ATLANTIS PRESS

Advances in Engineering Research, volume 138

2nd International Conference on Machinery, Electronics and Control Simulation (MECS 2017)

Research on Text Mining of Biomedical Field Based on Pubmed

Kang Li[1, a,*], Weidi Dai[1, b], Wenjun Wang[1, c], Ruixin Song[1, d]

[1]School of Computer Science and Technology, Tianjin University, Tianjin 300350, China;

[a]likang@tju.edu.cn, [b]wwj@pku.org.cn, [c]davidy@126.com, [d]afei@tju.edu.cn

Abstract: The field of biomedical science is one of the most studied areas of the 21st century, the field has published a huge number of research papers, which have averaged more than 600,000 articles a year. How to effectively obtain knowledge in the vast literature of research is a challenge for researchers in the field. As one of the branches of bioinformatics, technology of text mining of biomedical field is a new exploration of the efficient and automated acquisition of relevant knowledge. In this paper, we take the example of genetic enhancement, articles related to Genetic Enhancement from 2005 to 2016 were selected as datasets. We focus on two approaches, co-word analysis is used to identify research hotspot and complex network analysis is used to analyze the collaboration network to determine the status of research collaboration in this field. From co-word analysis, we find that there are four research hotspots, they are gene expression related field, bioethical issues related field, metabolic and protein engineering related area and cell related field. From research collaboration analysis, we find that the research collaboration network has a scale-free feature and the max connected subgraph of it has a small world phenomenon. We also find that research collaboration in this field is now recovering.

Keywords: Biomedical literature, research hotspot, co-word analysis, research collaboration, complex network analysis.

1. Introduction

At present, biomedical research is developing rapidly, a large amount of biomedical knowledge exists in unstructured forms in various forms of text. The total number of documents in the MEDLINE(Medical Literature Analysis and Retrieval System Online) database which is the most authoritative one in the biomedical field has reached 1.6 million. In recent years, more than 600,000 articles have been published annually. Therefore, how to effectively use the information and knowledge contained in these texts is of great importance in analyzing vast amounts of biomedical data. The traditional way is to search through the database or the Internet by keyword, however, this approach can't obtain knowledge from the text. Text mining can help the researchers to obtain implicit knowledge or information from the text automatically. Using text mining can effectively extract knowledge from biomedical database for further research, therefore, text mining has great application value in the biomedical field.

In this paper, the text of biomedical field is analyzed by means of text mining and focuses on two approaches. On the one hand, we use co-word analysis to identify research hotspot. The co-word analysis is mainly for pair words. Every two words is a pair and they are counted the numbers of times in the text. Then hierarchical clustering is carried out on the basis of co-word matrix to reveal the relationships which can be used to analyze the structural changes in the topics and subjects they represent. On the other hand, we use complex network analysis to analyze the scientific collaboration network, in order to determine the status of cooperation in this field. In a scientific collaboration network, if two scientists publish a collaborative document, it is defined as a link between them. In addition to the collaboration between authors, the network also includes the number of collaborators, the number of collaborative papers, the degree of clustering, and so on. Complex network analysis has put forward several quantitative analysis indexes, which mainly include degree, closeness, betweeness, community, etc. The author, subject force and group distribution can be reflected through these indicators.

2. Datasets

The PubMed database is a web-based biomedical information retrieval system developed by the National Institute of Biotechnology Information (NCBI), which is owned by the National Library of Medicine (NLM) in 2000 and the source of the literature is from MEDLINE which includes more than 5000 journals over 40 languages from nearly 70 countries since 1966. The Pubmed database provides the index, summary and full links of the literature which covers every area of biomedical science.

In this paper, we take the example of Genetic Enhancement research. Articles related to Genetic Enhancement from 2005 to 2016 were selected, the search strategy is Genetic Enhancement[Mesh] AND("1996 /01 /01"[PDat]:"2016 / 12 /31" [PDat]). A total of 1380 articles were retrieved and there are 8010 MESH words and 3310 authors in them. The details of the datasets are shown in table 1.

Table 1 Genetic Enhancement Datesets

Number of Articles	Number of MESH words	Number of authors	Time range
1380	8010	3310	2006-2015

3. Research hotspot analysis

3.1 Data pretreatment

In this part, first, we counted the word frequency. Then, we computed the boundaries of high-frequency and low-frequency words according to zipf's law and we got the high frequency list. Furthermore, according to the high frequency list, we counted the word frequency in every two words in every article so that we got the co-word matrix. Finally, correlation matrix is got by correcting co-word matrix with ochiai similarity.

3.1.1 Word frequency statistics

Each of the articles in the Pubmed database has 8-15 MESH words which are comprehensively summarized and described in terms of the content and form of the article. For the articles obtained, we counted the word frequency of each article and sorted them by frequency from high to low. In order not to affect research hot spots, we filter words like gender, country and age.

After filtering, statistics, sorting, the results are as follows in table 2. Given the length of paper, we only list a few words.

Table 2 Result of word frequency(Top 20)

Numble	MESH words	Word frequency
1	Protein Engineering	216
2	Recombinant Proteins	190
3	Genetically Modified	132
4	Escherichia coli	130
5	Genetic	72
6	Saccharomyces cerevisiae	58
7	Bacterial	57
8	Gene Expression Regulation	52
9	Genetic Engineering	49
10	Species Specificity	47
11	Signal Transduction	47
12	Escherichia coli Proteins	44
13	Molecular	44
14	Cloning	41
15	Ethanol	40
16	Glucose	37
17	Cell Proliferation	37
18	Bacterial Proteins	36
19	Transfection	35
20	Metabolic Engineering	35

3.1.2 Selection of high-frequency words

To determine the high frequency words that reflect research hot spots, we use zipf's law to get the threshold value that distinguish between high-frequency and low frequency words. The formula of zipf's law is as follows:

$$N = \frac{1}{2}(-1+\sqrt{1+I_1}) \tag{1}$$

N means the threshold value that distinguish between high-frequency and low frequency words, and I_1 means the number of words whose frequency is one. According to the formula, we get the threshold value $N = 78$, which means a MESH word is counted as a high frequency word only if its word frequency is greater than 78. According to the statistics, we identified 67 high frequency words at last.

3.1.3 Construct co-word matrix and correlation coefficient matrix

We constructed a 67×67 co-word matrix based on the 67 high frequency words, then we took the ochiai similarity and turned it into correlation coefficient matrix. The result is shown in table 3.

Table 3 correlation coefficient matrix(part of the matrix)

Numble	1	2	3	4	5	6	7	8
1	1.0000	0.6367	0.0000	0.3580	0.0000	0.2322	0.0000	0.1604
2	0.6367	1.0000	0.0000	0.3690	0.0000	0.2381	0.0000	0.1810
3	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	0.3580	0.3690	0.0000	1.0000	0.0000	0.0575	0.0000	0.1702
5	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
6	0.2322	0.2381	0.0000	0.5758	0.0000	1.0000	0.0000	0.3641
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
8	0.1604	0.1810	0.0000	0.0575	0.0000	0.1702	0.0000	0.9423

3.2 Hierarchical cluster analysis

The basic idea of hierarchical cluster analysis is to cluster variables of similar magnitude into one cluster. Variables that are closely related are aggregated into a small category, variables that are estranged are aggregated into a large category. The process is not stopped until all the variables are aggregated. Finally, the whole classification system is described as a hierarchical graph, which show the relative variation between all the variables. The cluster dendrogram of high frequency MESH words is shown in Fig-1.

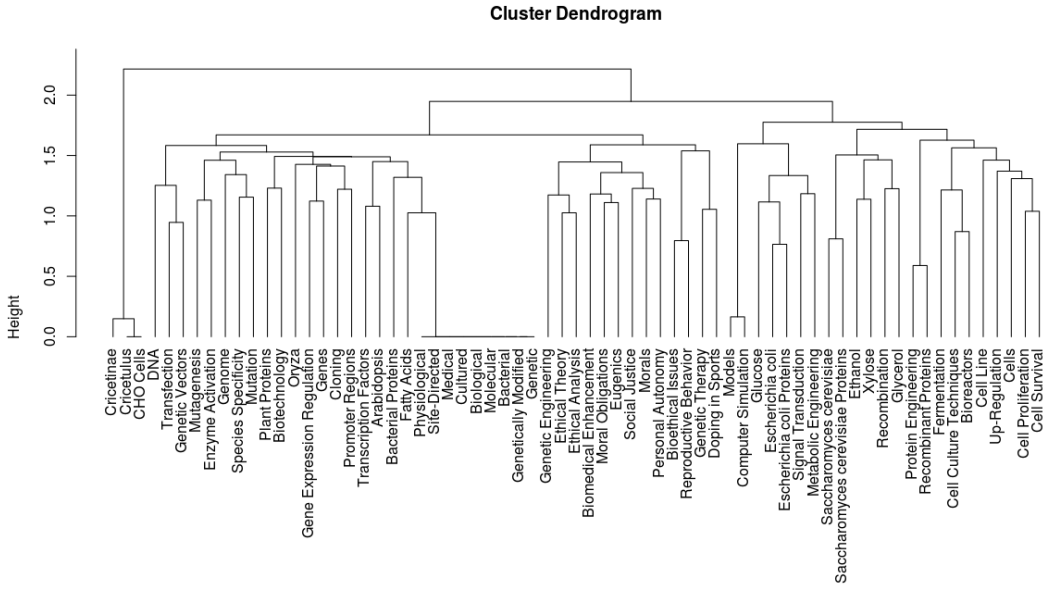


Fig. 1 Cluster dendrogram of high frequency MESH words of genetic enhancement field

4. Research collaboration analysis

4.1 Construct research collaboration network

In a scientific collaboration network, if two scientists publish a collaborative document, it is defined as a link between them. According to the definition, we extracted the authors of each paper and constructed an edge in every two authors who publish a collaborative document. So, we got the research collaboration network related to genetic enhancement field. The network topology index is shown in table 4. From the table, we know that the average clustering in max connected subgraph is very high, while the average shortest path length is low, which indicates that the max connected subgraph has a small world phenomenon.

Table 4 Three Scheme comparing

Number of nodes	Number of edges	Network density	Average clustering in max connected subgraph	Average shortest path length in max connected subgraph
3021	9255	0.002028	0.932108	2.272556

Because the research collaboration network is not a connected network, we got the max connected subgraph of research collaboration network in order to analyze more topology properties in the network. The Max connected subgraph of research collaboration network is shown Fig-2.

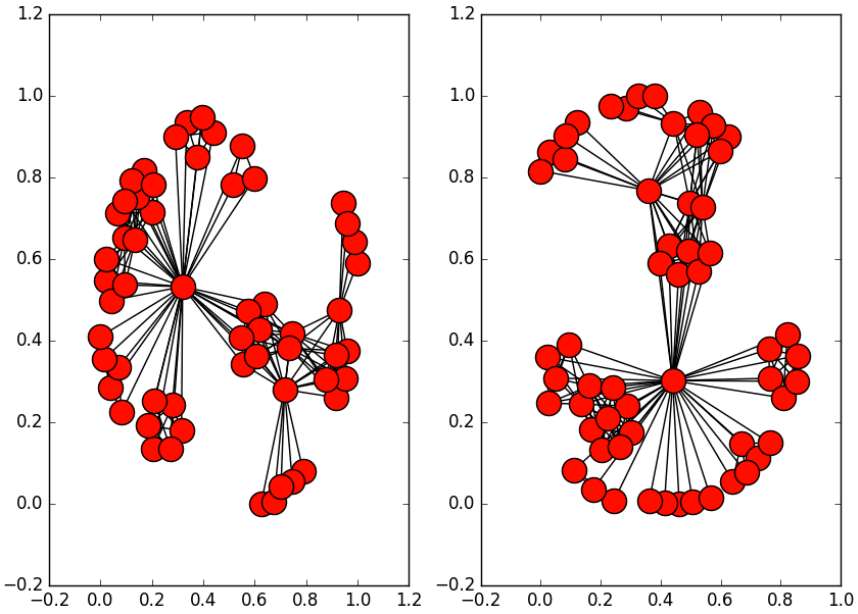


Fig. 2 Max connected subgraph of research collaboration network

4.2 Statistic analysis

In this part, we use three statistical analysis methods to analyze the current status of research collaboration in genetic enhancement field. They are degree analysis, k-core analysis and collaboration trend analysis.

4.2.1 Degree analysis

Degree distribution of research collaboration network is shown in Fig-3. From the degree distribution we know that the connections between the nodes have a severe uneven distribution. Nodes with low connections make up a large proportion of these nodes and the degree distribution of the nodes is in the power distribution. So, the research collaboration network has scale-free features.

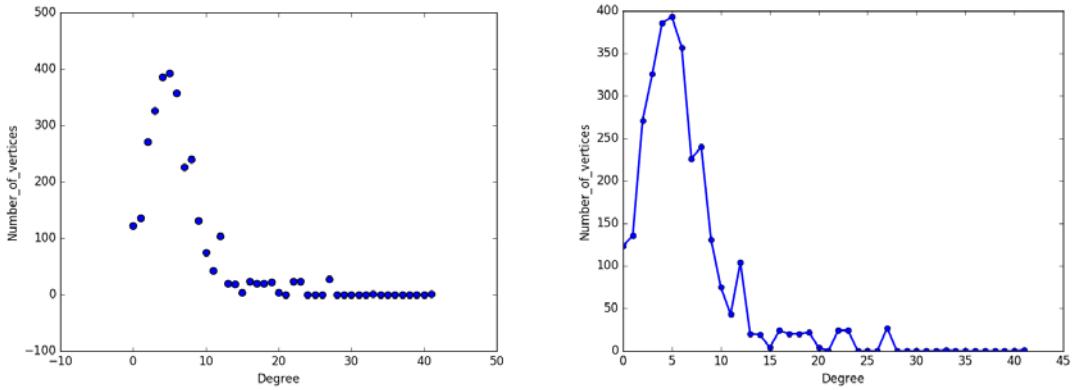
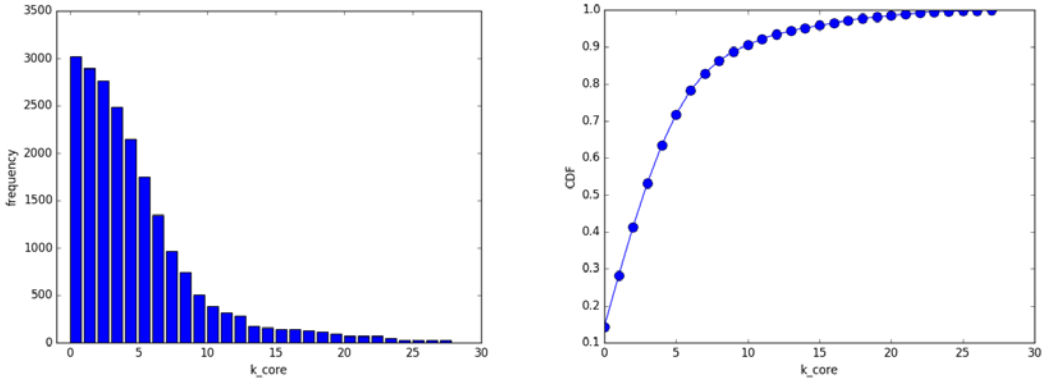


Fig. 3 Degree distribution of research collaboration network

4.2.2 K-core analysis

K-core distribution of research collaboration network is shown in Fig-4. From the cumulative distribution function we know that 90% of the author belong to the network with k-core less than10. It shows that the research collaboration network has a wide range of cooperation.

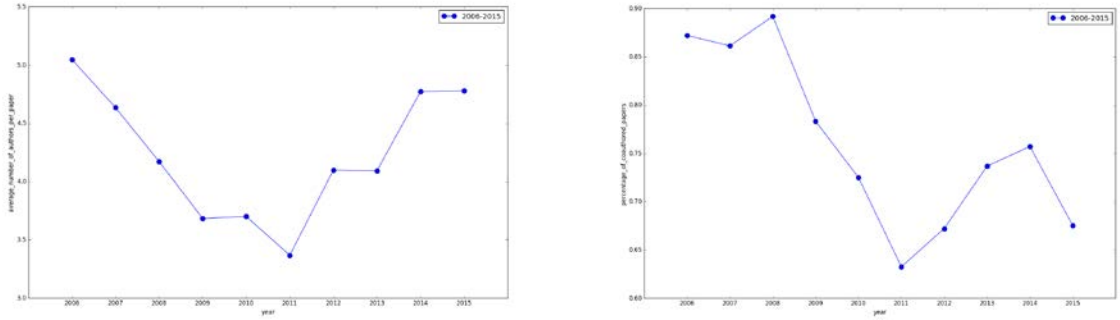


(a) (b)

Fig. 4 K-core analysis results. (a) K-core distribution of research collaboration network. (b) CDF of k-core distribution.

4.2.3 Collaboration trend analysis

Trend analysis results is shown in Fig 5-a. From the picture, we know that the number of authors per paper and percentage of co-authored papers has risen from 2011.



(a) (b)

Fig. 5 Trend analysis results. (a) Average number of authors per article from 2006-2015. (b) Percentage of co-authored papers per year from 2006-2015.

5. Conclusions

From the research hotspot analysis, we can draw the conclusion that there are four research hotspots, they are Gene expression related field, bioethical issues related field, metabolic and protein engineering related area and cell related field. From the research collaboration analysis, we find that the research collaboration network has a scale-free feature and the max connected subgraph of it has a small world phenomenon. We also find that research collaboration in this field is now recovering.

6. Acknowledgments

This work was financially supported by the National Social Science Fund(15BTQ056).

References

[1] Siamaki S, Geraei E, Zarefarashbandi F. A study on scientific collaboration and co-authorship patterns in library and information science studies in Iran between 2005 and 2009.[J]. International Journal of Health Promotion & Education, 2014, 3:99.
[2] Gaskó N, Lung R I, Suciu M A. A new network model for the study of scientific collaborations: Romanian computer science and mathematics co-authorship networks[J]. Scientometrics, 2016, 108(2):613-632.
[3] Okamoto J. Scientific collaboration and team science: a social network analysis of the centers for population health and health disparities[J]. Translational Behavioral Medicine, 2015, 5(1):12-23.
[4] Hou X N, Hao Y F, Cao J, et al. Scientific Collaboration in Chinese Nursing Research: A Social Network Analysis Study.[J]. Computers Informatics Nursing Cin, 2015, 34(1):47.
[5] Sin S C J. International coauthorship and citation impact: A bibliometric study of six LIS journals, 1980–2008[J]. Journal of the Association for Information Science and Technology, 2011, 62(9):1770–1783.
[6] Soria-Guerra R E, Nieto-Gomez R, Govea-Alonso D O, et al. An overview of bioinformatics tools for epitope prediction[J]. Journal of Biomedical Informatics, 2015, 53(C):405-414.
[7] Ravikumar S, Agrahari A, Singh S N. Mapping the intellectual structure of scientometrics: a co-word analysis of the journal Scientometrics (2005–2010)[J]. Scientometrics, 2015, 102(1):929-955.
[8] Gan C, Wang W. Research characteristics and status on social media in China: A bibliometric and co-word analysis[J]. Scientometrics, 2015, 105(2):1167-1182.
[9] Liu W J, Geng Y, Tian X, et al. A review of emergy studies based on bibliometric and social network analysis[J]. Journal of China Agricultural University, 2016..
[10] You G R, Sun X, Sun M, et al. Bibliometric and social network analysis of the SoS field[C]// International Conference on System of Systems Engineering. IEEE, 2014:13-18.
[11] Sedighi M. Using co-word analysis method in mapping of the structure of scientific fields (case study: The field of informetrics)[J]. Iranian Journal of Information Processing Management, 2015, 30(2):373-396.
[12] Chen C Y, Hua-Zhu W U, Chun-Mei P, et al. Subject Study of Papers on Sci-tech Novelty Retrieval in China Based on Co-word Analysis[J]. Journal of Library & Information Sciences in Agriculture, 2016.