Using Full-text of Academic Articles to Find Software Clusters

Heng Zhang[1] , Shutian Ma[2] and Chengzhi Zhang[3, *]

[1]zh_heng@njust.edu.cn
Department of Information Management, Nanjing University of Science & Technology, Nanjing 210094(China)

[2]mashutian0608@hotmail.com
Department of Information Management, Nanjing University of Science &Technology, Nanjing 210094(China)

[3]zhangcz@njust.edu.cn
Department of Information Management, Nanjing University of Science &Technology, Nanjing 210094(China)

Abstract

Scientific software is making contributions to modern science. To meet huge academic demands such as data analysis, modelling, visualization and so on, various software has been developed to help different steps in scientific work. In order to reveal the connections between scientific software, we conduct cluster analysis among scientific software based on the full-text data of 23,120 articles published in PLOS ONE. Firstly, we select some popular software whose mention times are over 50 to be our candidate software list for clustering analysis. Secondly, Word2Vec is applied to learn distributed representation for each software. Then, we apply Affinity Propagation to cluster software and tune different parameters to obtain better results. Silhouette coefficient is computed here to evaluate clustering performance under each parameter setting. According to our optimal results, software clusters with specific functions can be found. And software which have strong linkage between each other are mainly have functions in common.

Keywords: Scientific Software, Software Clustering, Distributed Representation

Introduction

Scientific software is a critical component in academic researches. It can analyze data, simulate the physical world and visualize the results, one single scientific work will need the support of several software with specific functions. As the unsung heroes (Chawla, 2016), researchers are paying attention to investigate what kind of important role it plays in the advancement of science. For example, Callahan et al. (2018) developed u-Index metric to measure the impact of informatics tools and databases. Smith et al. (2016) discussed on software citation principles that may encourage relevant policies for software citation across disciplines and venues. Since more and more full-text of literature has becoming accessible, some studies are focusing on utilization of text mining for this topic. Duck et al. (2016) identified mentions of databases or software in the PubMed Central full-text corpus through text mining. However, few of the researches are relevant about mining the connections between different software. So, in this paper, we try to reveal the connections between software by cluster analysis.

Methodology

Framework of our study

The main purpose of this paper is to reveal the connections between scientific software. As shown in Figure 1, firstly, from PLOS ONE, we collected 23,120 articles published in 2017. Our original software list comes from the previous work (Pan et al. 2015) and we further filtered out those that was mentioned in less than 50 articles. So, we get 260 software to be the candidate software list for cluster analysis. Secondly, Word2Vec (Mikolov et al. 2013), which can learn high-quality word vectors from huge corpora, is applied to learn distributed representation for these 260 software using full-text. Then, we apply Affinity Propagation (AP) to cluster software using the vector data. Finally, we analyze the characteristics of top-5 clusters which contain the largest number of software. Since the clustering is conducted using software vectors learned by Word2Vec, we want to investigate that if there really exist any relations between those software pair with strong linkage. Here, strong linkage refers to high cosine similarity between software pairs based on their distributed representations.

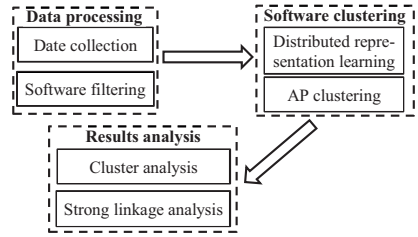

Figure 1. Framework of our work

Clustering evaluation

After clustering, we apply silhouette coefficient to evaluate the performance. Bigger silhouette coefficient value means better cluster result. Affinity Propagation is a clustering algorithm based on similarity matrix of data points (Frey et al. 2007). In

* Corresponding author.

this algorithm, *preference* is an important parameter, which controls how many exemplars are used. We set *preference* values from 0 to 0.9 and 0.1 as interval. and the clustering result with the maximum silhouette coefficient was used for the final analysis.

Results analysis

Cluster analysis

Silhouette coefficient of all different clustering results are shown in Table 1. We then further analyze the clusters obtained when the *preference* value is 0.4. The top-5 clusters with the largest number of software are shown in Table 2.

Table 1. Silhouette coefficient of clustering results in different *preference* values

preference	Silhouette coefficient	preference	Silhouette coefficient
0	0.1139	0.5	0.1445
0.1	0.1234	0.6	0.1256
0.2	0.1311	0.7	0.0992
0.3	0.1492	0.8	0.0417
0.4	**0.1576**	0.9	0.0058

Table 2. Top-5 clusters with the most software in the optimal clustering result

Clusters	Software
1	*SPSS, Prizm, Stata, SigmaPlot, Systat, MedCalc, StatSoft, G*Power, PASW, OriginLab, Minitab*
2	*ImageJ, Image ProPlus, NIS Element, AxioVision, Imaris, Aperio, MetaMorph, Feature Extraction, LAS AF, Leica Application Suite, Volocity*
3	*BLAST, SMART, Pfam, STRING, SignalP, Blast2GO, MEME, PANTHER, TMHMM, InterProScan*
4	*MEGA, MrBayes, RAxML, BEAST, STRUCTURE, FigTree, PAUP, TreeAnnotator, Modeltest*
5	*Clustal W, Geneious, MUSCLE, BioEdit, MAFFT, FASTA, Clustal X, Clustal Omega, Sequencher*

Firstly, we find that the software in each cluster is similar in function. In Cluster 1, software are mainly used for statistical analysis. Image processing software are gathered in Cluster 2. Software in Cluster 3 are relevant to protein research while software in Cluster 4 are used more in the study of heredity and evolution. Function of most software in cluster 5 is about multiple sequence alignment for DNA.

Strong linkage analysis

In order to understand more detailed relationships between software, we take software Clustal W as an example and analyze the other software which shows strong linkage with it. There are three software have high cosine similarity (>0.8) with Clustal W, they are Clustal Omega, MUSCLE and Clustal X. In gene sequencing domain, all of them are used for multiple sequence alignment. Besides, Clustal X, Clustal Omega and Clustal W are different versions of Clustal. MUSCLE is an alternative software of Clustal W. But in terms of the accuracy and speed of multiple sequence alignment, MUSCLE is better than Clustal W.

Conclusions

In this paper, we conduct AP clustering for 260 popular software using full-text from PLOS ONE. According to our experimental results, our method can find software clusters with specific functions and the software tend to have functional similarities between each other within each cluster. Since we used software mentioned times when selecting software for analysis, hidden research topics over current publication collection can also be inferred from these popular software based on their functions, such as protein analysis and DNA alignment.

Acknowledgments

This work is supported by Major Projects of National Social Science Fund (No. 17ZDA291), Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. SJCX18_0136), and Qing Lan Project.

References

Singh Chawla, D. (2016). The unsung heroes of scientific software. *Nature, 529*(7584), 115-116.

Callahan, A, Winnenburg, R. , & Shah, N. H. . (2018). U-index, a dataset and an impact metric for informatics tools and databases. *Scientific Data*, 5, 180043.

Smith AM, Katz DS, Niemeyer KE, FORCE11 Software Citation Working Group. 2016. Software citation principles. *PeerJ Computer Science 2*: e86. https://doi.org/10.7717/peerj-cs.86

Duck G, Nenadic G, Filannino M, Brass A, Robertson DL, et al. (2016). A Survey of Bioinformatics Database and Software Usage through Mining the Literature. *PLOS ONE 11*(6): e0157989. https://doi.org/10.1371/journal.pone.0157989

Pan, X, Yan, E, Wang, Q & Hua, W. (2015). Assessing the impact of software on science: a bootstrapped learning of software entities in full-text papers. *Journal of Informetrics, 9*(4), 860-871.

Mikolov, T, Chen, K, Corrado, G, & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Frey, B. J, & Dueck, D. (2007). Clustering by passing messages between data points. *Science, 315*(5814), 972-976.