Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A network-theoretic approach for decompositional translation across Open Biological Ontologies

Chintan O. Patel [a,*], James J. Cimino [a,b]

[a] Department of Biomedical Informatics, Columbia University, New York, NY, United States
[b] Laboratory for Informatics Development, NIH Clinical Center, Bethesda, MD, United States

ARTICLE INFO

Article history:
Received 1 September 2009
Available online 6 February 2010

Keywords:
Network theory
Biomedical ontologies
Ontology translation
Open biomedical ontologies

ABSTRACT

Biological ontologies are now being widely used for annotation, sharing and retrieval of the biological data. Many of these ontologies are hosted under the umbrella of the Open Biological Ontologies Foundry. In order to support interterminology mapping, composite terms in these ontologies need to be translated into atomic or primitive terms in other, orthogonal ontologies, for example, *gluconeogenesis* (biological process term) to *glucose* (chemical ontology term). Identifying such decompositional ontology translations is a challenging problem. In this paper, we propose a network-theoretic approach based on the structure of the integrated OBO relationship graph. We use a network-theoretic measure, called the clustering coefficient, to find relevant atomic terms in the neighborhood of a composite term. By eliminating the existing GO to ChEBI Ontology mappings from OBO, we evaluate whether the proposed approach can re-identify the corresponding relationships. The results indicate that the network structure provides strong cues for decompositional ontology translation and the existing relationships can be used to identify new translations.

1. Introduction

Biomedical ontologies are being increasingly used in a variety of informatics applications ranging from information retrieval, decision support and knowledge discovery. The size and scope of biomedical ontologies is rapidly expanding under the Open Biological Ontologies (OBO) Foundry [1]. There are over 60 ontologies with a total of 443,440 terms and 439,417 relationships within OBO. This volume is reflective of the increased use of ontologies in annotation, sharing and analysis of molecular biology datasets.

The process of annotation generally involves instances of biological entities such as proteins, genes or phenotypes being associated with one or more ontology terms. Consider, for example, gene *IGF1*, which is associated with *chondroitin sulfate proteoglycan metabolic process* in the Gene Ontology (GO). One of the common ways to analyze data annotated with such terms is to look at the orthogonal information such as chemicals, cellular locations or anatomy associated with a given annotation. In the aforementioned example, relevant chemical terms include *chondroitin sulfate*, *proteoglycan* and *proteochondroitin sulfates* from the Chemical Entities of Biological Interest (ChEBI) ontology [2]. We refer to this process of breaking down a composite term from a given source ontology to its constituent atomic terms in a target ontology as *decompositional ontology translation*.

The problem of ontology translation has been studied in context of identifying equivalent or similar meaning terms across a pair of ontologies [3,4]. In our previous work, we proposed a graph traversal algorithm over the UMLS Metathesaurus based on clustering coefficient to perform decompositional terminology translation [5]. One of the recurring themes emerging with the growing number of ontologies (and integration thereof) is the idea of using large-scale analytic methods such as machine learning and network theory to solve ontology translation problems.

In this paper, we seek to explore whether there is sufficient knowledge in OBO ontologies, specifically relationships and cross-ontology mappings (signifying a semantic overlap), such that a network-theoretic approach can be used for decompositional terminology translation. Specifically, we study the following two questions:

a. What is the extent of semantic overlap across OBO ontologies?
b. Can the existing OBO relationships be used for decompositional ontology translation?

Towards this goal, we develop an integrated version of OBO ontologies using cross-ontology mappings and relationships from

[*] Corresponding author. Address: Department of Biomedical Informatics, Columbia University, Vanderbilt Clinic, 5th Floor, 622 West, 168th Street, New York, NY 10032, United States.
E-mail addresses: chintan.patel@dbmi.columbia.edu, chintanop@gmail.com (C.O. Patel).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.01.007

all ontologies in the OBO Foundry. Using a dataset of GO-ChEBI mappings in the OBO, we evaluate a network-theoretic graph traversal algorithm to effect decompositional ontology translation.

2. Background

2.1. Decompositional ontology translation vs. ontology mapping

A significant body of research literature [3,4,6,7] focuses on ontology mapping involving identification of equivalent or nearly synonymous terms across ontologies. Noy et al. [3] developed a semi-automated approach that matched ontology classes and slots using a combination of human input and lexico-semantic matching techniques. Bodenreider et al. [4] developed an algorithm over the UMLS Metathesaurus that mapped an arbitrary UMLS concept to a set of MeSH terms with most similar meaning by traversing specific relationship types in the Metathesaurus graph.

Decompositional ontology translation [5] is a problem of identifying constituent atomic terms for a given composite term. There are generally two or more atomic terms associated with a single source composite term. The existing methods for ontology translation generally attempt to find the target terms with closest meaning, whereas decompositional ontology translation looks for the constituent atomic terms rather than synonymous terms, such as *proteoglycan* for *chondroitin sulfate proteoglycan metabolic process.* Various lexical [8,9] methods have been developed that perform sub-string matching across a set of terms to identify the atomic components, consider for example, *adenocarcinoma* is a sub-string of the term *adenocarcinoma of the eyelid*. The lexical approaches are limited in their inability to identify constituent terms (*proteochondroitin sulfates*) not present in the original string of the composite term (*chondroitin sulfate proteoglycan metabolic process*). The morphosemantic approaches [10,11] go beyond simple string matching by analyzing the morphemes such as "*leuk*", "*hepat*", "*anti-*" and "*-itis*" found in the medical terms such as *hepatitis*, *leukemia.* In our research, we approach the problem by using the structure of the ontology graph to decompose the terms.

2.2. Network theory

Social network analysis [12] provides methods for understanding interactions and social phenomena among people, groups and organizations. The relations are the primary object of analysis; the attributes of actors are not generally considered. Recently, these methods have been successfully used to analyze other types of networks such as the Internet, World Wide Web and various biological networks [13]. The network properties provide significant insights into the structure and function of the domain. In this research, we use two network-theoretic properties: the scale-free network property and the clustering coefficient measure.

2.3. Scale-free networks

Scale-free networks [13] are networks with a specific topology in which a small number of nodes (called hubs) have many relationships and large number of nodes have only a few relationships. One of the measures to identify a scale-free network involves plotting a power law degree distribution. The power law states that the probability $p(k)$ of a given node in the network connected to k other nodes is proportional to k^{-c}, where c is generally between 2 and 3 for scale-free networks. The power law implies that a few "hub" nodes are connected to a large number of nodes and that most other nodes in the network have only a small number of connections. The integrated OBO ontologies exhibit (see Fig. 1) a strong scale-free network property as evidenced from the slope of the plot (power law constant) equal to 2.87 (the value is between 2 and 3 for scale-free networks). The top 10 hub nodes in the OBO graph are shown in Table 1. The hubs are important in traversal of OBO graph as they connect different parts of the network. They can also potentially introduce large number of non-relevant target nodes during the traversal [14].

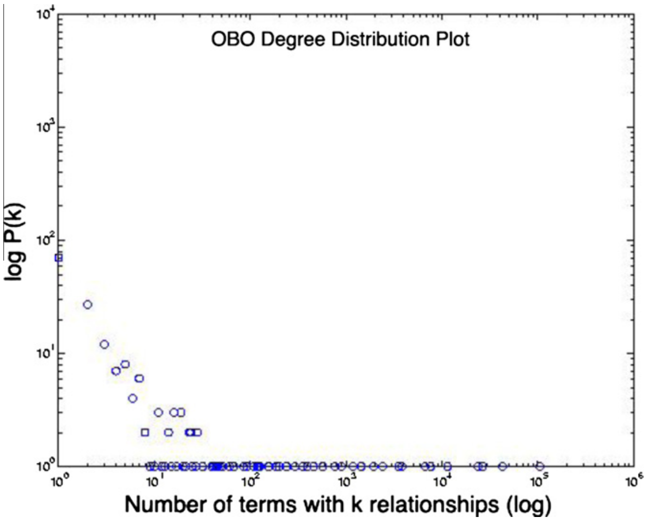

Fig. 1. The power-law distribution, $p(k) = k^{-c}$, where k is the degree of terms in the OBO relationship graph. The constant c is 2.87 indicating a scale-free network topology.

2.4. Clustering coefficient

The clustering coefficient [15] is a network analysis measure used to quantify the 'connectedness' of the neighborhood of a given node. The clustering coefficient is a ratio of the number of edges between the neighbors of a given node and the total possible number of edges among all its neighbors. To calculate the clustering coefficient (CC) for a given node n, let the degree (or number of immediate neighboring nodes) of n be k and let t be the total number of edges between the neighboring nodes, then

$$CC(n) = \frac{t}{k^{*}(k-1)},$$

i.e., the clustering coefficient is the ratio of number of edges between the neighbors of n and the total possible number of edges between the neighbors (if each neighbor was connected to every other neighbor). High (low) clustering coefficient indicates densely (sparsely) connected neighborhood. For example, the term *metabolic process* has 1027 relationships among its 882 unique neighbor terms, so its clustering coefficient is 1027/38,5521 = 0.0026. This

Table 1
The top 10 terms in OBO sorted by the descending number of relationships.

OBO term	Source ontology	Number of relationships
Regulation	Gene ontology	2869
Adult	Zebrafish anatomy	1480
Unknown	Zebrafish anatomy	1188
Protein complex	Gene ontology	1117
Embryonic cell	WBbt	916
Metabolic process	GO, MP, WBPhenotype	907
Cell	WBbt	890
Post-embryonic cell	WBbt	782
Catabolic process	Gene ontology	640
Anabolism	Gene ontology	638

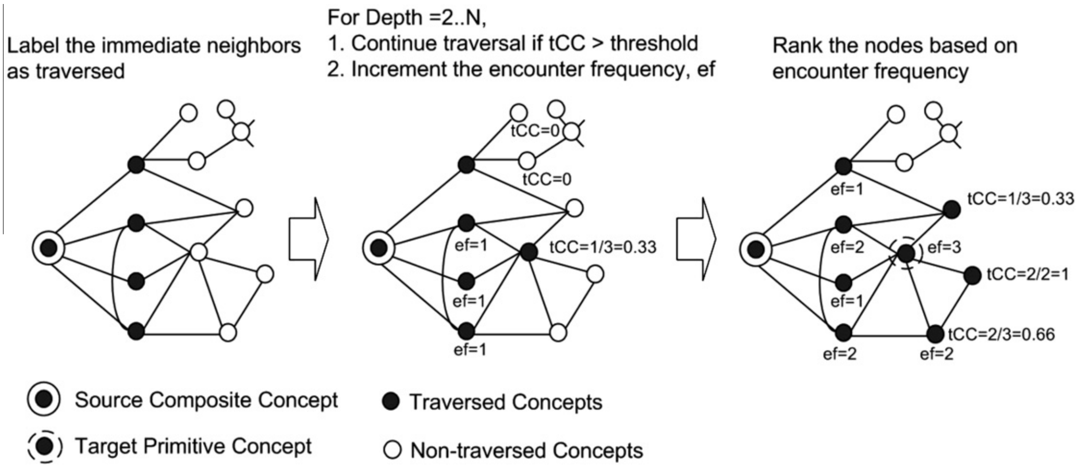

Fig. 2. An illustration showing how the TClustN algorithm favors more closely clustered neighborhoods over sparsely connected regions of graph.

measure is used to limit the traversal of the OBO graph as described in next section.

2.5. TClustN algorithm

In scale-free networks, traversing indirect edges for a given source node generally produces a large number of paths since hubs bring together different parts of the network. Using a cross-ontology traversal of indirect relationships in a scale-free network, such as the UMLS Metathesaurus or OBO, produces a large number of possible paths (which then have to be filtered in order to find relevant translations). To overcome this problem, in our previous work, we proposed an approach based on the clustering coefficient to limit the traversal around the closely connected neighborhood of a given source node. The key steps (Fig. 2) in the algorithm are:

1. For a given source node, traverse the outgoing edges transitively until a given depth, *D*.
2. At each next node in traversal step, calculate the traversal-based clustering coefficient and stop further traversal if the value is less than a given threshold, *T*.
3. Calculate the encounter frequency of each node during the traversal and use it to rank the target nodes in descending order.

Consider for example, that the source term *chondroitin sulfate proteoglycan metabolic process* has five direct relationships to other terms such as *metabolic process* (clustering coefficient of 0.0026) and *chondroitin sulfate proteoglycan anabolic process* (clustering coefficient of 0.18). If the threshold parameter, *T* is set to 0.01, then *metabolic process* will be eliminated from further traversal. A detailed description and rationale of algorithm can be found in our previous paper. The key idea behind the algorithm is to remove the effect of hubs or noisy nodes that are not within the semantic locality of a given node. In this paper, we evaluate this algorithm over OBO ontologies to perform decompositional ontology translation.

3. Methods

To evaluate the proposed network-theoretic approach over biological ontologies, we prepared an integrated OBO dataset using the existing cross-ontology mappings in the OBO Foundry. As a computationally-derived gold standard, we used the existing decompositional ontology translations between GO and ChEBI to evaluate the proposed TClustN algorithm based on a precision-recall measure (Fig. 3).

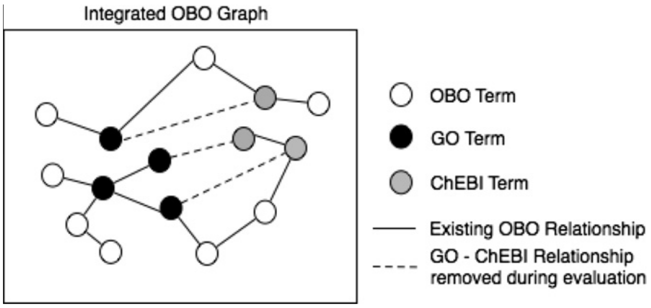
Integrated OBO Graph
OBO Term
GO Term
ChEBI Term
Existing OBO Relationship
GO - ChEBI Relationship removed during evaluation

Fig. 3. The integrated OBO graph is prepared by using the cross-ontology mappings in the OBO Foundry. The existing decompositional relationships from GO to ChEBI are eliminated from the OBO graph. The TClustN algorithm is used over the composite source GO terms to re-identify the atomic chemical terms in ChEBI.

3.1. Data preparation

1. *Integrated OBO Dataset:* From the OBO Foundry server,[1] we downloaded (and were able to parse) 50 ontology files and 44 ontology mapping files in the OBO format. We used the OBO Java API [16] to parse the following fields from the files: OBO ID, synonym, xref (exact reference) field, and the relationships (including ISA). The xref field was used to create an OBO Concept Unique Identifier (OBOCUI) across equivalent OBO IDs. The original relationships asserted across OBO IDs were re-modeled across OBOCUIs. A mapping table for OBOCUI to synonyms was also created.

2. *GO-ChEBI Test Dataset:* To evaluate the TClustN algorithm over the integrated OBO dataset, we used the existing 5996 GO-to-ChEBI mappings available in GO_to_ChEBI.obo file as the computationally-derived gold standard. We removed a subset (test set) of these mappings from the relationship table and applied the TClustN algorithm for the source GO terms (in the removed subset) to evaluate whether it can identify the corresponding target ChEBI terms. For example, an existing mapping from *chondroitin sulfate proteoglycan metabolic process* (GO) to *proteochondroitin sulfates* (ChEBI) is removed from the integrated OBO relationship graph to evaluate whether TClustN can re-identify the mapping. We developed three different versions of test datasets, G2C_20P, G2C_40P and G2C_80P, by randomly eliminating 20%, 40% and 80%, respectively,

[1] http://www.obofoundry.org/ [accessed 10.02.09].

of all GO-to-ChEBI mappings from the original OBO relationship graph. By eliminating a varying proportion of relationships, the goal was to evaluate the effect of existing knowledge (in form of relationships or mappings) towards discovery of future cross-ontology mappings.

3.2. Experiment steps

1. A binary sparse-matrix representation was used to represent the OBO relationship graph where each OBOCUI corresponded to a row or a column. The Java MTJ library [17] was used to implement the sparse in-memory matrix.
2. The TClustN algorithm was executed over the source concepts in the different GO-ChEBI datasets (with corresponding relationship matrices without the test dataset mappings). The TClustN parameters for depth, D was set to 3 and clustering coefficient threshold, T, was set to 0.01.
3. For baseline comparison, the TClustN algorithm was executed over G2C_20P with clustering coefficient threshold, $T = 0$. This is equivalent to simple transitive traversal of the OBO relationship graph.
4. To evaluate the performance of the algorithm, average top k precision and recall were calculated across all datasets as defined below:

$$k\text{-precision} = \frac{\text{True positives in top}k}{\text{Total terms in top}k},$$

$$k\text{-recall} = \frac{\text{True positives in top}k}{\text{Total relevant}},$$

where k refers the position of the target term in the sorted results of TClustN. A target term in the results is considered as true positive if the source ontology of the term is ChEBI. The total relevant terms refers to the number of ChEBI terms related to the given source GO term in the gold standard.

The top $k = 1$, 2 and 3 positions were used to calculate the true positives in three separate analyzes.

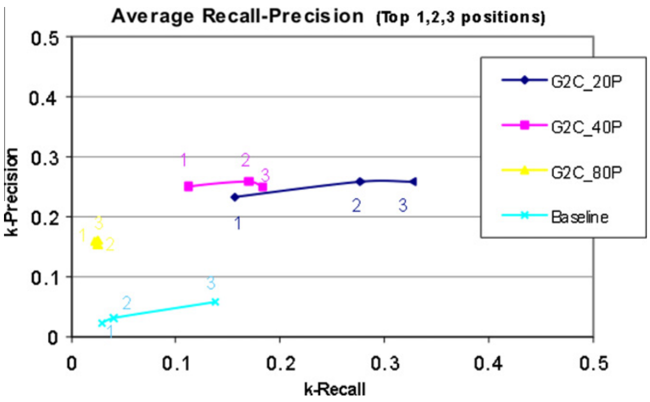

Fig. 4. The results of average precision-recall for the top 1, 2 and 3 positions across the datasets and parameter settings. The clustering coefficient based results are significantly higher than the baseline. The performance decreases as we eliminate higher proportion of existing mappings.

4. Results

We found 259,865 OBOCUIs across the total 439,417 OBO terms, indicating a 59.13% of overlap in meaning across the OBO ontologies under study. We found 144 distinct relationship types (*is_a*, *part_of*, *regulates*, *unit_of*, etc.) across all OBO ontologies. There were 895, 1777 and 2808 unique source OBOCUI in the 20%, 40% and 80% test datasets, respectively. The results of average precision and recall are shown in Fig. 4. The baseline result of simple transitive traversal was significantly lower than all other results indicating the usefulness of using network-theoretic measure of clustering coefficient towards ontology translation. The decreasing precision and recall upon removing higher proportions of GO-ChEBI mappings indicated the importance of the existing mappings in enabling identification of new translations. A sample of the results for the decompositional translation is shown in Table 2. Consider for example, the term *chondroitin sulfate proteoglycan metabolic process* was decomposed into relevant chemical terms *proteoglycan*, *chondroitin sulfate* and *proteochondroitin sulfates* (the term with gold standard mapping).

5. Discussion

The knowledge sources such as OBO ontologies have a rich set of terms, relationships and mappings largely created by costly manual processes. Automated methods can potentially use this existing knowledge to reduce the cost of manual knowledge engineering. Towards this direction, we have developed an automated method based on network theory to perform decompositional cross-ontology translations across OBO ontologies.

Understanding the network-based structure of the ontologies can provide a strong cue towards performing cross-ontology translation. As discussed earlier, the OBO ontologies exhibit scale-free properties with presence of hub concepts. Our results support this network property, for example, as shown in Table 2 under Baseline results, the top two concepts (*monoterpenoids*, *monoterpenes*) are hubs in ChEBI that are connected to many other nodes. The threshold parameter for clustering coefficient in TClustN eliminates such hub nodes as they have very small connectedness among their neighbors.

What is the critical number of relationships required to use a network-based method? This is an important question that we

Table 2
A sample of two source GO terms and top three target ChEBI terms using TClustN with and without clustering coefficient threshold. The baseline analysis produced hubs terms (*monoterpenoids*, *monoterpenes*) as top results that got eliminated with the threshold of clustering coefficient. The true positives are shown in bold. The algorithm also suggested other relevant atomic terms (*proteogylcan*, *chondroitin sulfate*) outside of the gold standard.

Source Gene Ontology term	Top 3 Target ChEBI terms	
	Clustering coefficient threshold = 0.01	Baseline (simple traversal), clustering coefficient threshold = 0
Chondroitin sulfate proteoglycan metabolic process	1. Proteoglycan 2. Chondroitin sulfate 3. Proteochondroitin sulfates	Monoterpenoids Monoterpenes Sulfur molecular entities
Purine deoxyribonucleoside diphosphate metabolic process	1. Purine 2-deoxyribonucleoside diphosphates 2. Purine 2-deoxyribonucleotides 3. Purine nucleoside diphosphates	Monoterpenoids Monoterpenes Pyrimidine 2-deoxyribonucleotides

attempted to answer by eliminating different proportions of the total number of relationships from the original OBO relationship graph. The results indicate a continuum of decrease in precision-recall as we remove higher percentage of existing relationships. We also observed that precision is relatively stable as compared to recall indicating that results obtained by algorithm are more accurate if not necessarily complete.

In our previous research [5], we had successfully applied TClustN over the UMLS Metathesuarus that has significantly higher number of relationships than the integrated version of OBO ontologies used in this paper. Nevertheless, the precision-recall results over OBO are comparable to the UMLS study indicating a similar network structure (scale-free) of both resources. Furthermore, it shows the domain-independent nature of the algorithm towards decompositional ontology translation.

In comparison to existing ontology translation methods [4,9], the precision-recall values obtained in our study were lower by 20–30%. The difference is significant and can be attributed to the use of domain specific properties by existing translation methods such as UMLS synonymy and semantics of relationships. Our network theory based approach is complementary that can be extended to include additional domain specific information to influence the traversal and ranking.

5.1. Limitations

The proposed approach is dependent on the traversal of cross-ontology relationships. This limits the applicability of the methods to ontologies with high semantic overlap with other ontologies in the integrated OBO graph. Further the proposed method is based on the structure of the ontology graph that resulted in lower precision and recall as compared to the existing methods using domain specific lexical, semantic or morphosemantic properties [11] of the terms and relationships.

5.2. Applications

The biomedical applications of the decompositional ontology translation go beyond data analysis. One important effort of the OBO Foundry is to formally define GO and other ontologies with logic based definitions based on primitive concepts. Consider for example, a complex term *chondroitin sulfate proteoglycan metabolic process* can be defined based on primitive chemicals concepts (*proteoglycan*, *chondroitin sulfate*) and the processes (*metabolic process*). Such definitions based on primitive concepts provide several benefits for ontology maintenance, ontology alignment and automated reasoning.

5.3. Future research implications

The use of a network-based feature (clustering coefficient) can enable terminology translation across the different ontologies in the OBO and the UMLS. Using such generic features that are intrinsic to the OBO and UMLS provide a powerful new mechanism to identify terminological links or relationships for different applications. The features can be learned to predict the new links or relationships using an existing training dataset from the application domain.

6. Conclusion

A network-theoretic approach was presented for decompositional ontology translation over the OBO ontologies. An algorithm based on clustering coefficient was used to identify relevant terms in target ontology. An integrated version of OBO ontologies was prepared and evaluated against a test dataset based on GO-to-ChEBI mappings. The results indicate that network structure provides a strong cue to perform decompositional ontology translation and the existing set of relationships in OBO can be used to identify new translations. The results were comparable to our previous UMLS study indicating the domain-independent nature of the algorithm.

Acknowledgments

We thank the anonymous reviewers of AMIA 2009 Annual Symposium for their insightful comments and the National Library of Medicine for their grant support (R21LM009638) that funded this research. Dr. Cimino is supported by funds from the intramural research program at the NIH Clinical Center and the National Library of Medicine.

References

[1] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25(11):1251–5.
[2] Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, et al. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res 2008;36(Database issue):D344–50.
[3] Noy N, Musen M. PROMPT: Algorithm and tool for automated ontology merging and alignment. In: Proceedings of the seventeenth national conference on artificial intelligence (AAAI-2000), Austin, TX.
[4] Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. In: Proceedings of the AMIA symposium; 1998. p. 815–9.
[5] Patel CO, Cimino JJ. Decompositional terminology translation using network analysis. In: AMIA annual symposium proceedings 2007 October 11. p. 588–92.
[6] Dinakarpandian D, Tong T, Lee Y. A pragmatic approach to mapping the open biomedical ontologies. Int J Bioinform Res Appl 2007;3(3):341–65.
[7] Johnson HL, Cohen KB, Baumgartner WA Jr, Lu Z, Bada M, Kester T, et al. Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. In: Pac symp biocomput; 2006. p. 28–39.
[8] McCray AT, Browne AC. Discovering the modifiers in a terminology data set. In: Proceedings of the AMIA symposium; 1998. p. 780–4.
[9] Barac'h V, Burgun A, Delamarre D, Le Beux P. A method for an automated decomposition of a complex term into its conceptual components. Stud Health Technol Inform 1999;68:875–80.
[10] Dujols P, Aubas P, Baylon C, Grémy F. Morpho-semantic analysis and translation of medical compound terms. Methods Inf Med 1991;30(1):30–5.
[11] Namer F, Baud R. Defining and relating biomedical terms: towards a cross-language morphosemantics-based system. Int J Med Inform 2007;76(2–3):226–33.
[12] Wasserman S, Faust K. Social networks analysis: methods and applications. United Kingdom: Cambridge University Press; 1994.
[13] Barabasi A, Albert R. Emergence of scaling in random networks. Science 1999;286(5439):509–12.
[14] Patel CO, Cimino JJ. A scale-free network view of the UMLS to learn terminology translations. Stud Health Technol Inform 2007;129(Pt 1):689–93.
[15] Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature 1998;393(6684):440–2.
[16] Day-Richter J, Harris MA, Haendel M, Gene Ontology OBO-Edit Working Group, Lewis S. OBO-Edit – an ontology editor for biologists. Bioinformatics 2007;23(16):2198–200. [Epub 2007 Jun 1].
[17] Matrix Toolkits for Java (MTJ), http://ressim.berlios.de/ [accessed 9.03.09].