**电 子 科 技 大 学**

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

**学士学位论文**

**BACHELOR THESIS**



论文题目  **基于机器学习的问答推荐算法设计**

专 业 **计算机科学与技术**

学 号 **2014060108020**

作者姓名 **伍峰**

指导教师 **俸志刚**

摘 要

当今互联网发展如此迅速,日新月异的各种技术层出不穷,什么大数据,云计算,人工智能等正潜移默化的进入了人们的生活,这里面我们都能看到一个技术的身影,那就是机器学习,面对互联网信息爆炸的今天,信息量爆炸,用传统的信息处理方式已经不足以满足现在的数据量,也就有了机器学习的用武之地.本文以问题答案匹配为机器学习的对象,以爬虫爬取数据,索引构建,模型构建,机器学习等为主要内容.

**关键词：**机器学习,问题答案推荐,模型构建,文本处理,关键词提取,爬虫,搜索引擎,索引

ABSTRACT

with the development of the Internet so quickly, so many new technology are born just like an endless stream,like big data,cloud computing,artificial intelligence and so on.

……

**Keywords:** time-domain electromagnetic scattering, time-domain integral equation (TDIE), marching-on in-time (MOT) scheme, late·-time instability, plane wave time-domain (PWTD) algorithm

第一章 绪 论

1.1 研究工作背景和意义

当今互联网发展如此迅猛,互联网的数据量越来越大,于是便产生了大数据,而大数据是为了人工只能所服务的,而人工智能一个重要的理论基础和分支,就是机器学习.这几年来,人工智能领域重新崛起,特别是当谷歌的alphago出现在世人面前并且击败了九段围棋高手李世石之后,机器学习算是正式进入了大众的视野,各大高校也开始纷纷开始了有关课题的学习和研究工作.

那究竟什么是机器学习呢,我们先从学习的定义说起. 按照人工智能大师西蒙的说法,学习就是系统在不断重复的工作中对本身能力的增强或者改进,使得下一次的执行相同任务或类似的任务的时候,会比之前做的更好更出色或者效率更高.

顾名思义，机器学习是研究如何使用机器来模拟人类学习活动的一门学科。稍为严格的提法是：机器学习是一门研究机器获取新知识和新技能，并识别现有知识的学问。这里所说的“机器”，指的就是计算机；现在是电子计算机，以后还可能是中子计算机、光子计算机或神经计算机等等。

1.2 机器学习国内外研究历史和现状

机器学习属于人工智能领域,人工智能是从20世纪40年代和50年代开始,逐渐走上历史舞台.1956年正式被确立为一门学科.而之后人工智能领域的发展,多多少少可以看到机器学习的身影.机器学习是人工智能发展到一定阶段的必然产物.

从20世纪50年代到70年代,当时人工智能领域的学者们普遍认为,只要给机器赋予逻辑推理能力,那么机器就有了智能,当时具有代表性的成果主要有”逻辑推理家”程序以及之后的”通用问题求解”程序等.其中”逻辑推理家”程序在1952年就证明了<数学原理>的38条定理,又过了11年,就证明了全部52条定理,而且定理2.85甚至比罗素和怀特海证明的证明更为巧妙,这个程序也给作者带来了1975年的图灵奖.

然而随着研究的深入,学者们发现,光有逻辑推理能力,是完全不够的.又掀起了一股新的认知狂潮,认为要让机器智能,必须得先让机器有知识,于是20世纪70年代开始,学者们的研究方向转向了赋予机器人类的知识,大量的专家系统出现,但是人们也逐渐发现,由人来把知识总结出来再教授给计算机的这个过程,是相当困难的,于是有的学者就想:机器要是能自己学习就好了.这也就是机器学习的必然性.

1.3 LTR研究现状

1.4 本论文结构安排

第二章 机器学习基础

2.1 机器学习分类

2.2 机器学习基本原理

2.3 深度神经网络

2.4 LTR分类

第三章 爬虫基础

3.1 爬虫基本原理

3.2 爬虫常见问题

第四章 搜索引擎基础

4.1 搜索引擎基本原理

4.2 文本特征值处理

4.3 索引

4.4 LTR

第五章 问答推荐系统研究与设计

5.1 数据的爬取

5.1.1 爬虫框架选取

5.1.2 数据结构设计

5.1.3 爬虫代码设计

5.1.4 爬虫中间件设计

5.2 数据预处理

5.2.1 文本特征值提取

5.2.2 建立正向索引

5.2.3 建立逆向索引

5.2.4 相关度标记

5.3 机器学习模型

5.3.1 模型的选取

5.3.2 模型的训练

5.4 优化

5.4.1 爬虫优化

5.4.2 数据库优化

第六章 全文总结和展望

6.1 总结

6.2 后续工作展望

致谢

参考文献