Identifying important nodes from content-associated heterogeneous graph by LeaderRank

Yulong Dai¹, Qiyou Shen², Xiangqian Xu³ and Jun Yang⁴

¹ College of Military Basic Education, National University of Defense Technology, Changsha, Hunan, P. R. China
² College of Information Communication, National University of Defense Technology, Wuhan, Hubei, P. R. China
³, ⁴ College of Systems Engineering, National University of Defense Technology, Changsha, Hunan, P. R. China

*2018111436@stu.gzhu.edu.cn

Abstract. Most real-world systems consist of a large number of interacting entities of many types. However, most of the current researches on systems are based on the assumption that the type of node or link in the network is unique. In other words, the network is homogeneous, containing the same type of nodes and links. Based on this assumption, differential information between nodes and edges is ignored. This paper firstly introduces the research background, challenges and significance of this research. Secondly, the basic concepts of the model are introduced. Thirdly, a novel type-sensitive LeaderRank algorithm is proposed and combined with distance rule to solve the importance ranking problem of content-associated heterogeneous graph nodes. Finally, the writer influence data set is used for experimental analysis to further prove the validity of the model.

1. Introduction

Information network[1] is a structured text way of knowledge representation. It contains a series of nodes and the links between nodes. Its structure reflects the structure of information stored in nodes, so it is called information network. Classic examples of information networks include academic networks[2], social media networks[3], social networks, etc. The traditional network model uses the homogeneous network to model the real system. This model ignores the difference information between entities, which leads to incomplete or missing information. Content-associated heterogeneous graphs[4] can accurately distinguish different meanings in information network and dig out more meaningful knowledge by analyzing various types of nodes in the network and various links between different types of nodes. Thus, it is of great significance to study content-associated heterogeneous graphs.

Discovering the core nodes in content-associated heterogeneous graph plays an important role in understanding the operating mode of the system and grasping the key factors of the system. It has important research value in epidemic prevention and control, network security and many other fields. At present, the research methods of node importance of homogeneous network have been relatively mature, such as PageRank algorithm[5] based on feature vector, HITS[6] algorithm, K-Shell method[7] based on node neighbor, degree centrality[8], etc. Compared with homogeneous networks, content-associated heterogeneous graphs contain richer topological structure and semantic information, which also poses new problems and challenges for the identification of important nodes. Therefore, the purpose
of this paper is to propose a node importance ranking method suitable for content-associated heterogeneous graphs.

In 2015, Shi et al.[9] summarized the heterogeneous information network analysis, introduced the basic concepts of heterogeneous information network analysis, discussed the development of heterogeneous information network analysis in different data mining tasks, and pointed out the future research direction. In 2019, Mohammad MehdiKeikha et al.[10] used deep learning to learn feature vectors of nodes in the network. In 2020, Soheila Molae et al.[11] proposed a method to measure the importance of nodes by integrating local and global information of network.

By adding a ground node to replace the page jump probability setting in PageRank, Lü et al.[12] proposed the LeaderRank algorithm, which effectively solves the problem of inconsistent ranking of different subnets in PageRank algorithm. However, the original LeaderRank algorithm is based on the homogeneous network and ignores the difference information between nodes. To address this deficiency, the authors improve the original LeaderRank algorithm and propose a type-sensitive LeaderRank algorithm for the richness and semantic complexity of content-associated heterogeneous graph nodes. Firstly, the importance of nodes is divided into the importance of each node type, and then the importance of each type is combined into a global importance by using the distance rule.

The overall structure of this paper is as follows: In section 2, we introduce the basic concepts of content-associated heterogeneous graphs, LeaderRank algorithm and distance rule. In Section 3, we build the model and provide a detailed description of the type-sensitive algorithm. In section 4, we give the process of mining the key nodes of writer influence network to verify the validity of the model. Finally, summarizing discussions and conclusions of the paper will be in the last Section 5.

2. Preliminaries

In this section, we introduce the concepts of content-associated heterogeneous graphs, Pagerank algorithm, and distance rule that will be used in this paper.

2.1. Content-associated heterogeneous graphs

Basic definitions and concepts of content-associated heterogeneous graphs have been illustrated by Zhang et al.[13] in 2019.

**Definition 1.** A content-associated heterogeneous graph is defined as a graph \( G = (V, E, V_0, E_0) \), The specific definition is as follows:

- \( V \): A set of nodes,
- \( E \): A set of links between nodes,
- \( V_0 \): A set of object types of nodes,
- \( E_0 \): A set of relation types of links

Each node or link belongs to a specific type in \( V_0 \) or \( E_0 \).

Figure 1 shows an academic graph, which is a typical content-associated heterogeneous graph. The node types \( V_0 \) include paper, place, and author, and the link types include write, cite, and publish. Since a paper can be written by a group of authors and published on a venue, the relationship between the paper and the author is one-to-many, and the relationship between the paper and the site is one-to-one.
2.2. LeaderRank algorithm
LeaderRank algorithm is a node ranking algorithm proposed by improving PageRank algorithm, which is mainly used to rank nodes of online social networks. LeaderRank algorithm proposes a solution to the problem of inconsistent ranking of disconnected subnet pages in PageRank algorithm. A ground node \( g \) is added into the whole network, connected with each node in the network bidirectionaly. The original network becomes a network of \( N+1 \) nodes and \( M+2N \) links. The addition of ground node makes the original network become a strongly connected network, which solves the problems of node ranking and probabilistic jump in PageRank and speeds up the convergence of algorithm.

The calculation process of the LeaderRank algorithm is an iterative process, which is as follows:

- Allocate one unit of resource to each node except the ground node.
- Each node allocates resources equally to the nodes it points to according to the output.
- Iterates until a steady state is attained.

2.3. Distance rule
Distance rule\[14\] is a multidimensional utility mergence method. When evaluating an object, it is necessary to combine multiple utilities. Take two-dimensional utility as an example, when the two utilities reach the maximum value at the same time, the combined utility reaches the maximum value. When both utilities reach the minimum value at the same time, the combined utility is zero. The distance rule is defined as follows:

**Definition 2.** Assume that the utility of \( n \) evaluation criteria are \( u_1, u_2, ..., u_n \), and the value range is \([0,1]\). The calculation formula of the distance rule is as follows:

\[
W(u_1, u_2, ..., u_n) = 1 - \left( \frac{1}{n} \sum_{i=1}^{n} (1 - u_i)^2 \right)^{\frac{1}{2}}
\]  (1)

3. Model and method
In this section, we improve the LeaderRank algorithm based on the characteristics of content-associated heterogeneous graphs, proposed a type sensitive LeaderRank algorithm, and combine it with distance rule to establish a model to identify important nodes of content-associated heterogeneous graphs.

**Definition 3.** Let \( G = (V, E, O_v, R_e) \) be a content-associated heterogeneous graph with a ground node. The adjacent matrix \( A \) of \( G \) can be expressed as:

\[
a_{ij} = \begin{cases} 
1, & \text{if } (v_i, v_j) \in E \\
0, & \text{if } (v_i, v_j) \notin E
\end{cases}
\]  (2)

Then the state transition matrix \( P \) can be expressed as:

\[
p_{ij} = a_{ij} \left( \sum_{v \in V} a_{ij} \right)^{-1}
\]  (3)
For general homogeneous graphs, the importance of each node can be directly calculated by constructing the adjacency matrix and state transition matrix, but this is not applicable to content-associated heterogeneous graphs. To address this problem, a set of type vector t is introduced to distinguish the type differences between nodes.

Definition 4. Let \( G = (V, E, O_v, R_v) \) be a content-associated heterogeneous graph with a ground node. For node type \( T_i \), type state vector \( t_{ij} \) can be expressed as:

\[
t_{ij} = \begin{cases} 
1, & \text{if } O_v = T_i \\
0, & \text{if } O_v \neq T_i
\end{cases}
\]  

(4)

As mentioned above, after continuous iteration, the importance of each type of node will converge to a fixed value and then be integrated. For \( T_i \) nodes, the calculation process is as follows:

**Algorithm 1. Type-sensitive LeaderRank for node type \( T_i \)**

Input: The set of nodes \( V \), the set of neighbor nodes \( P_i \) of each node, threshold \( \varepsilon \)
Output: influence value \( R_i \)
1: for each node i, initialize \( Q_i \leftarrow 1/|P_i| \)
2: for each node i in V
3: \( T_i \leftarrow 0 \)
4: for each node j in \( P_i \)
5: \( T_i \leftarrow T_i + Q_j \)
6: while True do
7: for each node i in V
8: \( R_i \leftarrow 0 \)
9: for each node j in \( P_i \)
10: \( R_i \leftarrow R_i + T_j \times Q_j \)
11: if the total amount change < \( \varepsilon \), then
12: end

Through the above calculation process. We calculated the matrix \( R \) containing the importance of each node, where \( r_{ij} \) represents the importance of node \( v_j \) in node type \( T_i \). We then use min-max scaling to normalize the importance of each node in type \( T_i \) to the interval \([0,1]\):

\[
u_{ij} = \frac{r_{ij} - \min(r_i)}{\max(r_i) - \min(r_i)}
\]  

(5)

Where \( \max(r_i) \) and \( \min(r_i) \) are maximum and minimum importance of type \( T_i \) nodes.

The normalized importance of each node in each node type is calculated and the importance matrix \( U \) is obtained, where \( u_{ij} \) represents the importance of node \( v_j \) in node type \( T_i \).

Then, the distance rule is used to calculate the global importance of node \( v_j \) as follows:

\[
w_j = 1 - \left( \frac{1}{n} \sum_{i=1}^{n} \left( 1 - u_{ij} \right)^2 \right)^{\frac{1}{2}}
\]  

(6)

Finally, the importance order of the content-associated heterogeneous nodes can be obtained by ranking the nodes by \( w_j \).

4. Experimental analysis

In this section, in order to further prove the validity of the importance ranking model for content-associated heterogeneous graphs, an experimental analysis is conducted with the writer influence dataset.
4.1. Data declaration
The data set comes from Kaggle.com, which records the influence relationships of more than 4,000 writers since 1800. A writer may be influenced by more than one other writers, and writers can be divided into five genres. Therefore, with writer as node, writer genre as node type, and influence relationship between writers as edge, the content-associated heterogeneous graphs of writer's influence are constructed. Some statistics of writer content-associated heterogeneous graph are shown in Table 1:

| Type          | Quantity |
|---------------|----------|
| Novelist      | 2531     |
| Poet          | 904      |
| Short Story Writer | 1038     |
| Playwright    | 371      |
| Essayist      | 294      |

4.2. Experiment process
According to the formula (3), we can calculate the state transition matrix $P$. $P$ is a sparse matrix, from which it can be seen that there is an obvious community division phenomenon in this graph.

Let $\epsilon_1 = \epsilon_2 = \cdots = \epsilon_5 = 1 \times 10^{-3}$ and $q_1 = q_2 = \cdots = q_5 = 0.2$. Then, importance matrix $U$ can be calculated. According to the importance matrix $U$, the top three writers in novelist, poet and playwright are listed in Table 2:

| Genre        | Writer          | Importance |
|--------------|-----------------|------------|
| Novelist     | J. K. Rowling   | 1.00       |
|              | Ernest Hemingway| 0.85       |
|              | Charles Dickens | 0.81       |
|              | J. R. R. Tolkien| 1.00       |
| Poet         | Edgar Allan Poe | 0.87       |
|              | Rabindranath Tagore | 0.79     |
| Playwright   | William Shakespeare | 1.00    |
|              | Oscar Wilde     | 0.86       |
|              | Arthur Miller   | 0.83       |

Next, we use the formula to calculate the combined importance of each writer node. For example, in the importance matrix $U$, Oscar Wilde's column values are $[0.87,0.96,0.81,0.89,0.82]^T$. Thus, the global importance for Oscar Wilde is 0.959. Some statistics of global importance calculation results are shown in Table 3:
Table 3. Experimental result

| Importance interval | Nodes                                      | Description                                      |
|---------------------|--------------------------------------------|---------------------------------------------------|
| 1                   | J. K. Rowling (Novelist), J. R. R. Tolkien (Poet), William Shakespeare (Playwright)… (5 in total) | The most influential writer in a particular field |
| [0.6,1)             | Ernest Hemingway (Novelist), Edgar Allan Poe (Poet), Oscar Wilde (Playwright)… (517 in total) | Writers with great influence in a particular field |
| [0.3,0.6)           | Charles Rhoton (Novelist), Denny Langon (Essayist)… (1928 in total) | Writers with moderate influence in a particular field |
| [0,0.3)             | Coy Pack (Short Story Writer), Nathanial Mankoski (Essayist),… (2688 in total) | Writers of little influence in any field |

In order to more clearly show the effect of the model on node importance classification, we draw the distribution histogram of node importance values, as shown in Figure 2 below. As can be seen from the figure, the importance of nodes is mainly concentrated in two intervals [0.1,0.25] and [0.6,0.7]. Very few nodes have very high ([0.95,1]) or very low ([0,0.05]) importance. Therefore, this experiment successfully distinguishes important nodes from non-important nodes.

Figure 2. Distribution histogram of node importance values

5. Conclusions

In this paper, we introduced the ranking of node importance in content-associated heterogeneous graphs. At present, most research objects are homogenous networks, and the differential information in the network is ignored, resulting in incomplete or missing information. In addition, the traditional Pagerank algorithm has the problem of inconsistent node importance ranking for disconnected subnets.

In order to address these problems, the authors propose a novel type-sensitive LeaderRank algorithm, and combine it with the distance rule to build the model, which makes full use of the difference information between nodes and improves the calculation efficiency. It provides an efficient, accurate and concise solution to the problem. The experimental results show that this model has a good effect in distinguishing important nodes and non-important nodes in Content-associated heterogeneous graphs.

However, this model does not make full use of the differential information of links, nor does it work as well in heterogeneous networks in other fields as it does in social networks. Thus, in future work, the full use of edge difference information will become the vital aspect of improvement, followed by the study of utility combination rules.
Acknowledgments
We are thankful to the Editor and the reviewers for their valuable comments and detailed suggestions to improve the presentation of the paper. Further, we also acknowledge the support in part by the National Natural Science Foundation of China under grant No.71690233 and No.71671186.

References
[1] HAO PENG, JIANXIN LI, YANGQIU SONG, et al. Streaming Social Event Detection and Evolution Discovery in Heterogeneous Information Networks. ACM Transactions on Knowledge Discovery from Data. 2021;15(5):1-33. doi:10.1145/3447585
[2] Bedru HD, Zhao W, Alrashoud M, Tolba A, Guo H, Xia F. TOSNet: A Topic-Based Optimal Subnetwork Identification in Academic Networks. IEEE Access, Access, IEEE. 2020;8:201015-201027. doi:10.1109/ACCESS.2020.3034997
[3] Jahromi, D. A. (2020). A Quantitative Study of the Perceived Impact of Social Media Networks on Bahraini Users’ English Language Learning. Teaching English with Technology, 20(4), 23–40.
[4] Li Y, Chen C, Duan M, Zeng Z, Li K. Attention-Aware Encoder–Decoder Neural Networks for Heterogeneous Graphs of Things. IEEE Transactions on Industrial Informatics, Industrial Informatics, IEEE Transactions on, IEEE Trans Ind Inf. 2021;17(4):2890-2898. doi:10.1109/TII.2020.3025592
[5] Jiung Lee, Mariana Olvera-Cravioto, PageRank on inhomogeneous random digraphs, Stochastic Processes and their Applications, Volume 130, Issue 4, 2020, Pages 2312-2348, ISSN 0304-4149.
[6] Kleinberg JM. Authoritative Sources in a Hyperlinked Environment. Journal of the ACM. 1999;46(5):604. doi:10.1145/324133.324140
[7] Kitsak M, Gallos L K, Havlin S, et al. Identification of Influential Spreaders in Complex Networks[J]. Nature Physics, 2012, 6(11): 888-893.
[8] Han L, Zhou Q, Tang J, Yang X, Huang H. Identifying Top-k Influential Nodes Based on Discrete Particle Swarm Optimization With Local Neighborhood Degree Centrality. IEEE Access, Access, IEEE. 2021;9:21345-21356. doi:10.1109/ACCESS.2021.3056087
[9] C. Shi, Y. Li, J. Zhang, Y. Sun, and P. S. Yu, “A survey of heterogeneous information network analysis,” arXiv:1511.04854, 2015.
[10] Mohammad Mehdi Keikha, Maseud Rahgozar, Masoud Asadpour, Mohammad Faghih Abdollahi, Influence maximization across heterogeneous interconnected networks based on deep learning, Expert Systems with Applications, Volume 140, 2020, 112905, ISSN 0957-4174.
[11] Soheila Molaei, Reza Farahbakhsh, Mostafa Salehi. Noel Crespi, Identifying influential nodes in heterogeneous networks[J]. Expert Systems with Applications, Volume 160, 2020, 113580, ISSN 0957-4174.
[12] Lü L, Zhang Y-C, Yeung CH, Zhou T. Leaders in Social Networks, the Delicious Case. 2011. doi:10.1371/journal.pone.0021202.
[13] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, Nitesh V. Chawla. Heterogeneous Graph Neural Network. KDD 2019.
[14] Mohammad Mehdi Keikha, Maseud Rahgozar, Masoud Asadpour, Mohammad Faghih Abdollahi, Influence maximization across heterogeneous interconnected networks based on deep learning, Expert Systems with Applications, Volume 140, 2020, 112905, ISSN 0957-4174.