* first requires building a bipartite graph with the links in the Cartesian product between the set of annotations of two scientific terms, and for each of these links compute a similarity. The aim is to design the best approach to solve the *Weighted Bipartite Matching*. We first consider model *AnnSim* as a *1-to-1 maximal weighted bipartite matching* [21], see Figure 2(b). We name this annotation similarity *AnnSim*, and it is defined it as follows: Given two concepts c_1 and c_2 annotated with the set of terms A_1 and A_2 in an annotated graph, and let *MWBG=(* $A_1 \cup A_2$*, WEr)* be *1-to-1 maximal weighted bipartite graph matching* for a *WBG*, where *WEr* $\subseteq$ *WE*, we have $AnnSim(c_1, c_2) = \frac{2 \cdot \sum_{(a_1,a_2) \in WEr} sim(a_1,a_2)}{|A_1|+|A_2|}$. This definition is in the style of the well-known Dice coefficient. The maximal similarity of 1.0 is achieved if and only if both annotation sets have the same cardinality ($|A_1| = |A_2|$) and all edge weights are equal to 1. This approach has limitations. We planned to obtain solutions to the many-to-many bipartite match problem to compute an enhanced metric. Initial results are reported at [17].



(a) Weighted Bipartite graphs for drugs Brentuximab vedotin and Catumaxomab. Shown similarity values are the highest values obtained with the metric $1 - d_{tax}$[2] on NCIt

(b) 1-to-1 Maximal Weighted Bipartite Graph Match for Brentuximab vedotin and Catumaxomab. The similarity value of *AnnSim* is 0.324

Fig. 2. Bipartite graphs for drugs Brentuximab vedotin and Catumaxomab.

We define a version of the *Annotation Signature Partition* problem as the partitioning of the edges of *BG* into clusters such that the value of the aggregated cluster density is maximized. We develop *AnnSigClustering*, a clustering solution that implements a greedy iterative algorithm to cluster the edges in *BG*. We note that such a clustering will result in N clusters of the edges of *BG* with potential overlap of nodes in different clusters.

Definition 1 (Cluster Density). *Given a labeled bipartite graph BG=(*$A_i \cup A_j$*, WE) with nodes* A_i *and* A_j *and edges* WE*, a distance metric d, and a subset p of WE, the cluster density of p cDensity(p)* $= \frac{\sum_{(e=(a,b)) \in p} 1-d(a,b)}{|p|}$.

Definition 2 (Similar Nodes $\sim$). *Given two nodes a and b, a real number θ in the range $[0:1]$, and a distance metric d, nodes a and b are similar, i.e., $a \sim b$, iff $1 - d(a,b) > \theta$.*

Definition 3 (The Annotation Signature Partition Problem). *Given a labeled bipartite graph BG=($A_i \cup A_j$, WE), a distance metric d, and a real number θ in the range [0,1]. For each $a \in A_i$ and $b \in A_j$, if $(a \sim b)$ and $\neg((a \in A_j \wedge b \in A_i) \wedge (a \neq b))$, then there is an edge $e = (a,b) \in$ WE. For each $e = (a,b) \in$ WE, label(e)= 1-d(a,b). The AnnSig Partition Problem identifies a (minimal) partition P of WE such that the aggregate cluster density P $AnnSig(P) = \frac{\sum_{p \in P}(cDensity(p))}{|P|}$ is maximal.*

We model *the Annotation Signature Partition Problem* using the Vertex Coloring Graph (VCG) problem. The Vertex Coloring Graph problem assigns a color to every vertex in a graph such that adjacent vertices are colored with different colors and the number of colors is minimized; this problem has been shown to be NP-hard [13]. Each component of the shared signature of the *the Annotation Signature Partition Problem* corresponds to a color in the VCG problem. We extend a well-known approximation named the DSATUR algorithm [3] to solve the VCG to obtain a signature and compute the *AnnSigClustering* value.

AnnSigClustering is a greedy iterative algorithm, based on DSATUR algorithm [3], to solve the *Annotation Signature Partition Problem. AnnSigClustering* adds an edge to a cluster following a greedy heuristic to create clusters that maximize the cluster density. *AnnSigClustering* assigns a score to an edge e in *WE* according to the number of edges whose adjacent terms are dissimilar to the terms of e, and that have been already assigned to a cluster. Then, edges are chosen in terms of this score (descendant order). Intuitively, selecting an edge with the maximum score, allows *AnnSigClustering* to place first the edges with more restrictions; this is one for which there is a smaller set of potential clusters. The selected edge is assigned to the cluster that maximized the cluster density function. Time complexity of *AnnSigClustering* is $O(|\,WE|^3)$.

5 Preliminary Results

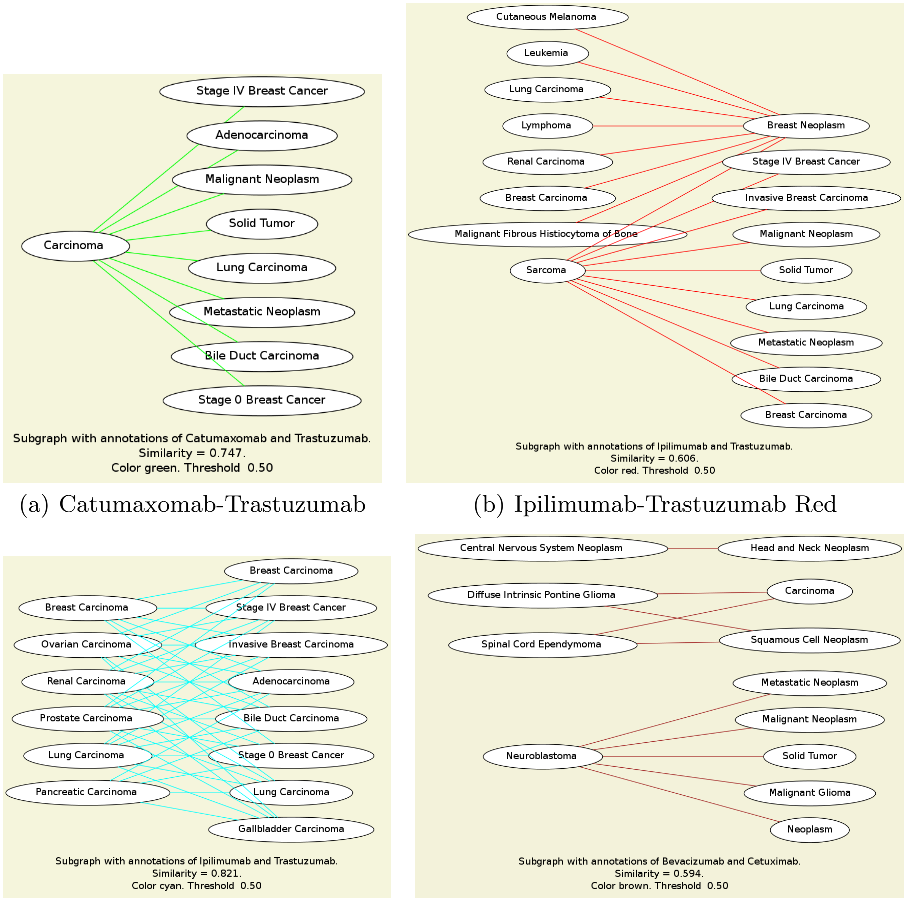
The goal of our evaluation is to validate if annotation signatures group together meaningful terms across shared annotations. Additionally, we evaluate the impact of the semantics encoded in the ontologies on the quality of the signature. We perform an evaluation of applying the *AnnSigClustering* results for link prediction. We consider two different datasets. Dataset 1 is based on a network of drugs and genetic targets and their interactions. The interactions were obtained from DrugBank [28]. The dataset has 315 drugs, 250 targets and 1306 interactions. We use 5 drug-drug similarities (Chemical-based, Ligand-based, Expression-based, Side-effect-based, and Annotation-based) and 3 target-target (Sequence-based, Protein-based and Gene Ontology-based) similarities, obtained from Perlman et al. [18]. Dataset 2 is comprised of twelve drugs within the intersection of monoclonal antibodies and antineoplastic agents; the name of the drug is followed by the abbreviation that we use in reporting results: Alemtuzumab, Bevacizumab, Brentuximab vedotin, Cetuximab, Catumaxomab, Edrecolomab, Gemtuzumab, Ipilimumab, Ofatumumab, Panitumumab, Rituximab, and Trastuzumab. The protocol to create the dataset is as follows: Each drug was used to retrieve a set of clinical trials in LinkedCT *circa* September 2011 (`linkedct.org`). Then each disease associated with each trial was linked to its corresponding term in the NCI Thesaurus version 12.05d; annotation was performed by NCIt experts. Our group of evaluators included two experts who develop databases and tools for the NCI Thesaurus, and two bioinformatics researchers with expertise on the NCIt and other biomedical ontologies.

We analyze the quality of *AnnSigClustering* predicting new iterations between drugs and targets in Dataset 1. Similarities between two drugs and two targets are considered to decide if they are or not related. We consider different thresholds between similarities drugs and targets. *AnnSigClustering* was used to compute the partition of the iterations between drugs and targets. Given a cluster, an edge between a drug and target in the cluster that was not included in the cluster was considered as a prediction. The graph density of the cluster was used as the probability that one edge was an interaction or not. We computed the Area Under the ROC Curve (AUC) to analyze the quality of our techniques. The state-of-the-art solution for this dataset is by Fakhraei et al.[8], and proposes a drug-target prediction supervised method based on PSL [4]. Table 1 shows the best result of our approach for Dataset 1. Figure 1(b) illustrates the cluster 149, the drugs DB00836 (Loperamide), DB01244 (Bepridil) and DB01100 (Pimozide) are associated with three gene targets. Predicted interactions are shown as broken edges.

Table 1. *AnnSigClustering* best result versus Fakhraei et al.[8] on the our approach prediction. Similarity drug-drug: Expression-based and target-target: Sequence-based

Method	**AUC**	**Execution Time**
AnnSigClustering	0.9431	7 min (Intel i7 3.3 Ghz)
Fakhraei et al.[8] on this prediction	0.9269	3h + 10h of learning (Xeon 2.9 GHZ)

In Dataset 2, our challenge is to identify connectivity patterns and knowledge encoded in each component. The connectivity pattern within each cluster provides insight into the ontological relatedness of the diseases. In Figure 3(a) `Carcinoma` on the left is connected to 8 terms on the right. In Figure 3(b) is a more complex pattern, where `Sarcoma` and `Breast Neoplasm` show high betweenness centrality. `Sarcoma` on the left is connected to 9 drugs on the right, and `Breast Neoplasm` on the right is connected to 8 diseases on the left. None of the other drugs has more than one adjacent drug in this subgraph. In contrast, in Figure 3(c), we see a much more general many-to-many connection pattern between the diseases on the left and right. Finally, Figure 3(d) shows a more complex connectivity pattern where the terms are ontologically related but they are placed within three disconnected graphs. The four terms `Diffuse Intrinsic Pontine Glioma`, `Spinal Cord Ependymoma`, `Carcinoma` and `Squamous Cell Neoplasm` form the most well connected cluster. Comments from the evaluators noted that while



(a) Catumaxomab-Trastuzumab

(b) Ipilimumab-Trastuzumab Red

(c) Ipilimumab-Trastuzumab Cyan

(d) Bevacizumab-Cetuximab Brown

Fig. 3. Connectivity Patterns within Each Cluster for $\theta = 0.5$

groups such as Figure 3(a) that included generic terms such as `Carcinoma` were valid, they did not convey useful information. In contrast, groups in Figures 3(c) and (d), that had more specific terms and were more densely connected, had the potential to be more meaningful.

6 Evaluation Plan

We will develop a semantic metric by the many-to-many connection pattern between two concepts. Evaluation of the our approach will be performed in other biomedical datasets that represent diverse type of relationships between drugs, diseases, targets, and enzymes. Our mining methods will use general knowledge base and ontologies as OpenCyc[1] and Yago [11] and other specialized as

[1] http://www.cyc.com/platform/opencyc

SNOMED CT[2]. We will develop a algorithm for learning threshold and a machine learning framework to obtain semantically related structures. Furthermore, we will compare with state-of-the-art learning-based approaches [7] and predicting system as PSL [4] for predicting drug-target interactions.

7 Conclusions and future work

We showed the feasibility of mining patterns semantically related in the LOD. We have defined the *Annotation Signature Partitioning Problem* and the *AnnSigClustering* algorithm to develop the components of a signature based on shared annotations and ontological relatedness. We have analyzed the effects of considering knowledge encoded in the ontologies used to annotate Linked Data. We have identified clusters can be used for link prediction and discover complex patterns. In the future we plan to conduct a deeper evaluation, as indicated in the previous section, and thus determine the potential discovery capability of the approach.

Acknowledgment. We thank Prof. Louiqa Raschid for all the insightful discussions and ideas to define the proposed approach.

References

1. Bauer, F., Kaltenbock, M.: Linked Open Data: The Essentials. edition mono/monochrom (2013)
2. Benik, J., Chang, C., Raschid, L., Vidal, M.E., Palma, G., Thor, A.: Finding cross genome patterns in annotation graphs. In: Proceedings of Data Integration in the Life Sciences (DILS) (2012)
3. Brélaz, D.: New methods to color vertices of a graph. Commun. ACM 22(4) (1979)
4. Broecheler, M., Mihalkova, L., Getoor, L.: Probabilistic similarity logic. In: Conference on Uncertainty in Artificial Intelligence (2010)
5. Brohee, S., van Helden, J.: Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7(1), 488 (2006)
6. Cook, D.J., Holder, L.B.: Mining graph data. Wiley-Blackwell (2007)
7. Ding, H., Takigawa, I., Mamitsuka, H., Zhu, S.: Similarity-based machine learning methods for predicting drug–target interactions: a brief review. Briefings in bioinformatics p. bbt056 (2013)
8. Fakhraei, S., Raschid, L., Getoor, L.: Drug-target interaction prediction for drug repurposing with probabilistic similarity logic. In: ACM SIGKDD International Workshop on Data Mining in Bioinformatics (BIOKDD) (2013)
9. Fortunato, S.: Community detection in graphs. Physics Reports 486(3-5), 75–174 (2010)
10. Giles, C.L.: The future of citeseer: citeseer x. In: Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases. pp. 2–2. Springer-Verlag (2006)
11. Hoffart, J., Suchanek, F.M., Berberich, K., Lewis-Kelham, E., De Melo, G., Weikum, G.: Yago2: exploring and querying world knowledge in time, space, context, and many languages. In: Proceedings of the 20th international conference companion on World wide web. pp. 229–232. ACM (2011)

[2] http://www.ihtsdo.org/snomed-ct/

12. Ji, M., Sun, Y., Danilevsky, M., Han, J., Gao, J.: Graph regularized transductive classification on heterogeneous information networks. In: Machine Learning and Knowledge Discovery in Databases, pp. 570–586. Springer (2010)
13. Karp, R.: Reducibility among combinatorial problems. In: Miller, R., Thatcher, J. (eds.) Complexity of Computer Computations, pp. 85–103. Plenum Press (1972)
14. McInnes, B., Pedersen, T., Pakhomov, S.: Umls-interface and umls-similarity : Open source software for measuring paths and semantic similarity. Proceedings of the AMIA Symposium pp. 431–435 (2009)
15. Mencıa, E.L., Holthausen, S., Schulz, A., Janssen, F.: Using data mining on linked open data for analyzing e-procurement information. In: Proceedings of the International Workshop on Data Mining on Linked Data, with Linked Data Mining Challenge collocated with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013) (2013)
16. Mougel, P.N., Plantevit, M., Rigotti, C., Gandrillon, O., Boulicaut, J.F.: Constraint-based mining of sets of cliques sharing vertex properties. In: Workshop on Analysis of Complex NEtworks (ACNE 2010) co-located with ECML/PKDD. Citeseer (2010)
17. Palma, G., Vidal, M.E., Haag, E., Raschid, L., Thor, A.: Measuring relatedness between scientific entities in annotation datasets. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics. p. 367. ACM (2013)
18. Perlman, L., Gottlieb, A., Atias, N., Ruppin, E., Sharan, R.: Combining drug and gene similarity measures for drug-target elucidation. Journal of Computational Biology 18(2), 133–145 (2011)
19. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal Of Artificial Intelligence Research 11, 95–130 (1999)
20. Saha, B., Hoch, A., Khuller, S., Raschid, L., Zhang, X.N.: Dense subgraphs with restrictions and applications to gene annotation graphs. In: Conference on Research on Computational Molecular Biology (RECOMB) (2010)
21. Schwartz, J., Steger, A., Weißl, A.: Fast algorithms for weighted bipartite matching. In: WEA. pp. 476–487 (2005)
22. Shavitt, Y., Weinsberg, E., Weinsberg, U.: Estimating peer similarity using distance of shared files. In: International workshop on peer-to-peer systems (IPTPS). vol. 104 (2010)
23. Shi, C., Kong, X., Yu, P.S., Xie, S., Wu, B.: Relevance search in heterogeneous networks. In: EDBT. pp. 180–191 (2012)
24. Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T.: Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB 4(11), 992–1003 (2011)
25. Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., Wu, T.: Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In: Proceedings of the 12th EDBT. ACM (2009)
26. Thor, A., Anderson, P., Raschid, L., Navlakha, S., Saha, B., Khuller, S., Zhang, X.N.: Link prediction for annotation graphs using graph summarization. In: ISWC. pp. 714–729 (2011)
27. Von Luxburg, U.: A tutorial on spectral clustering. Statistics and computing 17(4), 395–416 (2007)

28. Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., Hassanali, M.: Drugbank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research 36(suppl 1), D901–D906 (2008)