Research on Community Detection and Community Correlation of Academic Large Data

Zhang Xilun
Shanghai Jiao Tong University
zhangxilun@sjtu.edu.cn

ABSTRACT

In this semester, "Wireless Communication Principles and Mobile Networks" course, I joined the map group theory group, this report mainly reported my work in the laboratory and progress. I mainly engaged in things on the academic data of community detection and community relevance research, in the laboratory to do three things: First, we first understand the model of the experiment, and participate in the modification of a small amount of code; Secondly, In the guidance of the students simply learn JavaScript, and participated in the script front-end scripting; Finally, I understand some of the t-SNE algorithm on the content, and help the seniors to find some of the available data on this algorithm.

COMMUNITY TESTING INTRODUCTION

In real life, the community is usually defined as a set of internal has close contact with a group of nodes and the nature of the common features, such as love cats who can form a community, love dog who can form another community. By the above definition that can overlap occurs between different communities, such as the cat dog who love love in the two communities of overlapping position. A difficult problem now is the discovery of communities in the network, known as community testing, as a fundamental issue of network science, which has attracted a great deal of attention in the past few decades. In recent years, with a large number of studies on large data, creates another related but different problem, called community search, its purpose is to find the most likely contains the query node community, attention of academic and industry fields.[1]

This study is carried out by us, in many models, the authors are based on the assumption that the community between the internal connection density is far greater than the density of the area between the overlapping community, which can be introduced if the K communities overlap between the two nodes in the K, the greater the probability of connecting two nodes is low. But the actual situation and the reality of life in the contrary, for example, two of the more common interests, so the two men more likely to become friends, rather than the opposite, so in the BIGCLAM [2] model, the author carried out the opposite assumption and by setting between the community and the node weights model to establish, with visible [2].

In addition, in community search, there is a classic MMSB [3] algorithm, the full name A, Mixed-Membership, Stochastic, Blockmodel. The main idea is to construct a membership vector $\theta_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{iK})$ of node i in the community, where θ_{iK} represents the probability that node i belongs to the community K, and the sum of the components of the vector is 1. For node pairs (i, j) in the network, select community indicators $z_{i \to j}$ and $z_{i \leftarrow j}$ pointing to one of the K communities. When both are k, the probability of connection for the two nodes is β_k.

T-SNE THEORY

T-distributed random adjacent embedding (t-SNE) is a machine learning algorithm for dimensionality reduction. It is a nonlinear dimensionality reduction technique, which is especially suitable for embedding high-dimensional data into 2D or 3D space. Visualize in scatter charts. Specifically, it models each high-dimensional object by two-dimensional or three-dimensional points, so that similar objects are modeled by nearby points, and non-similar objects are modeled by remote points.

The t-SNE algorithm consists of two main phases. First, the t-SNE constructs the probability distribution of the high-dimensional object pairs in such a way that similar objects have a high probability of being selected, and different points have very small selected probabilities. Second, the point in the low-dimensional map of t-SNE defines a similar probability distribution, and it minimizes the position of the Kullback-Leibler divergence between the two distributions relative to the point in the map. Note that although the original algorithm uses the Euclidean distance between objects as the basis for its similarity measure, it should be changed as appropriate. For detailed modeling, see [4].

EXPERIMENTAL CONTENT AND SIMULATION

The main principle of the preparation of the front of the map is that we first sort each of the processed images in accordance with the time, and then use JavaScript code to control each picture in a certain time interval appears on the page, the specific effect shown in the figure.

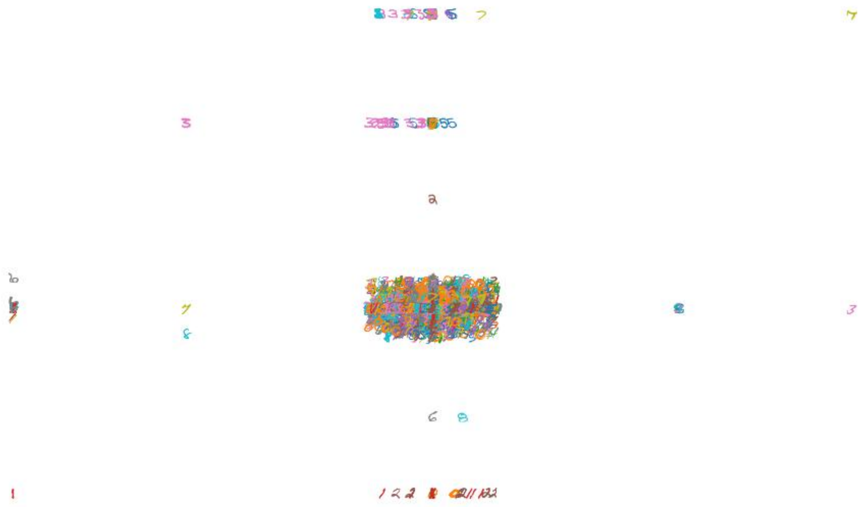


Figure 1 at the beginning of the map front-end situation

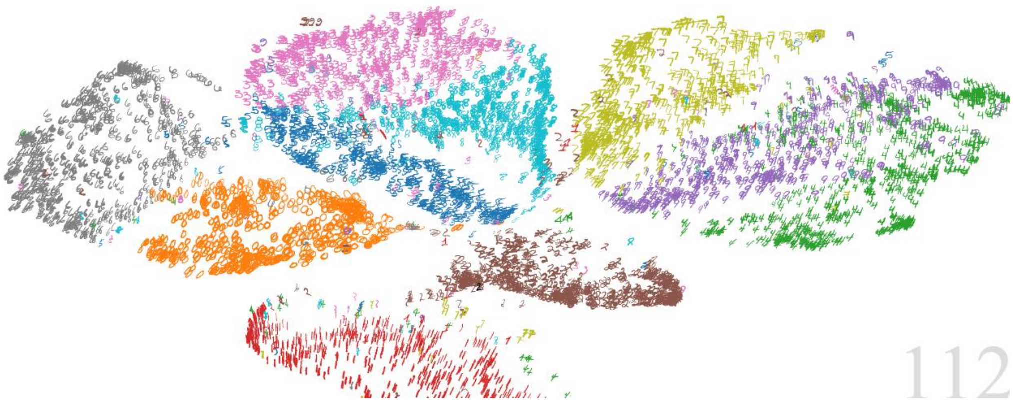


Figure 2 The situation of the map front end during the change

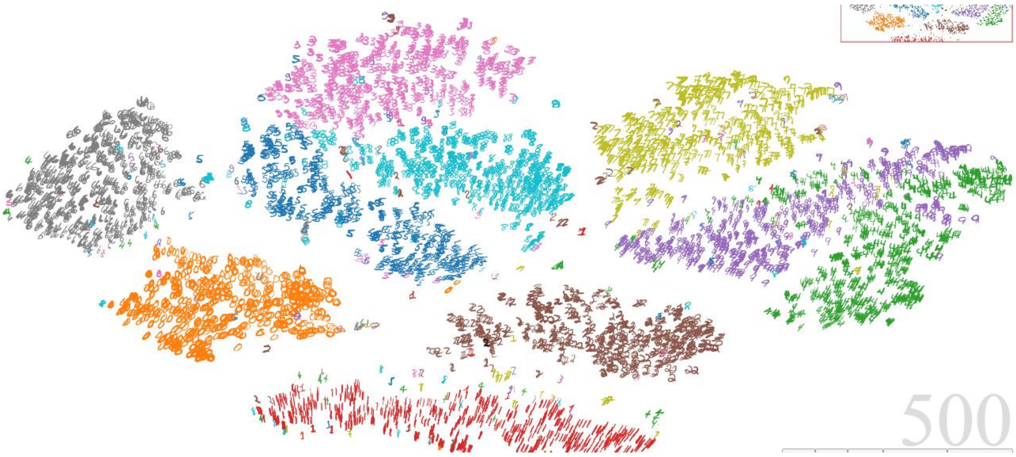
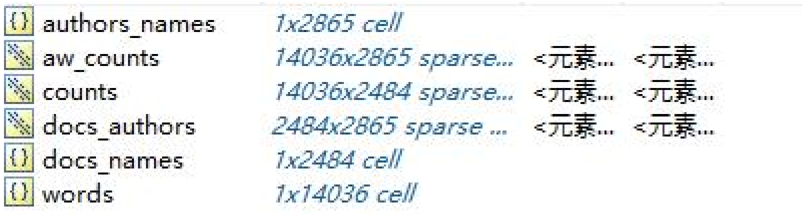


Figure 3 The final effect of the map front end

The entire map front end uses the t-SNE clustering algorithm, each point in the map is close to its corresponding class, to the final stability.

Next, I found some information on the academic articles, including the beginning and ending year for the 1988 and 2003, the data shown in the figure.



```
authors_names   1x2865 cell
aw_counts       14036x2865 sparse...  <元素...  <元素...
counts          14036x2484 sparse...  <元素...  <元素...
docs_authors    2484x2865 sparse ...  <元素...  <元素...
docs_names      1x2484 cell
words           1x14036 cell
```

Figure 4 Academic article data

Here are some of the variables on the introduction:

words: Words that appear in academic articles, which are a list of words that can analyze whether the article belongs to the same field based on the ratio of the same word in the two articles.

docs_name: the name of the article, the name of the format for the "year / article name."

Authors_name: the author's name.

docs_authors: it is the number of articles x the number of sparse binary matrices. If an article belongs to an author, the element value at the corresponding location is 1, otherwise it is 0.

counts: it is the number of words x the number of articles count matrix.

aw_counts: it is a number of words x the number of authors of the counting matrix.

The basic data of the academic map can be obtained by processing and classifying the above data, dividing it into multiple communities by using an algorithm, and attaching the result to the front of the map.

CONCLUSIONS AND FUTURE PROSPECTS

In community testing, the method we usually use is to run a set of data with a program, and then use the algorithm to get the community division. This process is called training. Then run the community with a prediction system, restore the resulting system network, and then compare the resulting system network with the original network to see if it can be restored with less error. The smaller the gap between the restored network and the actual network is, the more successful the community testing method is.

We have implemented the above process, but in the process we do not see the changes of intermediate variables, such as the BIGCLAM model between community and node weights and MMSB membership vector, can only see the scoring algorithm finally prediction system, which is not conducive to our code debugging and modification of the algorithm. So we want to make the program when the output value of the intermediate variables, but due to the large complex network, so that all the values of all output variables is not realistic, so we hope to be able to let the node can program every fixed output intermediate results, this is our expectation for the future of the program.

In this semester "wireless communication network" course, I mainly in Jia Yuting seniors and Ren Xiangyu came under the guidance of laboratory research projects have a preliminary understanding, in accordance with the arrangement of the seniors read some articles about the community search paper, and do their best to participate in some tasks, but feel from the skilled there is a big gap, should continue to study hard in the future direction of their love. Thank you very much for being able to set up such a platform for me to see the gap between you and other people. Thank you very much for the help from other teachers and students in the course.

REFERENCES

[1] Wikipedia: en.wikipedia.org

[2] Jaewon Yang and Jure Leskovec. Overlapping Community Detection at Scale: A Nonnegative Matrix Factorization Approach,2013

[3] Prem K. Gopalan1 and David M. Blei. Efficient discovery of overlapping communities in massive networks,2012

[4] van der Maaten, L.J.P.; Hinton, G.E. Visualizing Data using t-SNE,2008