文献综述 2500字

参考文献英文不低于3条，总共不低于15条

文献综述

1、引言

在数字化、网络化、全球化的科技迅猛发展的时代下，互联网上人与人之间的沟通交流方式越来越来丰富多彩，产生的数据总量及其复杂度都呈现出“爆炸式增长”的模式。面对海量、结构复杂的文本数据，通过以往的人工浏览的方式从大量的文本中获取有效的信息变得十分困难。近些年来，自然语言处理(Natural Language Processing)逐渐兴起，融合了计算机科学，统计学，代数学，语言学等多门学科，其目的是为了让计算机学会认知、理解、运用自然语言，从而帮助人们快速、准确的获取蕴含在文本中的信息。这里的“自然语言”是指自然世界中存在的各种人类语言，如中文、英文、德语、日语等自然界中人们沟通交流所使用的语言，并非指c、c++、java、goland等计算编程语言。由于其高效，快速，准确的特性，自然语言处理广泛应用于信息检索(Information Search)、数据挖掘(Data Mining)、机器阅读理解(Machine Reading Comprehension)、机器翻译(Machine Translation)、情感分析(Sentiment Analysis)、文本分类(Text Classification)、语音识别(Speech Recognition)等领域。随着计算机的算力不断提升以及存贮容量的增加，自然语言处理取得的效果和准确率也在不断的提升，能够成为科研人员或者课堂老师分析文本数据的有效利器。

由于长篇小说《平凡的世界》的篇幅过长，其中的人物关系复杂，故事情节跌宕起伏，对其感兴趣的读者在阅读的过程中没有足够的时间时，难以将整本书读完。即便读过的读者也需要往复阅读这本小说，才能理清其中的故事脉络和人物关系网络。而《平凡的世界》这本小说其中所蕴含的人情事故的处事原则和人生哲理也将会使读者受益。为帮助《平凡的世界》的读者能够掌握其中的知识，且鉴于自然语言处理技术的强大，本文利用自然语言处理中较为流行的技术帮助对《平凡的世界》长篇小说感兴趣的读者进行文本分析，帮助其掌握小说的主题思想，故事情节发展线以及小说人物之间的关系网络。

1. 国内外研究现状

文本分析属于自然语言处理下的一个重要领域，而自然语言处理属于深度学习和机器学习的范畴。本小节主要将自然言语的发展历程划分为5个阶段进行介绍。

萌芽期

在第二次世界大战后，图灵(Turing)于1936年首次提出“图灵机”得概念。图灵模型的诞生带动了自动化的萌芽，这被认为是当今计算机科学发展的根基。图灵在自动化研究上的工作引领了同时期的科学家们在该领域的不断思考和研究。其中数学家沃尔特皮茨(Pitts)与医学博士麦卡洛克(McCulloch)的共同工作使得神经网络得最初模型问世；以及美国数学家、逻辑学家家Kleene在有限自动机和正则表达式上的研究为后来的机器学习奠定了基础。1956年，香农(Shannon)将离散马尔可夫过程的概率模型应用在描述语言的自动机上。1956年，语言学家乔姆斯基博士在香农的工作上，提出使用有限状态机来识别句法以及定义有限状态机过程产生的语言为有限状态语言，包括上下文无关语法。这些工作形成了前期机器语言处理的理论基础。

该时期，基于概率算法的语音和语言处理也得到了发展，香农提出了在信道中传输的信息的编解码方式以及噪声的存在对信道的干扰。香农将热力学中的“熵”的概念应用到统计信道中信息容量的大小，开创了将概率方法应用到英文信息测度领域的先河。

1952年，贝尔实验室创建了第一个基于统计的语音辨识系统，该系统通过说话人的语音来辨识输出0到9之间的数字。在特定的条件下，该系统能够达到97%至99%的正确率

1. 研究展望
2. 参考文献