Brain Knowledge Graph Analysis Based on Complex Network Theory

Hongyin Zhu[1], Yi Zeng[1,2(✉)], Dongsheng Wang[1], and Bo Xu[1,2]

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
{zhuhongyin2014,yi.zeng}@ia.ac.cn
[2] Center for Excellence in Brain Science and Intelligence Technology,
Chinese Academy of Sciences, Shanghai, China

Abstract. Domain knowledge about the brain is embedded in the literature over the whole scientific history. Researchers find there are intricate relationships among different cognitive functions, brain regions, brain diseases, neurons, protein, gene, neurotransmitters, etc. In order to integrate, synthesize, and analyze what we have known about the brain, the brain knowledge graph is constructed and released as part of the Linked Brain Data (LBD) project, to reveal the existing and potential relationships of brain related entities. However, there are some incorrect and missing relationships in the extracted relations, and researchers also cannot find the key topics overwhelmed in the massive relations. Some researchers analyze the properties of vertices based on the network topology, but they cannot verify and infer the potential relations. In order to address the above problems, we propose a framework which consists of 3 parts. Firstly, based on complex network theory, we adopt the embeddedness to verify the relations and infer the potential links. Secondly, we use the network topology of existing knowledge to build the self-relations graph. Finally, the structural holes theory from sociology is adopted to discover the key and core vertices in the whole brain knowledge graph and we recommend those topics to users. Compared with logic inference methods, our methods are lightweight and capable of processing large-scale knowledge efficiently. We test the results about relation verification and inference, and the result demonstrates the feasibility of our method.

Keywords: Complex network · Brain knowledge graph · Relation inference · Network analysis · Linked Brain Data

1 Introduction

There is a long history of the research on the brain from the perspectives of its cognitive functions, its building blocks, and related brain diseases, etc. Brain research is not only useful because it is highly related to answer the question of who we are, the understanding of the brain is also important for the development of Artificial Intelligence. There is massive known and unknown knowledge about the brain, while knowledge engineering can help to extract, organize, and

G.A. Ascoli et al. (Eds.): BIH 2016, LNAI 9919, pp. 211–220, 2016.
DOI: 10.1007/978-3-319-47103-7_21

analyze these domain knowledge. Under this background, The Linked brain data (LBD) project is developed and the platform is released[1]. It's aim is to extract, synthesize, and analyze the data and knowledge about the brain from the World Wide Web [16]. However, it is inevitable that errors and missing relations exists in the LBD knowledge base. Besides, it is also hard to find the key topics which are overwhelmed in the enormous knowledge network.

Our work focuses on the network topology analysis to obtain new knowledge and new understandings based on the existing LBD brain knowledge graph. In [7], clustering coefficient is used to analyze the network topology of extracted information. In [8], graph theory based method is used to generate the document summarization. Their works mainly analyze the properties of vertices or relations according to their degrees. However, they cannot infer the potential links or verify the relations. Our contribution is the relation verification and inference based on the complex network theory.

In this paper, we propose a framework of analyzing brain knowledge graph by complex network theory. Firstly, the embeddedness is adopted to improve the accuracy of extracted relations and infer potential relations. Secondly, as an extension to the existing brain knowledge graph in Linked Brain Data, which focused on category inter-relationship, this paper extract category intra-relationship construction. Namely, the correlation of entities in the same categories (i.e. the category of cognitive functions, brain diseases, brain regions, neurons, proteins, genes, neurotransmitters). Finally, the structural holes theory [2] is adopted to find key topics for users.

2 Related Works

From the spatial perspective, domain knowledge on the brain is distributed around the world, such as different universities, laboratories and institutes, different literature sources, different databases. From the temporal perspective, they have been distributed almost in the whole history of Science. Although they are physically distributed, these knowledge on the brain are connected implicitly by nature, and they collectively provide a more comprehensive understanding of the brain. Nevertheless, the brain is still a mystery, and scientists are still on the way to provide a hologram of the brain. Most brain scientists focus on specific directions and scales for the investigation, and it is impractical for a brain scientist to know every scientific conclusion of existing brain research.

Under this background, the Linked Brain Data platform makes an effort to integrate and extract distributed knowledge on the brain and make a 10 million scale brain knowledge base accessible to all academic and industry communities. It integrates multi-source data and knowledge and links them semantically [16]. For the next stage, we not only plan to provide a brain knowledge graph that users could explore, but also want to provide domain knowledge based services (such as research recommendations).

[1] Linked Brain Data: http://www.linked-brain-data.org/.

For the relation verification, Liu et al. propose a method to verify "isa" relation based on specific features and rules [9], while the method is relation specific and cannot generalize to other relations. Zhang et al. propose an ontology based method to verify semantic relations, and their work needs a domain ontology and a vector space model [17]. Our paper proposes a model free method to verify the relations merely depending on the topology of the knowledge graph. As for the relation inference, Schoenmackers et al. propose a method to learn the inference rules from Web text [13]. Our method applies the existing topological structure to infer potential relations without rules. Currently, many efforts on recommender system focus on the adaptability to users [11]. Nevertheless, to the best of our knowledge, the work concerning recommending the key topics in the knowledge graph attracts little attention. Catanese et al. adopt the clustering coefficient to analyze the structural properties of Facebook Graph [4]. Here, we adopt clustering coefficient to find key topics in the brain knowledge graph.

3 Relation Verification and Inference

Since the domain knowledge is automatically extracted from scientific literatures, uncertainty are inevitable due to the reason that understanding of the brain may be inconsistent and the limitation of current automatic knowledge extraction techniques. The embeddedness [5] is the number of common neighbors of 2 vertices. The high embeddedness means high confidence, stability and consistency, and vice versa [1,6,12]. As for the knowledge graph, the relation confidence can be represented by embeddedness which also represents the strength or probability of a relationship.

The embeddedness of relations is calculated by Algorithm 1. Our first step is to find the corresponding entity pair according to the relation list in the knowledge graph. After getting the specific vectors, we can calculate their summation. If there is a common vertex, the corresponding element is 2 in the summation of the 2 vectors. For example, dementia is correlated with working memory. At the same time, the dementia is also correlated with white matter which is also correlated with working memory. So the white matter is the common vertex of the relation between dementia and working memory. It also means there is a triadic closure.

The higher the embeddedness value is, the stronger the binary relationship is. This method can support the correctness of the existing relations from a specific perspective. In addition, embeddedness can be used to infer currently unknown relations. More common vertices are available, more likely that a binary relation exists between the vertex pair. For example, based on the current brain knowledge graph in LBD, there is no direct relationship between the Zona incerta and the Lysine, but they have 36 common vertices, so the relationship between them may exist with a high probability. Hence, the method can support researchers to validate existing relations and predict unknown relationships.

We propose that we acquire new relations, the embeddedness calculation process is being carried out simultaneously as a supporting factor. We propose

Algorithm 1. Binary Relation Embeddedness Calculation Algorithm

```
Require: The adjacency matrix of vertices and the relation lists between those vertices
Ensure: The embeddedness of every vertices pair
  procedure CE($String[][]\ matrix$,$List\ relation$)
      for $i \leftarrow 0, relation.length - 1$ do
          $row[2]$=findRows($relation[i]$)
          for $j \leftarrow 1, relation.length - 1$ do
              $ele[j] = matrix[row[0]][j]+matrix[row[1]][j]$
          end for
          for $j \leftarrow 1, relation.length - 1$ do
              if $ele[j] > 1$ then
                  $multi$++
              else if $ele[j] == 1$ then
                  $single$++
              end if
              $emr = multi/(multi + single)$
          end for
      end for
  end procedure
```

this method as statistical topology inference (STI) which investigate on the probability of relations from a completely different perspective compared to logic inference. It transforms the topological properties of a graph into statistical features to infer the potential relations and support analysis on existing relations.

4 Category Intra-relationship Inference and Verification

For the previous version of the brain knowledge graph in Linked Brain Data, links are mainly established between entities in different categories, since for the first stage, we want to obtain relationships among different cognitive functions, brain diseases, and brain building blocks at multiple scales. However, links within the same category are also very important. For example, connections among different type of neurons are essential to understand the structural connectivity mechanism of the brain.

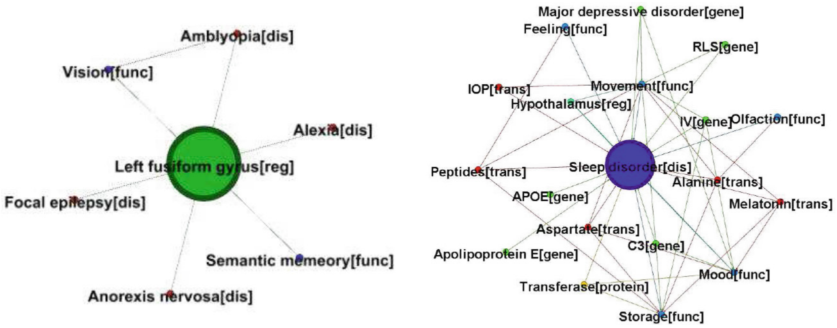
Category intra-relationship for cognitive functions (such as correlated relations of different cognitive functions) are also very important. Sometimes one kind of cognitive function does not play a separate role. Many cognitive functions serve as closely related building blocks to complex cognitive tasks. For example, Moscovitch et al. took the experiments to investigate on the relationship between long-term memory and episode memory in the same patient [10]. Now, by statistical topology inference, we may obtain possible relationships among different cognitive functions even before the experimental studies. Besides, possible correlated relations among brain diseases are also very important. This effort can be used to help doctors and medical researchers find potential relationships among different brain diseases to support their medical diagnosis and treatment.

5 Key Topics Discovery and Recommendation

Within the brain knowledge graph, some of the topics (domain terms) are essential from the topology point of view. The topology structure of a vertex can reflect its degree of importance compared to others. Here, we adopt the structural holes theory to find the key topics. Our method can find some topics located at a significant or special position of this knowledge graph.

In Sociology, there are some vertices with low embeddedness which are called structural holes [2,3]. The structural hole has many properties. For example, it is the connection vertex between several communities [2,3]. Based on its structural characteristic, the structural hole, the traffic hub of information, has higher power than other vertices. In the social network, the person, who is in the position of the structural hole, has a lot of interpersonal relations and becomes the key to communicating among several communities [2,3]. As for the knowledge graph, the structural holes are key concepts playing an important role in the connection of different local knowledge networks. In our experimental brain knowledge graph, there are only a few structural holes in the strict sense. In order to extend the result, we make some improvement to increase the number of candidates, and we can also get some vertices which are very similar with structural holes from topology perspective. When we increase the threshold of being a structural hole, the key vertices are more likely to show up.

Given the special position of structural holes in a network, their presence or disappearance will greatly affect the connectivity of the network. For example, it makes human more vulnerable to some extent that structural holes sometimes establish a shortcut for the diseases. It also reminds us of an effective way to eliminate the factors that can cause brain disease. Finding the structural holes would help people to prevent diseases with more explicit targets. If we removed the structural holes, a specific disease would only occur when several other conditions are satisfied together. Because we have already cut off this short path so that this disease only appears when it finds another complete pathway. It means that we can reduce the probability of a specific disease once we cut off the connectivity to structural holes. For example, the left fusiform gyrus correlated with various brain diseases and cognitive functions, as illustrated in Fig. 1(a). However, based on the partial knowledge graph, the semantic memory does not connect to brain diseases directly and it only connects to the left fusiform gyrus directly. If the semantic memory disorder symptom occurs in a patient, we may need to pay attention to the left fusiform gyrus, although it may be with no problem. The correlated vertices have the higher probability to be affected than those uncorrelated ones. If we take care of the left fusiform gyrus, we can predict or prevent the diseases, since based on the partial knowledge graph, semantic memory is not directly related to anorexia nervosa, amblyopia, etc. through the left fusiform gyrus, as shown in Fig. 1(a). It means paying special attention to these key vertices may can decrease the disease incidence, especially when the related vertices are starting lesion in patients.



(a) (b)

Fig. 1. (a) An example of structural hole, Left fusiform gyrus, and its related nodes. (b) An example of core vertex, the Sleep disorder, which has more triadic closures and is topologically very different from the structural holes.

The formation of structural holes is decided by the existing knowledge graph, so the relationships in this area may have not been revealed completely by scientists. We adopt the Algorithm 2 to find the structural holes.

There is another kind of important vertices, core vertices, which have more triadic closures around and present a totally different characteristic with the structural holes. Those core vertices can be found by the clustering coefficient [15], as shown in Eq. (1). c_i represents the clustering coefficient of vertex i. t_i is the number of edges among the neighbors of vertex i, and k_i is the number of its neighbors [14].

$$c_i = \frac{t_i}{C_{k_i}^2} \tag{1}$$

We adopt Algorithm 2 to calculate the clustering coefficient. Those vertices with the higher clustering coefficient are the core of stable communities which influence the whole network stability [15]. For example, the sleep disorder has many triadic closures around, as Fig. 1(b) shows. The sleep disorder has the capacity to form an intensive correlation with the surrounding vertices. This special structure characteristic represents special meaning to the whole structure. The key and core vertices can be recommended to the users.

6 Experiments

We take all the none duplicated correlated relations (265,946 relations) and related vertices (16,890) in Linked Brain Data to perform our experiments (The original data are brain related literature titles and abstracts from PubMed, ranging from the year 1874 to 2014). We take the above vertices and relations as seeds to generate 142,627,605 possible relations. In the relations, we found that 597,946

Algorithm 2. Calculating the clustering coefficient

```
Require: The adjacency matrix of vertices
Ensure: The structural holes and the clustering coefficient of every vertex
  procedure Clus(String[][] matrix)
      for $row \leftarrow 1, matrix.length - 1$ do
          List<Integer> $li$ = findOnes($matrix[row]$)
          for $i \leftarrow 0, li.length - 1$ do
              initialize indexList
              for $j \leftarrow i + 1, li.length - 1$ do
                  $x = li$.get($i$)
                  $y = li$.get($j$)
                  indexList.add(combinationIndex($x$,$y$))
              end for
              clusteringCofficient(indexList)
          end for
      end for
  end procedure
```

relations have more than 20 common vertices in their neighbors between different categories and 602,389 relations in the same category, some examples are shown in Table 1. S represents the number of the neighbors which only have one relationship with Entity 1 or Entity 2. The EM represents the number of common vertices of a specific entity pair. The EMR is the embedding ratio. The gene, reg, dis, protein, trans, func and neu represent the gene, brain regions, brain diseases, protein, neurotransmitters, cognitive functions and neurons respectively. When we sort relations by EMR (with a threshold $EMR > 0.5$), there are only 8,250 relations between different categories and 204,345 relations without category limitation. The huge difference indicates that there are extensive relations in the same category and the portion of common vertices of many entity pairs is small.

In the existing relations, 155,729 relations have more than 20 common vertices. The cardinal number of common vertices can be very big, but the embedding ratio of most relations is less than 40 %. It implies that most of the neighbors are correlated with only one entity of the two entities in a specific relation pair. According to the various situation mentioned above, we design some rules to find the relations with both high cardinal number and embedding ratio. These relations are considered as the highly confident ones.

We randomly select 1000 verified relations about brain regions, brain diseases and cognitive functions from the extracted relations, and we manually check the correctness of them. Our experimental results show the verification precision is 95.3 % when we set the EM > 20. This method can filter some of the incorrect relations. As for the inferred relations, they are to some extent generated hypothesis, and we expect and invite Brain Scientists to investigate on these hypothesis

Table 1. Example relations and their corresponding parameters

Entity1	Entity2	S	EM	EMR
Schizophrenia [dis]	Encoding [func]	6672	1140	0.14592
Atherosclerosis [dis]	Encoding [func]	6534	865	0.11690
CA2 [reg]	Encoding [func]	6290	1226	0.16311
Hippocampus [reg]	Movement [func]	1798	616	0.25517

and verify them by biological experiments[2]. The above inference function can be considered as a novel way to find the potential links.

Table 2. Some examples of the inferred brain region correlations which are not extracted directly

Entity1	Entity2	S	EM	EMR
CA2 [reg]	Hippocampus [reg]	1391	737	0.34633
Hypothalamus [reg]	CA2 [reg]	1278	558	0.30392
Cerebellum [reg]	Hippocampus [reg]	1091	763	0.41154
CA2 [reg]	CA1 [reg]	1179	553	0.3192
Hypothalamus [reg]	Forebrain [reg]	810	469	0.36669

In the category intra-relation inference experiment, some examples of the inferred relations about brain regions are shown in Table 2. We randomly select 100 inferred relations between brain regions and manually check the correctness of them. The precision is currently 85 % when we set EM > 20.

Table 3. Examples of the key vertices in the brain knowledge graph

Structural holes	Num
NO [protein]	47
GABA [protein]	43
Knowledge retrieval [func]	7
Core of nucleus accumbens [reg]	7
Barbiturate dependence [dis]	5

Table 4. Some examples of the core vertices in the knowledge graph

VERTICES	R	V	CC
Encoding [func]	195765	7277	0.0074
Movement [func]	96921	1630	0.0730
Alzheimer [dis]	91519	1582	0.0732
Schizophrenia [dis]	68227	1675	0.0487

As for the topics discovery experiments, some examples of the key vertices are shown in Table 3 where *Num* denotes the number of neighbors of a specific vertex. Most of the key vertices have high value of *Num* and many relations with

[2] Inferred relationships can be accessed through Linked Brain Data.

their neighbor vertices. Some examples of the core vertices in the brain knowledge graph is shown in Table 4. R represents the number of relationships with the neighbors of a corresponding vertex. V represents the number of neighbor vertices. CC is the value of the clustering coefficient. Some vertices with low CC value but high R value also can be considered as the core vertices since they also have many triadic closures. Finally, users can get some structurally important vertices and relations overwhelmed in the massive knowledge on the Brain.

7 Conclusion and Future Work

Based on complex network theories, we propose a framework to address the problems of relation verification, inference and key topics discovery on brain knowledge graph. Firstly, the verification and inference of relation extraction are investigated based on the embeddedness. We test our verified results based on the annotated data. The experimental results demonstrate the feasibility of our method. Secondly, we investigate on the category intra-relations and use embeddedness for verification. Finally, the discovery function of key and core topics is realized by the structural holes algorithm which is borrowed from sociology.

Our future work will consider extracting the specific types of the correlated relations in the brain knowledge graph. We will also invite brain scientists to verify the potential links that we generated based on the prediction model introduced in this paper.

Acknowledgments. This study was funded by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB02060007), and Beijing Municipal Commission of Science and Technology (Z151100000915070, Z161100000216124).

References

1. Bearman, P.S., Moody, J.: Suicide and friendships among American adolescents. Am. J. Pub. Health **94**(1), 89–95 (2004)
2. Burt, R.S.: Structural holes and good ideas. Am. J. Sociol. **110**(2), 349–399 (2004)
3. Burt, R.S.: Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge (2009)
4. Catanese, S., Meo, P.D., Ferrara, E., Fiumara, G., Provetti, A.: Extraction and analysis of facebook friendship relations. In: Abraham, A. (ed.) Computational Social Networks, pp. 291–324. Springer, Berlin (2012)
5. Granovetter, M.: Economic action and social structure: the problem of embeddedness. Am. J. Sociol. **91**, 481–510 (1985)
6. Kossinets, G., Watts, D.J.: Empirical analysis of an evolving social network. Science **311**(5757), 88–90 (2006)
7. Li, X.: Graph-based learning for information systems. Ph.D. thesis, The University of Arizona (2009)
8. Li, Y., Cheng, K.: Single document summarization based on clustering coefficient and transitivity analysis. In: Proceedings of the 10th International Conference on Accomplishments in Electrical and Mechanical Engineering and Information Technology, pp. 26–28 (2011)

9. Liu, L., Zhang, S., Diao, L., Yan, S., Cao, C.: Automatic verification of "ISA" relations based on features. In: Proceedings of the Sixth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 2, pp. 70–74. IEEE Press (2009)
10. Moscovitch, M., Nadel, L., Winocur, G., Gilboa, A., Rosenbaum, R.S.: The cognitive neuroscience of remote episodic, semantic and spatial memory. Curr. Opin. Neurobiol. **16**(2), 179–190 (2006)
11. Nanda, A., Omanwar, R., Deshpande, B.: Implicitly learning a user interest profile for personalization of web search using collaborative filtering. In: Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 2, pp. 54–62. IEEE (2014)
12. Rapoport, A.: Spread of information through a population with socio-structural bias: Iii. Suggested experimental procedures. Bull. Math. Biophys. **16**(1), 75–81 (1954)
13. Schoenmackers, S., Etzioni, O., Weld, D.S., Davis, J.: Learning first-order horn clauses from web text. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 1088–1098. Association for Computational Linguistics (2010)
14. Soffer, S.N., Vazquez, A.: Network clustering coefficient without degree-correlation biases. Phys. Rev. E **71**(5), 057101 (2005)
15. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**(6684), 440–442 (1998)
16. Zeng, Y., Wang, D., Zhang, T., Xu, B.: Linked neuron data (lnd): a platform for integrating and semantically linking neuroscience data and knowledge. In: Frontiers in Neuroinformatics. Conference Abstract: The 7th Neuroinformatics Congress (Neuroinformatics 2014), Leiden, The Netherlands, pp. 1–2 (2014)
17. Zhang, X., Chen, H., Ma, J., Tao, J.: Ontology based semantic relation verification for TCM semantic grid. In: Proceedings of 2009 Fourth ChinaGrid Annual Conference, pp. 185–191. IEEE Press (2009)