Journal of Southeast University (English Edition) Vol. 26, No. 1, pp. 31 – 35 Mar. 2010 ISSN 1003—7985

Discovering hidden information of gene ontology based on complex networks analysis

Tang Jintao[1] Wang Ting[1] Wang Ji[2]

([1]School of Computer, National University of Defense Technology, Changsha 410073, China)

([2]National Laboratory for Parallel and Distributed Processing, National University of Defense Technology, Changsha 410073, China)

Abstract: To resolve the ontology understanding problem, the structural features and the potential important terms of a large-scale ontology are investigated from the perspective of complex networks analysis. Through the empirical studies of the gene ontology with various perspectives, this paper shows that the whole gene ontology displays the same topological features as complex networks including "small world" and "scale-free", while some sub-ontologies have the "scale-free" property but no "small world" effect. The potential important terms in an ontology are discovered by some famous complex network centralization methods. An evaluation method based on information retrieval in MEDLINE is designed to measure the effectiveness of the discovered important terms. According to the relevant literature of the gene ontology terms, the suitability of these centralization methods for ontology important concepts discovering is quantitatively evaluated. The experimental results indicate that the betweenness centrality is the most appropriate method among all the evaluated centralization measures.

Key words: gene ontology; complex network analysis; centrality measure

An ontology is an explicit shared specification of the conceptualization of a domain [1], which has attracted a surge of researches and has been used in many disciplines. In recent years, more and more ontologies have been published and widely used in the web. As a result, the scale and the complexity of the ontology are increased rapidly, which require more prior knowledge and efforts to understand. The gene ontology(GO) is a widely used ontology in bioinformatics which is appropriate to unify the representation of gene and gene product attributes across all species[2]. However, with the rapid development of bioinformatics, the tremendous scale of the gene ontology has become one of the most common obstacles for its understandability and usability. So it is necessary to study the methodology of ontology understanding[3] to enhance the usability of large-scale complex ontologies such as the gene ontology.

Complex network analysis has been a hot research area in recent years. In the empirical studies of real-world networks, some topological features that never occur in simple networks such as lattices or random graphs have been observed, for example, six degrees[4], scale-free[5] and so on. All the real-world networks that have such non-trivial features are named as complex networks. The topological features of complex networks are further studied. Thus many network properties and analysis methods have been proposed and widely used in many real networks applications, such as link analysis[6], biological network analysis[7] and so on. Large-scale ontologies are developed and maintained by experts; the network representations of such ontologies may also have some non-trivial features. So it is natural to investigate the structure and important concepts/relations by utilizing the complex network analysis methods. According to this idea, this paper studies the topological features of the gene ontology insight from the complex network analysis and utilizes various networks centralization methods to discover the potential important terms in the gene ontology. Furthermore, this paper retrieves the relevant biological papers in the MEDLINE database to evaluate whether the discovered important terms are really widely studied by biologists.

1 Discovering Complex Network Features of Gene Ontology

1.1 Graph representation of ontology

An ontology defines the concepts, individuals and relations in a certain domain; it is natural to represent the concepts/individuals as vertices and relations as directed edges. Given ontology O, the corresponding graph representation G is defined as follows:

Definition 1 A directed graph $G = (V, E)$ is a corresponding graph of a given ontology O, where V is the set of vertices representing all the terms in O, and E is the set of edges representing all the relations.

Definition 1 is suitable for simple ontologies. However, the gene ontology is more complex. Unlike the simple ontology, the gene ontology contains three domains: a cellular component, a biological process, and a molecular function. So the gene ontology can be referred to three ontologies or an ontology consisting of three sub-ontologies. Whether the sub-ontologies have the same topological features as the gene ontology is also needed to be investigated. Furthermore, there are three types of links in the gene ontology such as is_a, part_of and regulates. Different types of links represent different semantics of the relationships between the terms, which suggests that the sub-graphs only containing a certain type of links may have some distinct structural features. For a comprehensive analysis of the gene ontology, this paper defines the concept of view for a given ontology.

Definition 2 A view of a graph $G = (V, E)$ is a directed sub-graph $V_G = (V_v, E_v)$ of G, where $V_v \subseteq V$, $E_v \subseteq E$. Each vertex and edge in V_G is selected based on its semantic/domain.

In this paper, seven views of the gene ontology are gener-

Received 2009-08-30.

Biographies: Tang Jintao(1981—), male, graduate; Wang Ting(corresponding author), male, doctor, professor, tingwang@nudt.edu.cn.

Foundation items: The National Basic Research Program of China (973 Program) (No. 2005CB321802), Program for New Century Excellent Talents in University (No. NCET-06-0926), the National Natural Science Foundation of China (No. 60873097, 90612009).

Citation: Tang Jintao, Wang Ting, Wang Ji. Discovering hidden information of gene ontology based on complex networks analysis[J]. Journal of Southeast University (English Edition), 2010, 26(1): 31 – 35.

ated from different perspectives. These views include the sub-graphs only containing a certain domain, the sub-graphs only containing a certain type of relationship, and the graph of the whole gene ontology. The statistical information and the topological features are shown in Tab. 1 and discussed in section 1.2.

Tab. 1 The statistic information of different views of the gene ontology graph

Graph	Vertices	Edges	Diameter	Average distance	Max degree	α	R^2
Gene ontology	27 650	48 425	15	3.78	574	2.54	0.957
Biological process	16 667	33 867	15	3.74	145	2.62	0.927
Cellular component	2 366	4 531	10	3.16	524		
Molecular function	8 602	10 024	11	3.28	263	2.07	0.958
Is_a relation	27 636	40 327	13	3.77	524	1.72	0.911
Part_of relation	5 262	4 347			37		
Regulates relation	5 185	3751			3		

1.2 Topological features of gene ontology

As shown in Tab. 1, the diameter of the whole gene ontology network is 15, while the average distance is only 3.78. The diameters of the sub-graphs of the three domains are also comparatively small. For example, the diameter of the molecular function sub-ontology is 11, and the average distance of this big graph containing over 8 000 vertices is only 3.28. So it is believable that the gene ontology and its three sub-ontologies obviously show the "small-world" feature. However, the views based on relations do not display the same feature. These views except the is_a relation based view are divided into more than 1 000 components and isolated vertices. So we consider that the views based on relations do not display the small-world feature, and also do not correctly reflect the nature of the gene ontology because of the lack of connectivity.

The scale-free feature indicates the fact that the structural feature and the dynamics of the networks are independent of the scale. A distinguishing characteristic of scale-free is the distribution of degrees following a power-law, which suggests that there exist a few active nodes in the network connected with many edges and there exist many common nodes only connected with a few edges. For example, an empirical study of the Internet indicates that the in-links of most pages are no more than 4, while 0.01% pages occupy over 80% of the in-links[5].

The power-law distribution suggests that the probability that a node is connected to other nodes k is $p(k) = Ck^{-\alpha}$, where C is a constant parameter, and α is the exponent of the power-law.

Fig. 1(a) shows the degree distributions of the whole gene ontology. We can learn from this figure that the degrees of the gene ontology follow a power-law distribution. The dotted curve shows the best power-law curve that is most in accord with the degree distribution of the gene ontology. It also illustrates the scale-free feature of the gene ontology. As listed in Tab. 1, the exponent α of the fitted curve is 2.54, which follows the rule that the exponent of most real-world complex networks ranges from 2 to 3[5]. The corresponding determination coefficient R^2 is 0.957, which demonstrates the high quality of the fitting cure on the distribution of data.

However, the degree distributions of the sub-ontologies except for those of the molecular function do not obviously follow the power-law. As shown in Fig. 1(b), the degree distribution of the biological process sub-ontology is not in accordance with the best fitting curve, especially at the small degrees. The corresponding determination coefficient R^2 shown in Tab. 1 also indicates the quality of the fitting curve is not so good as that in Fig. 1(a).

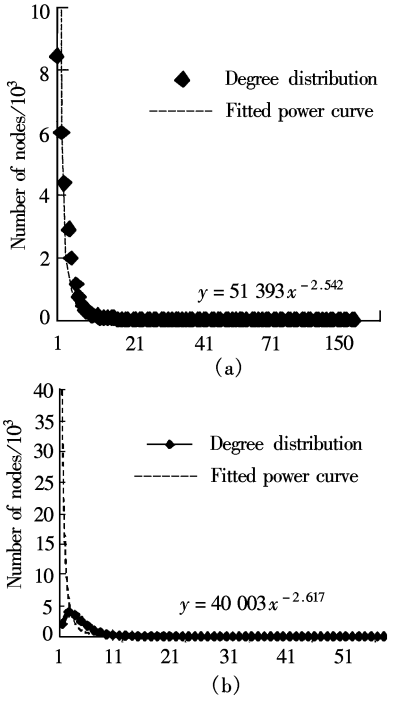

Fig. 1 Degree distribution of the gene ontology graph. (a) Gene ontology; (b) Biological process sub-ontology

In conclusion, the graph representation of the whole gene ontology is a complex network, which displays the same topological features since complex networks include "small-world" and "scale-free". Three sub-ontologies show obvious small-world features as being the same as the whole ontology. However, the degree distributions of the sub-ontologies except for those of the molecular function do not obviously follow the power-law. Furthermore, the views from the link types may not correctly reflect the nature of the gene ontology because of the lack of connectivity.

2 Finding Important Terms

How to discover the potential important concepts in a large-scale ontology is important for ontology understanding. Sociologists and mathematicians have proposed a surge of complex network centrality measures, such as closeness centrality and so on. In this section, we adopt and evaluate these centrality measures to discover the potential important terms of the gene ontology.

2.1 Centrality measures

In the social network analysis area, there are two well-known measures for evaluating the importance of nodes: the betweenness centrality, the proportion of all shortest paths in the network that run through a given node, and the closeness centrality, the average distance from the given node to every other node in the network[8]. Sociologists have also proposed a simple but effective measurement to evaluate the importance of nodes at local environments and local centrality[9], which evaluate the importance of nodes according to the nodes' degrees.

For a given graph $G=(V, E)$ with n vertices, the local centrality $C_{\mathrm{D}}(v)$ for vertex v is defined as

$$C_{\mathrm{D}}(v)=\frac{\deg(v)}{n-1}$$

where $\deg(v)$ is the degree of v.

The closeness centrality $C_{\mathrm{C}}(v)$ for vertex v is defined as

$$C_{\mathrm{C}}(v)=\frac{1}{\sum_{t\in V\setminus v}\mathrm{dis}(v, t)}$$

where $\mathrm{dis}(v, t)$ is the geodesic distance from vertex v to t.

The betweenness centrality $C_{\mathrm{B}}(v)$ for vertex v is defined as

$$C_{\mathrm{B}}(v)=\sum_{s\neq t\neq v}\delta_{st}(v),\ \delta_{st}(v)=\frac{\partial_{st}(v)}{\partial_{st}}$$

where ∂_{st} is the number of the shortest paths from vertex s to t, and $\partial_{st}(v)$ is the number of the shortest paths from s to t that pass through v.

In the information retrieval research area, many algorithms based on hyperlinks have been proposed to rank the importance of web pages, such as PageRank[10], HITS[11] and so on. Recently, these algorithms have also been widely used in complex network analysis applications. In this paper, we evaluate the effectiveness of the HITS algorithm for discovering the important terms in the gene ontology. The HITS algorithm is an iterative algorithm that exploits a mutual reinforcing relationship between hub pages and authority pages. The hub score and authority score for a vertex is calculated by the following steps[11]:

1) Assign each vertex with the same hub score and authority score, and usually the score is 1;

2) Update each vertex's authority score to be equal to the sum of the hub score of the neighbors that points to it;

3) Update each vertex's hub score to be equal to the sum of the authority score of the neighbors that it points to;

4) Normalize the values of the hub scores and authority scores;

5) Repeat 2) for k times.

2.2 Measure selection

The aforementioned centrality measures depict the potential structural important vertices in different ways. For instance, the local centrality measures the importance of vertices at local environments, while the betweenness centrality measures the importance of vertices based on the contributions on the graph connectivity. Which measure is more appropriate for discovering the real important terms in the gene ontology at the semantic level is an interesting problem. We develop a centrality measure evaluation strategy to evaluate these methods in two aspects: the effectiveness and the efficiency.

To evaluate the effectiveness of the aforementioned centrality measures, we turn to get help from the MEDLINE[12] database. MEDLINE is a literature database of life sciences and biomedical information, which covers most literature in biology and biochemistry, and other fields such as molecular evolution and so on. As the most famous and the biggest scientific literature database, MEDLINE contains more than 18 million records from more than 5 000 selected publications since the 1950s. The database is freely accessible via the PubMed information retrieval interface.

It is clear that the most attention-getting concepts are probably important concepts in the domain. So this paper utilizes the number of related scientific papers to evaluate the importance of gene terms. For a given gene term, this paper retrieves the relevant papers in MEDLINE that mentioned the term in title or abstract. The number of papers is an objective and reasonable measure to evaluate whether a concept is more important than another. We use the aforementioned centrality measures to rank the importance of the terms in the gene ontology. For each ranking list, we select the first ten terms with the highest value to count their relevant papers in MEDLINE. We also randomly select ten terms for comparative study.

The gene ontology is a large-scale ontology with more than 27 000 terms; the execution time of the centrality measures is an important factor that should be considered. We investigate the computational complexity of the aforementioned algorithms to consider that whether these algorithms are tractable when running on the gene ontology. If the quality of ranking results is almost at the same level, we consider the algorithm that costs less time is more suitable for large-scale ontology centralization. Tab. 2 shows the computational complexity of the aforementioned measures for a given graph with n vertices and e edges.

Tab. 2 The complexity of the centrality measures

Measure	Complexity
Local centrality	$O(e)$
Closeness centrality	$O(n^3)$
Betweenness centrality	$O(n^2\log n+ne)$
HITS Algorithm	$O(kn^2)$

As shown in Tab. 2, the computational complexity of the local centrality only relies on the number of edges e, which is much smaller than that of other methods. The complexity of the HITS algorithm relies on the square number of vertices n and the number of iterative steps k. However, the betweenness centrality and the closeness centrality are much more extortionately and computationally expensive. These two measures need more than several minutes to rank the importance of all the vertices on large-scale ontologies such as the gene ontology, which is unacceptable for many real applications. To avoid this problem, we adopt the approximate centrality analysis methods based on CDZ shortest paths approximation[13], which is especially effective and efficient for complex network analysis. With this approxi-

mation, the computational complexity of both the closeness centrality and the betweenness centrality measures can be reduced to $O(n\log n + e)$.

2.3 Experiments

At first, the four mentioned important concepts ranking methods are executed on the gene ontology and the first ten terms with the highest ranking score are obtained. The top 10 most important terms of these methods are shown in Tabs. 3 to 6. Then these potential important terms are reviewed by domain experts to evaluate the semantic coverage of the gene ontology. For each ranking list given by different methods, three groups of 10 randomly selected terms not in top 500 are used to compare with the top 10 terms discovered by those centrality measures. These terms are used to query the MEDLINE, and then the number of relevant papers is summed by group. The average number of relevant papers of three random groups and the number of relevant papers of the top 10 terms are then used to evaluate the correctness of the discovered elements. The comparative results are shown in Fig. 2.

Tab. 3 The top 10 most important terms based on betweenness centrality

Id	Terms	Papers
9467	Biological process	1 614
8754	Catalytic activity	17 039
5974	Molecular function	867
2063	Protein modification process	1 456
24097	Cellular process	743
23044	Metabolic process	812
5992	Regulation of biological process	155
11253	Cellular protein metabolic process	48
18880	Transfer activity	879
8446	Regulation of cellular process	363

Tab. 4 The top 10 most important terms based on closeness centrality

Id	Terms	Papers
9467	Biological process	1 614
24097	Cellular process	743
5974	Molecular function	867
23044	Metabolic process	812
5992	Regulation of biological process	155
24027	Developmental process	786
11203	Response to stimulus	2 784
4967	Biological regulation	1 532
21507	Immune system process	452
16032	Cellular metabolic process	137

Tab. 5 The top 10 most important terms based on HITS algorithm

Id	Terms	Papers
20969	Membrane coat	155
9716	Respiratory chain complex Ⅲ	205
22220	Proton-transporting ATP synthase	266
8653	NADH dehydrogenase complex	132
12413	APC-IQGAP1-Cdc42 complex	42
3215	Acetate CoA-transferase complex	13
23096	Biotin carboxylase complex	22
8271	Junctional membrane complex	167
20022	Karyopherin docking complex	16
1499	Nup107-160 complex	6

Tab. 6 The top 10 most important terms based on local centrality

Id	Terms	Papers
19916	Protein complex	747
10531	Oxidoreductase activity, acting on the CH-OH group of donors	0
20831	Cytoplasmic part	254
20845	Plasma membrane part	195
24471	Anatomical structure development	61
10618	Oxidoreductase activity, acting on paired donors	0
10735	Hydro-lyase activity	209
20812	Intracellular part	642
10294	Kinase activity	1 255
6649	S-adenosylmethionine-dependent methyltransferase activity	86

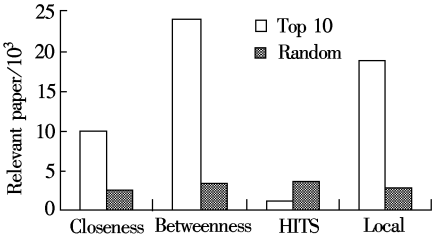

Fig. 2 Number of retrieval papers of the most important terms given by different measures

As shown in Fig. 2, the relevant papers in MEDLINE of the gene ontology terms suggest that the betweenness centrality and the local centrality are more suitable for potential important terms discovering than the other two algorithms. The relevant papers of the top 10 most important terms with the highest betweenness/local centrality value are much more than that of random selected 10 terms. The important terms discovered by the closeness centrality also display a similar phenomenon but not so dramatically. However, the relevant papers of the important terms discovered by HITS indicate that the importance ranking based on HITS does not reflect the popularity of the gene ontology terms in MEDLINE.

However, based on the human evaluation, most of the top 10 terms discovered by the local centrality are not the basic concepts of the gene ontology, such as the term "6649" and "10531". Furthermore, these top 10 terms cannot describe the main semantics of the gene ontology. For instance, most of these terms are in the domains of the cellular component and the molecular function, while there are scarcely any terms in the domain biological process which is regarded as one of most important concepts. The human evaluation suggests that the top 10 terms in Tab. 3 and Tab. 4 are more suitable to represent the outline of the gene ontology than the top 10 terms in Tab. 6. The terms, especially in Tab. 3, include the most basic concepts in the gene ontology and represent the major semantics of the gene ontology, which suggests that the importance ranking list generated by the betweenness centrality are more reasonable.

According to the human evaluation and the number of relevant papers, this paper suggests that the betweenness centrality is the best important terms discovering method among the aforementioned methods. To overcome the complexity limitation, this paper suggests adopting the shortest path approximate algorithms such as CDZ. If the scale of

the ontology is too large to run the betweenness centrality method, the local centrality which can obtain a good balance between effectiveness and efficiency is suggested.

However, the gap between the number of relevant papers for important terms and those for randomly selected terms is not sufficient enough. For instance, the random selected terms can obtain even more than 20 000 relevant papers when evaluating closeness centrality. The human review of these terms also indicates that some terms are not so important to the semantics of the gene ontology. Because the gene ontology is more of a hierarchy structure than a network, while the centrality measures are proposed to discover the important vertices in a network structure. Furthermore, the importance of the ontology terms relies more on the semantics than the structures, which indicates that the discovering methods only taking the structure information into account may be not sufficient. So it is necessary to find more reasonable importance ranking based on both the topological features and the semantics of the ontology.

3 Conclusion

This paper investigates the structural features and the potential important terms of large-scale ontology from the perspective of complex networks analysis. Through the empirical studies of the gene ontology with various perspectives, this paper shows that the whole gene ontology displays the same topological features as complex networks including "small world" and "scale-free", while some sub-ontologies have the "scale-free" property but no "small world" effect. This paper also adopts and evaluates some centralization methods to discover the potential important elements of the gene ontology. According to the relevant papers of a given gene ontology term in MEDLINE, this paper evaluates which centralization method is more suitable for ontology important concepts identification. The experimental results indicate that the betweenness centrality is the best method among the evaluated centralization measures. As future work, we plan to focus on the further study of potential important terms discovering based on both the complex network features and the semantic of ontology to obtain a more reasonable importance ranking of concepts on the large-scale ontology.

References

[1] Uschold M, Gruninger M. Ontologies: principles, methods and applications [J]. *Knowledge Engineering Review*, 1996, **11**(2): 93 - 136.

[2] The Gene Ontology Consortium. The gene ontology project in 2008 [J]. *Nucleic Acids Research*, 2008, **36**(database issue): 440 - 444.

[3] Bontas E P, Mochol M. Towards a cost estimation model for ontology engineering [C]//*Berliner XML Tage*. Berlin, Germany, 2005: 153 - 160.

[4] Milgram S. The small world problem [J]. *Psychology Today*, 1967, **2**(1): 60 - 67.

[5] Barabasi A L, Albert R. Emergence of scaling in random networks [J]. *Science*, 1999, **286**(5439): 509 - 511.

[6] Benoit G. Link analysis: an information science approach [J]. *Journal of the American Society for Information Science and Technology*, 2006, **57**(13): 1855 - 1858.

[7] Van Someren E P, Wessels L F A, Backer E, et al. Genetic network modeling [J]. *Pharmacogenomics*, 2002, **3**(4): 507 - 525.

[8] Freeman L C. Centrality in social networks: conceptual clarification [J]. *Social Networks*, 1979, **1**(3): 215 - 239.

[9] Neminen V. On centrality in a graph [J]. *Scandinavian Journal of Psychology*, 1974, **15**(1): 332 - 336.

[10] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine [J]. *Computer Networks*, 1998, **30**(1): 107 - 117.

[11] Kleinberg J M. Authoritative sources in a hyperlinked environment [J]. *Journal of ACM*, 1999, **46**(5): 604 - 632.

[12] Brian S, Katcher. *MEDLINE: a guide to effective searching in PubMed and other interfaces* [M]. 2nd ed. San Francisco: Ashbury Press, 2006: 1 - 136.

[13] Tang Jintao, Wang Ting, Wang Ji, et al. Efficient social network approximate analysis on blogosphere based on network structure characteristics[C]//*Proceedings of the Third ACM SNA-KDD Workshop*. Paris, France, 2009: 55 - 62.

利用复杂网络分析方法研究基因本体隐藏结构信息

唐晋韬[1] 王 挺[1] 王 戟[2]

([1] 国防科学技术大学计算机学院, 长沙 410073)

([2] 国防科学技术大学并行与分布处理国家重点实验室, 长沙 410073)

摘要: 为解决大规模本体理解问题, 提出了一个从复杂网络分析的角度研究大规模本体结构信息和重要概念挖掘的方法. 通过将基因本体的各种视图转换为网络进行全面分析, 证明了整个基因本体具有明显的复杂网络特征, 尤其是"小世界特性"和"无标度特性"; 但其子本体的复杂网络特性没有这么明显, 往往只具有"无标度特性"而没有"小世界特性". 同时, 利用网络分析中常用的节点重要性度量算法对本体中的重要概念进行挖掘. 在此基础上, 提出了基于 MEDLINE 信息检索结果的概念重要性评价算法, 评估几种节点重要性算法用于本体重要概念挖掘任务的正确性. 实验结果表明介数中心性算法在各种节点重要性度量算法中最适合于本体重要概念挖掘.

关键词: 基因本体; 复杂网络分析; 中心性度量

中图分类号: TP311