A Multi-dimensional Scientometrics Analysis on Alfalfa Research

Heying Zhu1, Yanxiang Zong1, Yanping Yang2, Huoyou Li3, Yu Dong2 and Qingqiang Wu1*
1 School of Informatics, Xiamen University, Xiamen, Fujian, 361005, China
2 National Science Library, Chinese Academy of Sciences, Beijing, Beijing, 100000, People’s Republic of China
3 Informatics Technology, Longyan University, Longyan, Fujian, 360802, People’s Republic of China
*Corresponding author’s e-mail: wuqq@xmu.edu.cn

Abstract. Alfalfa is known as "the king of forage". It not only has high yield, but also has good quality. Studying the development process of alfalfa will help scholars in this field grasp the future research direction. In this paper, the data of alfalfa field from 2009 to August 14, 2020 retrieved on Web of Science is taken as the research object. Firstly, this paper analyses the current research status by using basic statistical methods. The selected analysis dimensions include time, country, funding source etc. Then, we conduct topic clustering research based on Latent Dirichlet Allocation (LDA) model and the scientometrics method to compare and analyze its interests from both the overall and time perspectives. Finally, the experimental results are presented in diagrams. The results reveal that the development of alfalfa field tends to be stable. China and the United States have made remarkable achievements in the field of alfalfa. The topics integration and decomposition trend is weak, and the research content has strong coherence.

1. Introduction
Alfalfa is the most widely planted legume pasture in the world and is known as the "king of pasture". It is also an important source of high-quality protein for animal husbandry. In addition, alfalfa can also bring many environmental benefits to people, such as increasing soil carbon sequestration, preventing the loss of nitrogen and phosphorus, improving crop nutrient access, and breaking the life cycle of pests [1]. In recent years, as the world's demand for high-quality dairy products (such as grass-fed livestock products) continues to grow, the global demand for alfalfa has also increased [2]. Therefore, the current problem to be solved is to analyze the progress of alfalfa research and explore future research directions.

Bibliometrics is one of the scientific methods for studying the evolution of topic, including word frequency analysis, co-occurrence analysis, and co-citation analysis. It regards the literature as the object of data analysis and can help scholars understand the development of the subject and grasp the frontier direction of the subject in time. "Co-occurrence" refers to the phenomenon that the information described by the feature item of the document co-occurs, which can effectively discover some hidden information in complex networks and has been widely used in many fields[3-5]. Co-word analysis is an important branch of co-occurrence analysis. It is generally believed that the more the lexical pairs appear in the same document, the closer the relationship between the two topics is. The common statistical methods used in co-word analysis include cluster, association, word frequency, burst word monitor, etc.
The topic clustering research on the literature is a measure of the research content of scientific research papers. The topic model is a statistical model [6] that is used to discover "potential topics" that exist in a collection of documents. Latent Dirichlet Allocation (LDA) can extract the semantic information hidden from discrete data. It is an unsupervised probability generation model. The LDA topic model has great advantages in topic discovery. Since its introduction, it has been used in the analysis and application of literature topics[7-9].

Using bibliometrics method to carry out thematic clustering research on alfalfa literature will help us analyze the research interest of alfalfa and guide future research. This paper collects the literature in the alfalfa field, and draws a chart of the results through basic statistical analysis and subject clustering to analyze the development process and future development direction of the alfalfa field.

2. Method
In this paper, the process of data processing was divided into four parts:

- **Data localization**: Download literature data in the alfalfa field from the thesis database according to the search expression. After data splitting, field extraction, and data deduplication, the data is imported into the local database.
- **Data specification**: Use the synonym table, expert knowledge, and thesaurus to standardize the extracted fields into a standard form.
- **Data analysis**: This part is to analyze standardized data with domain knowledge. The analysis methods used include basic statistical analysis, co-occurrence analysis, LDA topic cluster analysis, and expert interpretation.
- **Visualization**: Display the analysis results in the form of intuitive charts by using tools.

The specific process of the above data processing method is shown in figure 1.

![Figure 1. Research framework.](image)

3. Experiments
The detailed experimental flow chart is shown in figure 2. The code used in experiment is available online1.

1 https://github.com/Thinking-Lab-XMU/Scientometrics/tree/master/Alfalfa
3.1. Dataset
The data of this paper is collected from the web of science core collection paper (SCI paper for short) in the Science Citation Index Expanded database of web of Science (WOS). By the end of August 14th in the year 2020, 5035 retrieval papers were obtained with the retrieval time range from 2009 to 2020. When searching for alfalfa data, considering its English and Latin expressions, the final search expression is written as follows:

\[(TI=(Medicago \text{ or alfalfa or lucerne or } "\text{purple medic}" \text{ or burclover or } "\text{bur clover}" \text{ or medick or barrelclover or } "\text{barrel clover}" \text{ or } "\text{barrel medic}" \text{ or } "\text{gama medic}"") \text{ OR } AK=(Medicago \text{ or alfalfa or lucerne or } "\text{purple medic}" \text{ or burclover or } "\text{bur clover}" \text{ or medick or barrelclover or } "\text{barrel clover}" \text{ or } "\text{barrel medic}" \text{ or } "\text{gama medic}"") \text{ NOT } TS=(\text{alfalfa-free or Kriens-in-Lucerne or Lucerne-seminase or } "\text{radio lines: galaxies}" \text{ or ALFALFA-Dw1 or ALFALFA-UCHVC or galaxies})\]

After removing duplicate and irrelevant data, 5011 articles remain.

3.2. Basic statistical analysis
For the 5,011 articles, we have made basic statistics based on some basic characteristics of the documents, including the number of articles issued year by year, country, funding sources. The result is shown in the figure 3.
As shown in figure 3, the number of papers published in the field of alfalfa is on the rise, especially in recent years. The two most active countries in the alfalfa field are the United States and China. And the proportion of papers published alone is high. The annual number of papers issued in the United States is around 90-120. Since 2009, especially in 2015-2016, the number of papers issued has increased significantly in China. In 2016, China successively released the "National Herbivorous Animal Husbandry Development Plan (2016-2020)", "National Dairy Industry Development Plan (2016-2020)" and "National Alfalfa Industry Development Plan (2016-2020)". Began to vigorously support the improvement of the market competitiveness of grass products and livestock products, and promote the sustainable and healthy development of the pasture industry including alfalfa. The National Natural Science Foundation of China began to increase investment and support research and development in the alfalfa field during 2011-2012. And in recent years (2018-2020), it has reached an annual output of about 100 articles. This shows that the vigorous development of a field is also closely related to national policy support. It can be seen the importance of the alfalfa field to the development of agriculture in our country.
In contrast, other sources of funding for the alfalfa field are average and stable. This may indicate that the development of alfalfa in other countries/regions are stabilizing or it is not its main research field.

3.3. Domain topic identification

In this paper, at first, we use the method “bag of words” to treat each document as a word frequency vector. In this paper, using the 17537-dimensional vector obtained by splitting and cleaning the ID and DE fields of 5011 papers as the feature vector, calculate a 0-1 matrix for each paper. After expert judgment, set the number of topics to 10($\alpha=1, \beta=0.02$). After a maximum of 1000 iterations, the model is trained to obtain the preliminary topic training results and the literature topic distribution matrix. After expert interpretation, the research topics and the number of papers included in the 10 topics are shown in table 1.

Table 1. The result of topic clustering.

| Research Topic                          | ID | Number of papers |
|-----------------------------------------|----|-----------------|
| Bioremediation                          | 1  | 486             |
| Mechanism of mycorrhizal symbiosis      | 2  | 690             |
| Field production1                       | 3  | 295             |
| Insect ecology                          | 4  | 417             |
| Feeding value                           | 5  | 562             |
| Seed treatment                          | 6  | 336             |
| Bioactive ingredients                   | 7  | 433             |
| Field production2                       | 8  | 615             |
| Resistance to abiotic stress            | 9  | 549             |
| Protein genomics                        | 10 | 628             |

3.4. Stage topic analysis

A total of 12 years from 2009 to 2020 are divided into 3 stages according to a slice of 4 years. Analyzing the changes in research topics at each stage will help researchers grasp the changes in the development direction of alfalfa in the past 12 years, and explore potential research directions.

Comparison figure 4 and table 1, it can be seen from that, in the three stages, the similarity between...
topics with the same name is higher, such as Mechanism of mycorrhizal symbiosis. On the other hand, because LDA clustering will produce clusters with similar sizes, which may contain more noise, the similarity ratio between topics with the same name is only higher than the similarity between others. It is worth noting the following three points:

- "Growth development and synthesis regulation" appeared in both the first and second stages, but the similarity between the two topics is not high. This topic disappeared in the third stage, and it has the highest similarity with "Mechanism of mycorrhizal symbiosis".
- The two topics of "Bioremediation" and "Field production" appeared separately in the first and third stages, but in the second stage, the two topics were merged into a topic that cannot be named precisely. It shows that the research content of the two topics at this stage may be relatively similar, and the research scale is reduced.
- "Seed treatment" first appeared in the second stage, and it remained the same in the third stage. "Biologically active ingredients" appear in the third stage, and have the highest similarity with the topics that cannot be named precisely in the second stage, which indicates that they are emerging research topics at this stage.

4. Conclusion
The scientific literature is one of the final presentation methods of most scientific research projects. In this paper, the relevant methods of bibliometrics are used to carry out basic statistical analysis of the scientific literature in the field of alfalfa and combined with the LDA topic clustering model to analyze the topic and its changing process. It can be seen from the results of basic statistical analysis that China and the United States have outstanding achievements in the field of alfalfa, and the number of documents and sources of funding for documents has obvious advantages compared with other countries or regions. From the results of LDA topic clustering, we can see that the evolution of topic in the past 12 years is relatively small, and the inheritance of topics is better. The research method proposed in this article can be used as a reference for other research fields. The limitation of this article is that co-word analysis is very sensitive to word selection and requires certain expert knowledge.

References
[1] USDA, P. H. Z. M. (2013). Agricultural research service. US Department of Agriculture. http://planthardiness.ars.usda.gov. Accessed, 1.
[2] Pankhurst, G. (2018). A preliminary investigation of the effects of hoof shape on the health of the equine hoof (Doctoral dissertation, Charles Sturt University).
[3] Bornmann, L., Haunschild, R., & Hug, S. E. (2018). Visualizing the context of citations referencing papers published by Eugene Garfield: A new type of keyword co-occurrence analysis. Scientometrics, 114(2), 427-437.
[4] Ali, I., & Gölgeci, I. (2019). Where is supply chain resilience research heading? A systematic and co-occurrence analysis. International Journal of Physical Distribution & Logistics Management.
[5] Deng, S., & Xia, S. (2020). Mapping the interdisciplinarity in information behavior research: a quantitative study using diversity measure and co-occurrence analysis. Scientometrics, 124(1), 489-513.
[6] McCullagh, P. (2002). What is a statistical model?. Annals of statistics, 30(5), 1225-1310.
[7] Yau, C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. Scientometrics, 100(3), 767-786.
[8] Zhang, Y., Tao, J., Wang, J., Ding, L., Ding, C., Li, Y., ... & Zhang, H. (2019). Trends in diatom research since 1991 based on topic modeling. Microorganisms, 7(8), 213.
[9] Han, X. (2020). Evolution of research topics in LIS between 1996 and 2019: an analysis based on latent Dirichlet allocation topic model. Scientometrics, 125(3), 2561-2595.