Feature Detection in Hierarchical Structure Based on Tensor Factorization

Zhe Li¹*, Dong Wang² and Ke Zhang³

¹School of Control Science and Engineering, Shandong University, Jinan, 250061, P. R. China
²Shandong Tengdong Shengjian Coal Mine, Tengzhou, 277599, P. R. China
³Beijing Convenient Bee Chain Business Co., Ltd, Beijing, 100000, P. R. China

*Corresponding author email: lz910213@163.com

Abstract. Feature is an important connection point of hierarchical structure. Large-scale network feature detection becomes an important way to analyse hierarchical structure. In this paper, we use tensor decomposition based large-scale network feature detection to study hidden relationships among different users and infer its potential feature structure. Data set experiment is used to verify the effectiveness of the algorithm from two aspects of accuracy and efficiency. At the same time, the two algorithms of tensor decomposition and Fast Newman and SLPA are compared to find that the hierarchical structure feature detection algorithm based on tensor decomposition is more efficient and more accurate.

Keywords: Hierarchical structure; Tensor factorization; Feature detection; Data set experiment.

1. Introduction
Hierarchical structure has become an important research area which has attracted tremendous researches [1][2]. There has been much work done both in industry and academia on developing new approaches to analyse the networks over last decade. In the paper, we present our work towards addressing feature detection efficiently in a large-scale network, which is based on tensor factorization [3][4][5]. To this end, we extend the concept of matrix factorization to the concept of tensor factorization. A tensor is a generalization of matrix concept to multiple dimensions. In the example of DBLP data (see section 4.1), the usual author-author two-dimensional matrix is converted into a multiple dimensional tensor. We first present a brief survey of feature detection in hierarchical structure Then, we propose our method to detect communities based on Tensor factorization. A tensor is a generalization of matrix concept to multiple dimensions. The examples apply the method to a real network with a known characteristic structure. We further discuss the proposed method and draw conclusions.

2. Feature Detection
Many hierarchical structure are found to divide naturally into modules, namely communities, groups of nodes with relatively dense connections within groups but sparser connections between them[6][7]. Early efforts at feature detection, going back to the 1970s, which adopts graph partitioning technology such as spectral partitioning [8] and the Kernighan-Lin algorithm [9]. Another line of feature detection research is to build a hierarchical structure of communities, which is early and still widely used. Girvan and Newman first proposed a divisive algorithm based on edge between [6]. In recent years,
some optimization methods are proposed, such as the so-called E/I ratio, likelihood-based methods and others. However, the most widely used is the measure known as the modularity [10]. In the paper we propose and demonstrate an algorithm for feature detection that can detect hidden relations and take multi-dimensional attributes into account, which satisfies all of the demands above. On DBLP date the algorithm has performed better than the best previous algorithms in detecting known feature structure.

3. Tensor Factorization

For the sake of simplicity, we will describe the model for a single contextual variable $C$, and therefore $Y$ the tensor containing contextual information will be a 3-dimensional tensor. The generalization to larger numbers of context variables and $N$ dimensions is trivial. In the following we denote the sparse tensor of contextual observations by $Y \in \mathbb{R}^{n \times m \times c}$, where $n$ are the numbers of authors, $m$ the number of papers, and $c$ where $c_i \in \{1, \ldots, c\}$ the number of contextual variables. Specially, the contextual information is given on a ten star scale and thus $Y \in \{0, \ldots, 9\}^{n \times m \times c}$, where the variables indicate that the research field of the papers, according to the classification criteria provided by the Chinese Computer Association.

The two most commonly used tensor factorizations are the Tucker model [3] and CP model [4], both of which can be regarded as higher-order generalizations of the matrix singular value decomposition. In our approach, we follow the Tucker factorization, own to its more flexibility, where the 3-dimensional tensor is factorized into a core tensor transformed by factor matrices, each of which is for one mode of the tensor.

The Tucker factorization of a 3-dimensional tensor $Y$, which can be formulated as

$$Y = G \times_1 A \times_2 B \times_3 C$$

(1)

where $A \in \mathbb{R}^{n \times r_1}, B \in \mathbb{R}^{m \times r_2}, C \in \mathbb{R}^{c \times r_3}$ are the latent factor matrices respectively for the 1-mode, 2-mode, and 3-mode unfolding of $Y$. $G \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the core tensor showing the interaction between the latent components from the three factor matrices. For example, the element $g_{pqr} \in G$ shows the interaction between latent component $a_{p}b_{q}c_{r}$. Correspondingly, each component of can be computed via

$$x_{ijk} = \sum_{p=1}^{r_1} \sum_{q=1}^{r_2} \sum_{r=1}^{r_3} g_{pqr} a_{ip} b_{jq} c_{kr}$$

(2)

Where $i = 1, \ldots, n$, $j = 1, \ldots, m$. Tucker factorization is most suitable for modelling relationships and previous work has shown that the power of factorization approaches can predict potential relations by employing tensor factorization. The task investigated in this paper is to extend factorization approaches to include contextual information which can detect feature structures in hierarchical structure.

4. Experiments

First, we download DBLP.xml data from the official websites, analyse the file format, parse data sets, and then store the data into the database. Second, according to the classification of Chinese computer association, we divide the conference journals into three levels (A,B,C). The data of the top 500, 400, 300, 200 and 100 of the maximum number of published papers are taken out respectively as our experimental data. Lastly, we construct a relation matrix according to the three levels, and the value of the matrix is represented by the number of papers published by author $I$ and author $J$, shown as Figure 1.
The second dataset is classic datasets of movieLens. According to the time users wrote comments, we divided the users into three levels (A, B, C). To ensure the randomness of the experiment, we randomly selected 100, 200, 300, 400 and 500 people for experiments, and then expressed the strength of the relationship by counting the number of viewing the same movie between the users in different time periods. This constructs a user-user-time tensor and the data therein represents the number of the same movie viewed by user i and user j in the first period of time, shown as Figure 2. Finally, the accuracy of the feature classification is determined by the user's occupation.

This article selects the two contrastive algorithms, one kind is Fast Newman, and another is SLPA. Both are classic algorithm of feature discovery. We do experiments on two dataset, the results shown in the table.
Table 1. MovieLens accuracy

|       | 100   | 200   | 300   | 400   | 500   |
|-------|-------|-------|-------|-------|-------|
| Zhang liang | 0.312 | 0.3056| 0.3255| 0.3376| 0.3436|
| SLPA   | 0.2254| 0.2534| 0.2657| 0.2486| 0.2897|
| Fast Newman | 0.281 | 0.284 | 0.271 | 0.297 | 0.321 |

Table 2. DBLP accuracy

|       | 100   | 200   | 300   | 400   | 500   |
|-------|-------|-------|-------|-------|-------|
| Zhang liang | 0.57  | 0.582 | 0.593 | 0.5975| 0.638 |
| SLPA   | 0.513 | 0.486 | 0.489 | 0.521 | 0.519 |
| Fast Newman | 0.56  | 0.525 | 0.573 | 0.53  | 0.552 |

From the above table it can be seen that the three algorithms of accuracy is not high, because of MovieLens dataset. In combination with table 1 and table 2, tensor factorization algorithm is significantly higher than the other two algorithms in accuracy. Based on above analysis, this shows tensor factorization algorithm is feasible in feature detection.

Figure 4 and Figure 5 give the accuracy and runtime curve respectively compared with SLPA, Fast Newman, tensor factorization Algorithm.

Figure 4. MovieLens line chart of accuracy

From the above Figure 4 it can be seen that although the accuracy of the three algorithms is not very high, the tensor decomposition algorithm has obvious advantages over the other two algorithms. The accuracy of the tensor decomposition algorithm has a clear upward trend with the number of people increasing, while the accuracy of the SLPA algorithm in 400 people and Fast Neman algorithm in 300 people has a significant decline. The overall accuracy of the two algorithms has large ups and downs, while the tensor decomposition algorithm inherits a steady growth trend.

From the DBLP dataset experiments, it can be seen that the tensor decomposition algorithm is still superior to the other two. In the Fast Neman algorithm and the SLPA algorithm, however, there is no indication that the accuracy rate increases with the number of people. In particular, Fast Neman algorithm accuracy has a large fluctuation with the increase of the number of people. Because of the Fast Neman running time is too long, so we just compare SLPA and Fast Newman algorithm. And tensor factorization algorithm not only run time is shorter than SLPA, and growth is slower than the SLPA.

Figure 5. DBLP and runtime line chart of accuracy
5. Conclusions
In this paper, we present a novel feature detection algorithm in hierarchical structure based on tensor factorization. Using tensor factorization, the feature structure of the original network becomes more obvious since the hidden relations separated by different users are detected. The experimental results show that the proposed algorithm is more efficient on real data on social, compared with other algorithms. Furthermore, the proposed algorithm has a good scalability in more dimensional models. Further efforts are needed to explore the algorithm more. The first research goal is how to select the threshold value from the relationship matrix to better predict the feature, and get a higher accuracy rate. Second, although the tensor factorization algorithm has a lot of advantages over Fast Newman algorithm in the running time, but the running time is still a lot of restrictions. The next step is to improve the efficiency of the algorithm.

Acknowledgement
This work was supported by the General Program of National Natural Science Foundation of China under Grant U1613223, National Key Research and Development Project under Grant 2017YFC0806505 and Grant 2015AA042201, the Key Research and Development Program of Shandong Province under grant 2019GGX104091.

References
[1] H.Kwak, C.Lee, H.Park, and S.Moon. “What is twitter, a hierarchical structure or a news media?” in Proceedings of the 19th international conference on World Wide Web.ACM,591-600, 2010.
[2] DJW. Guorgi Kossinets. “Empirical analysis of an evolving hierarchical structure,” Science, vol. 311, no.5757, pp. 88-90, 2006.
[3] LR. Tucker. “Some mathematical notes on three-mode factor analysis,” Psychometrika, vol. 31, no.3, pp. 279-311,1996.
[4] RA. Harshman. “Foundation of the PARAFAC procedure: Models and conditions for an “explanatory” multimodal factor analysis,” UCLA working papers in phonetics, vol.16, pp. 1-84, 1970.
[5] JD. Carroll, J.Chang. “Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young”decomposition,” Psychometrika, vol. 35, no.3, pp. 283-319, 1970.
[6] M. Girvan, ME.Newman. “feature structure in social and biological networks,” Proceedings of the National Academy of Sciences of the United States of America, vol. 99, no.12, pp. 7821-7826, 2002.
[7] S.Fortunato. “feature detection in graphs,” Physics Reports, vol. 486, no.3, pp. 75-174, 2002.
[8] A.Pothen, H.Simon, KP. Liou. “Partitioning sparse matrices with eigenvectors of graphs,” SLAM Journal on Metrix Analysis and Applications, vol. 11, no.3, pp. 430-452, 1990.
[9] BW. Kernighanm, S.Lin. “An efficient heuristic procedure for partitioning graphs,” Bell system Technical Journal, vol. 49, no.2, pp. 291-307, 1970.
[10] MEJ. Newman. “Fast algorithm for detecting feature structure in networks,” Physical Review, vol.69, No.6Part2, pp. 69066133, 2004.