Chapter 8
Review of Chinese Text Mining in Agriculture

Xinyue Zhao
https://orcid.org/0009-0000-1528-7137
Shandong Agricultural University, Taian, China

Yunsheng Song
Shandong Agricultural University, Taian, China

ABSTRACT

Agricultural text mining refers to the utilization of natural language processing and deep learning techniques to extract and analyse useful information from a vast amount of existing agricultural texts. Nowadays, there is a tremendous quantity of text information related to Chinese agriculture; however, there is a scarcity of summarization and generalization in the field of agricultural text mining. This leads to duplicated and redundant agricultural textual information, making it difficult to obtain a holistic perspective on agricultural texts. Therefore, this chapter aims to categorize and summarize the latest developments in text mining in Chinese agricultural domain. Firstly, following the order of Chinese agricultural text mining, various methods for agricultural text preprocessing and representation are introduced. Then, research work related to Chinese agricultural text classification and matching is discussed in detail. Finally, the chapter concludes with a summary of the current state of research in agricultural intelligent services. This chapter contributes to a better understanding of the research trends in this field and provides suggestions and inspiration for future studies.

INTRODUCTION

With the rapid development of the Internet, the volume of textual information in the agricultural domain continues to increase. The substantial amount of agricultural text data faces challenges such as data dispersion, insufficient text mining, and complex content. Consequently, the efficient extraction and analysis of valuable information from large-scale agricultural datasets have become crucial issues in the agricultural domain.

DOI: 10.4018/979-8-3693-1582-8.ch008

As the world's population continues to grow, how to increase agricultural production on limited land area is crucial. In the current era of globalization and digitization, agricultural text mining is not only a means to enhance agricultural production efficiency but also a key tool providing profound insights for agricultural decision-makers, researchers, and practitioners. By mining textual data in the agricultural domain, we can gain a comprehensive understanding of various aspects such as crop growth, meteorological conditions, market trends, and more. This enables the optimization of agricultural production processes, improvement of crop quality, and even effective responses to global challenges like climate change and food safety.

Text mining can be understood as the process of utilizing computer technology and natural language processing (NLP) methods to discover, extract, analyse, and infer useful information from a large amount of textual data (Guo 2023). The development of text mining can be divided into three stages. The first stage is the rule-based stage, where early text mining primarily relied on manually designed rules used to extract specific information from the text. Rule-based methods include keyword-based searches, information extraction based on vocabulary and grammar rules, among others. Text mining in this stage was mainly applied to information retrieval, keyword search, and simple text classification tasks. The second stage is the statistical-based stage, where text mining gradually shifted towards statistical techniques with the rise of statistical methods, relying more on the statistical distribution of the data itself. This period emphasized the application of probability models and vector space models. Statistical-based text mining is primarily applied to tasks such as text classification, clustering, information retrieval, and various NLP problems. The third stage is the deep learning stage, where deep learning models, through the learning of multiple layers of neural networks, can automatically learn more complex and abstract features and representations from large-scale text data, achieving breakthroughs in the field of text mining. Deep learning-based text mining is mainly applied to complex NLP tasks such as sentiment analysis, machine translation, named entity recognition, and others. Currently, due to the unique characteristics of the agricultural domain, the performance of existing text mining techniques transferred to the agricultural domain is not satisfactory. This paper summarizes relevant research on Chinese agricultural text (CAT) mining up to now, to provide readers with more valuable and targeted information.

Text mining is the key technical support for agricultural intelligent services, and research on CAT mining directly affects the process of digital-driven agricultural development. Although there have been comprehensive literature reviews on CAT mining, the existing English review papers do not cover the mainstream deep learning methods and technologies (Drury, 2019). While the existing Chinese papers provide a summary of various aspects of agricultural text mining, they lack specific elaboration on these topics (Zhao, 2023). This paper will start from the characteristics of CATs to clarify the status of research on CAT mining technology and the development status of agricultural intelligent services. Based on NLP, agricultural intelligent services can be classified into two types: classification and prediction. The former is mainly used for the classification and information extraction of CATs, such as agricultural news classification, agricultural intelligent Question Answering System (Q&A), and agricultural public opinion monitoring. The latter is mainly used for agricultural product price prediction.

CORE TECHNOLOGUES FOR CHINESE CAT MINING

To extract potential information effectively and efficiently from CATs, it is necessary to have a deep understanding of the characteristics of agricultural domain texts. CAT mining should start from the