张军亮 著

面向“三农”问答系统的关键技术研究

社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS(CHINA)

张军亮 著

面向“三农”问答系统的关键技术研究

社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS(CHINA)

图书在版编目(CIP)数据

面向“三农”问答系统的关键技术研究 / 张军亮著
. -- 北京 : 社会科学文献出版社, 2017.12
ISBN 978-7-5201-1836-1

Ⅰ. ①面… Ⅱ. ①张… Ⅲ. ①三农问题-汉字信息处理系统-研究 Ⅳ. ①TP391.12

中国版本图书馆 CIP 数据核字（2017）第 291511 号

面向“三农”问答系统的关键技术研究

著　　者 / 张军亮

出 版 人 / 谢寿光
项目统筹 / 许葆华　周志静
责任编辑 / 周志静　许葆华

出　　版 / 社会科学文献出版社 · 人文分社(010) 59367215
地址：北京市北三环中路甲 29 号院华龙大厦　邮编：100029
网址：www. ssap. com. cn
发　　行 / 市场营销中心（010）59367081　59367018
印　　装 / 三河市尚艺印装有限公司

规　　格 / 开 本：787mm× 1092mm　1/16
印 张：12. 25　字 数：157 千字
版　　次 / 2017 年 12 月第 1 版　2017 年 12 月第 1 次印刷
书　　号 / ISBN 978-7-5201-1836-1
定　　价 / 89. 00 元

本书如有印装质量问题，请与读者服务中心（010-59367028）联系

序

以数字化、网络化、智能化为特征的信息化浪潮为“三农”信息化发展营造了强大势能。政府和研究机构针对农业生产、农民生活以及农村建设方面的事务提供了大量的信息资源，对促进农村社会经济发展、提高农民的生产能力和生活水平都产生了十分重要的帮助作用。问答系统（Question Answering System，QA）是信息检索系统的一种形式，它能用准确、简洁的自然语言回答用户提出的问题，是目前人工智能和自然语言处理领域中一个具有广泛发展前景的研究方向。针对我国“三农”领域信息资源服务中尚未全面、深入的引入问答系统的相关理论和方法的现状，本书较为系统地阐述了问答系统的技术原理和中文信息处理的相关知识，将 FAQ 系统和 Web 自动问答技术应用到当前的“三农”信息资源服务中，研究满足问答系统的“三农”知识表示方式；研究融合 HowNet 以及“三农”概念簇等计算 FAQ 问句匹配算法；研究综合利用自然语言处理、机器学习等方法实现“三农”问句分类和答案抽取的理论和方法；构建了面向“三农”FAQ 和 Web 自动问答系统模型。

本书是作者在参与国家社科基金重点项目过程中的研究成果，相关的方法研究和技术研究颇具新意。该书将问答系统的理念和技术应

用于“三农”信息服务中，特别是“三农”问答系统的构建，“三农”概念簇知识表示、FAQ 检索匹配，以及自动问答系统的“三农”问句分类和答案抽取等关键技术，拓展了信息服务的理论方法；对“三农”信息资源充分利用能产生积极的推动作用，从而能进一步促进现代信息技术在农村发展中的应用，有利于缩小我国城乡间的信息鸿沟。

本书的主要贡献是从“三农”信息需求出发，将问答系统融合到“三农”信息资源服务中，为我国“三农”信息资源服务提供一种新的服务模式的理论和实践，对其他领域开展类似研究也具有较好的参考价值和借鉴意义。希望本书的出版，有助于促进问答系统在“三农”信息服务广泛、深入应用，也希望有更多的领域、机构参与到“三农”问答系统的理论和实践研究中来。

朱学芳

南京大学信息管理学院教授，博士生导师

摘　要

随着“三农”信息资源需求的大量提升、信息资源数量的急速增长和农村信息基础设施的不断完善，如何提供有效的“三农”信息资源服务以满足信息需求，已成为一个亟待解决的问题，“三农”信息化建设成为我国信息化工作的重要组成部分。由于高效的问答系统能够从广泛的信息资源中，较准确地自动抽取提问问题的答案，因此，如果能有针对性地将问答系统技术应用到“三农”信息资源服务中，构建面向“三农”的问答系统，就能对解决“三农”信息资源利用问题产生积极的推动作用，能够为农民生活、农村生产、学者研究和管理者决策提供有效的“三农”问题信息服务。

在此背景和基础上，总的说来，本书以构建面向“三农”的问答系统为目标，首先，阐述了问答系统及其系统框架的基本相关概念和研究，以及由此展开的本书研究的内容、方法和意义等；其次，总结了本书研究的基础理论——中文信息处理基础理论；再次，分别研究了“三农”概念簇的知识表示、基于混合策略的“三农”FAQ 系统、面向“三农”问句分类以及面向“三农”的答案抽取等关键技术；最后，构建出面向“三农”问答系统。具体而言，本书的主要研究工作包括以下几个方面。

第一，基于K最近邻（K-Nearest Neighbor，KNN）分类算法的“三农”概念簇的研究。本书主要进行“三农”知识组织的研究，首先，用“三农”概念簇表示“三农”知识，利用基于DOM（Document Objecct Model）树从网络《农业大词典》抽取词条和释义部分的方法，通过正则表达式从释义部分抽取词条的口语名称和设计“三农”词表的结构；其次，从词条释义部分抽取、人工选择和合并特征词，生成特征向量，并利用KL（Karhunen-Loeve）变换对特征向量降维；最后，生成KNN的“三农”概念簇，并通过实验验证出，本书的特征向量的生成、降维和基于KNN的“三农”概念簇方法是有效的。

第二，基于混合策略的面向“三农”常见问题问答（Frequently Asked Questions，FAQ）系统的研究，以FAQ系统的检索匹配方法为主要研究对象。首先，通过问句之间的表层和语义相似度计算问句之间的相似度、利用LSA计算用户提问问句和常见问题集的答案部分间的相似度；其次，采取混合策略法将这两个相似度组合到一起，形成本书的“三农”FAQ系统的检索方法：基于混合策略匹配方法，并通过实验验证了这种方法的有效性。

第三，面向“三农”问句分类体系和分类方法研究。本书参考开放域问句的分类体系和“三农”领域知识，设计了面向“三农”自动问答系统的问句分类体系；把疑问词、“三农”概念簇、HowNet义原作为问句分类特征，将信息熵作为特征值，并设计了基于模板的粗分类和基于支持向量机（Support Vector Machine，SVM）的精细分类算法；并通过实验表明本书选取的特征向量和分类方法能够有效地满足需求。

第四，面向“三农”自动问答答案抽取方法研究。本书针对不同的“三农”问句类别和答案选择源，提出了不同的答案抽取解决方

式。对事实性问句，可采用基于“三农”知识库的抽取；对原因性问句，利用原因性线索词的模板指导抽取；对于方式性问句，则采用基于自动文摘的方式性的抽取。实验验证了本书的答案抽取方法的有效性。

第五，面向“三农”问答系统的构建与实现。介绍了面向“三农”问答系统构建的网络环境和服务器端技术，以及实现所应用的相关技术和结果。

第六，本书还对研究的主要工作进行了总结，指出了研究的不足之处，并提出了下一步研究工作的构想。

关键词：“三农”自动问答；“三农”概念簇；“三农”常见问题集；“三农”问句分类；答案抽取

Abstract

With the promotion of the information needs, the rapid growth of the information resources of "Agriculture, Farmers, Rural Area" (AFR), and the constant improvement of the AFR information infrastructure in rural areas, how to enhance information service to meet the information needs has become an urgent problem. The informatization of AFR is the important part of China's informatization. Question Answering (QA) system can more accurately and automatically extract the answer of the question, which was questioned in natural language, from a wide range of information resources. So, to build a QA system serving AFR will be able to promote the application of AFR information and has a positive significance for famers, researchers and policy makers by applying the QA into the AFR information service.

On the basis of the backgroup and the technology, the paperaims at building the QA system serving AFR. Firstly, the paper elaborates the basic concepts and framework of QA system and research topics both at home and abroad, the research contents and methods, significance and the basic structure of this paper. Secondly, the basic theories of Chinese

information processing are summarized, and it is also the basis of the study. Thirdly, the AFR concept clusters which represent the knowledge, FAQ system severing AFR based on the mixed strategy, the classification of AFR question, and answer extraction severing AFR are the key technologies of the QA system severing AFR. Finally, building a QA system severing AFR is described. The main research works of this paper are as follows:

First, the research on the AFR concept cluster based K-Nearest Neighbor (KNN). This part focuses on the AFR knowledge organization and presents the AFR concept cluster. First of all, the method that extracts the entry and interpretation section from the online "Agriculture Dictionary" and the other method that extract the spoken name using the regular expressions are elaborated. The AFR table is designed. Then, the feature words of entity are extracted, artificial selected and merged from the interpretation section. The feature vector and dimensionality reduction using KL transforms are executed. Finally, experiment shows the method is valid.

Second, FAQ system severing AFR based on the mixed strategy. This part is mainly about research on FAQ search matching method. The similarity of the surface and semantic similarity between the questions and the similarity between the user's question and the answer section of question answer pairs are calculated. Then take a mixed strategy to group the two similarities and form the retrieval of the FAQ severing AFR. Finally, the effectiveness of the method is verified by experiments.

Third, the AFR question classification system and method. This paper designs the questionclassification of automatic QA system severing AFR,

referring to the classification system of open domain and the AFR domain knowledge. We consider Wh-word, the AFR concept cluster and HowNet sememes as classification features, calculate characteristic value by the information entropy and design the algorithm of a template-based coarse classification and classification based SVM. The experiments show that the feature vector and classification method in this article can effectively meet the demand.

Fourth, the answer extraction of QA system severing AFR. According to different question category and answer source, the paper proposed different method. The method AFR knowledge-based is for factual questions. The method using template cues words of reason is for question of reason. For the “how” question, the method based automatic summarization extraction is proposed. These algorithms are also validated by experiment.

Fifth, the construction and realization of QA system severing AFR. The part describes the network environment, the server-side technologies, the related technologies applied in the system and the results of the system.

Sixth, we draw up the contribution of the research, and we indicate the shortcomings of the research and discuss the future work.

Keywords: Question Answering (QA) Serving “Agriculture, Farmers, Rural Area” (AFR); AFR concept cluster; Frequently-Asked Question (FAQ) of AFR; AFR question classification; answer extraction

目 录

图目录