Mapping Research Funding in 2D and 3D by t-SNE

Ting Chen chenting@casipm.ac.cn
Insitutes of science and development, Chinese academy of sciences
Xiaomei Wang wangxm@casipm.ac.cn
Insitutes of science and development, Chinese academy of sciences
Guopeng Li liguopeng@casipm.ac.cn
Insitutes of science and development, Chinese academy of sciences

Introduction

With an exponentially growing number of research awards funded each year, a visualization tool for exploring funding's hotspots and gaps is becoming indispensable. However, revealing the landscape of funding is a very challenging task. One researcher created the map of funding using the tree map[1]. Others used the pLSA or paragraph vector to extract relationships between all pairs of awards. Then the network maps are created with force direct layout[2,3].

The network map is the most widely used visualization method in bibliometrics, but the relationship between awards is a non-sparse distance matrix, the threshold needs to be manually set in the visualization task. Besides, when high-dimensional text features are converted into relationships between pairs, some information in the high-dimensional space will be lost.

This paper is creating a new way to sort and view the research funding by mapping high-dimensional representation of awards in a 2D and 3D space with a nonlinear dimensionality reduction technique t-SNE. 4669 NSF awards data from 2008 to 2017 were downloaded from the Information and Intelligent Systems department in this research.

Methods

Document feature extraction

After the data has been cleaned, a TF-IDF (BOW) vector is generated for each title and proposal following lemmatization and removing the stop word, the total dimensions of the TF-IDF vector space is 6000. The TF-IDF vectors space contains high dimensional sparse features that may affect the generalization performance of mapping, the original feature space is converted to a more compact new space using latent semantic analysis (LSA). The latent semantic analysis is a simple topic modeling technique used to reduce noise and create more overlap between document vectors from similar topics. To this end, the dimensionality of the TF-IDF vectors is reduced to a fixed number of components using singular value decomposition.

Mapping with t-SNE

t-Distributed Stochastic Neighbor Embedding[4] (t-SNE), which is a nonlinear dimensionality reduction technique that is well-suited for embedding high-dimensional data into 2D or 3D mapping. This is in contrast to methods such as PCA and MDS that use the same linear mapping. As a result, t-SNE provides much better visualization results than linear techniques, such as PCA or MDS over the complicated dataset technique.

The t-SNE algorithm comprises of two main stages. First, t-SNE constructs a probability distribution over pairs of high-dimensional objects, similar objects should have a higher probability of being picked, and dissimilar points should have a lower probability of being selected, t-SNE treated the probability of pairs as similarity between points. Second, t-SNE defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the Kullback-Leibler divergence[5] between the two distributions with respect to the locations of the points in the map. We ran 1000 times and selected the solution with the lowest KL divergence in this study.

Results

Before we visualize NSF awards, we divided them into 21 topics by using K-means clustering and human interpretation, the topic label of each award will be used as follow-up mapping verification references. To create the visualization map, the awards' TF-IDF and LSA vectors are respectively embedded in 2D space using t-SNE. In this case, 20 dimensions were used to represent LSA vectors according to K-means clustering result.

The comparison of TF-IDF(A) and LSA(B) with t-SNE mapping results are shown in Fig1; each dot represents an award in 2D space, the color is the pre-classified topic label of the award. As we can see, the map of LSA and t-SNE had better class-separability than TF-IDF. It successfully uncovers hidden structures in the data, exposing the pre-classified award topics.
A 3D space allowed the map to represent more detailed structures than 2D; the t-SNE algorithm can easily map high-dimensional data into any low-dimensional space. Therefore, the author constructed a 3D interactive funding map (Fig 2). The user can zoom and rotate the map in 3D space, as well as click the dot to see more detail and locate the most similar awards by measuring the Euclidean distance in 3D space.

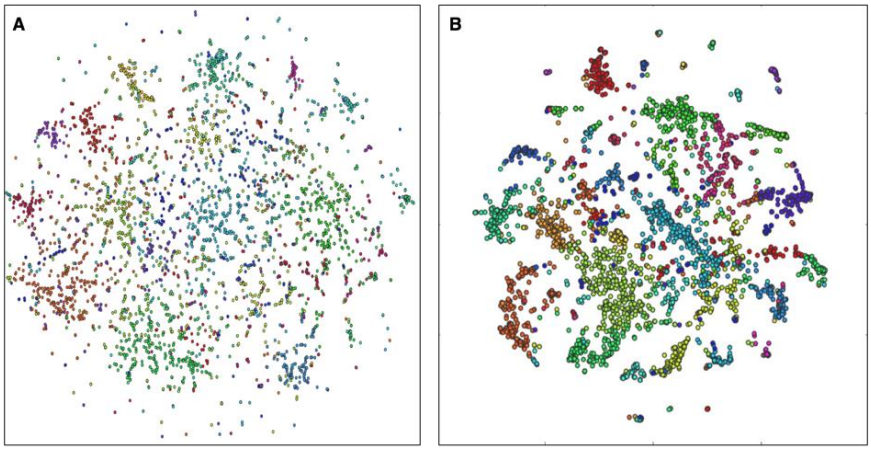

Fig 1. Mapping TF-IDF(A) and LSA(B) vectors in 2D space

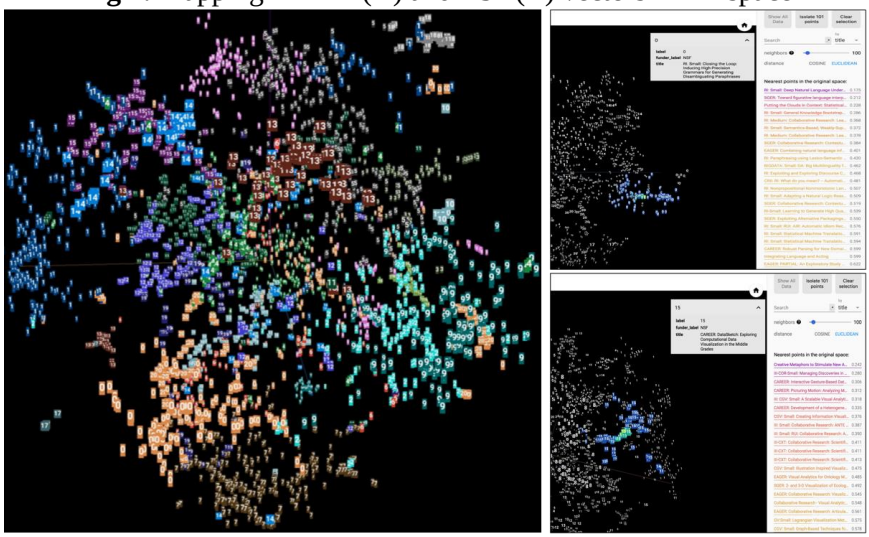

Fig 2. Interactive funding map in 3D space

Discussion

The t-SNE algorithm with LSA Feature extraction provides an effective method to visualize the funded awards. It successfully uncovered hidden structures in NSF awards data, exposing natural topics. By zooming and clicking in the 3D map, decision makers could observe the funding hotspots and funding gaps effectively. Also, we can apply the clustering algorithm directly in the map to discover funding topics, such as K-means, DBSCAN. As this study is only at the explorative stage, we will further verify the methods in different datasets of different funding agencies. Furthermore, we plan to perform additional comparisons with multiple funding sources, such as NSF, NSFC (China) and Framework Programme (EU).

References

[1]Liu S, Cao N, Lv H. Interactive Visual Analysis of the NSF Funding Information[C]// Visualization Symposium, 2008. PacificVIS '08. IEEE Pacific. IEEE, 2008:010502.

[2]Ii B W H, Talley E M, Burns G A P C, et al. The NIH Visual Browser: An Interactive Visualization of Biomedical Research[C]// Information Visualisation, 2009, International Conference. IEEE, 2009:505-509.

[3]Takahiro K, Katsutaro W, Naoya M. Funding Map for Research Project Relationships using Paragraph

Vectors[C], 16th International Conference on Scientometrics & Informetrics (ISSI), Wuhan, China, 2017.
[4]Maaten L V D, Hinton G. Visualizing Data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(2605):2579-2605.
[5]Kullback S, Leibler A. On Information and Sufficiency[J]. Annals of Mathematical Statistics, 1951, 22(1):79-86.