学校编码：10384　　分类号_____密级_____
学号：24320121152273　　UDC_____

硕 士 学 位 论 文

基于模板匹配需求识别的方法研究与应用

Research and Application of Demand Recognition Method Based on Template Matching

陈康

指导教师姓名：王 备 战 教 授
专 业 名 称：计算机软件与理论
论文提交日期：2015 年 4 月
论文答辩日期：2015 年 5 月
学位授予日期：　　年　　月

指 导 教 师：_____

答辩委员会主席：_____

2015 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘要

传统的搜索引擎的搜索方式是基于倒排索引的全文检索，也就是根据搜索语句查询索引库中的检索方式，并没有很好地利用搜索语句所表达的含义，这样就不能准确识别出用户的具体需求，势必会给用户带来更大的搜索成本。垂直搜索的引入解决了传统搜索引擎的这一不足，而实现垂直搜索首先就是要识别用户搜索语句的含义，这也是自然语言处理所要解决的问题。

本文设计了基于模板匹配的需求识别算法，并在这个需求识别算法的基础上针对股票垂直类目词典挖掘的具体应用进行了设计与验证，提出了相关的数据结构和算法。为了设计需求识别算法和股票垂直类目词典挖掘方案，本文研究了相关词典查找技术，并介绍了本文中使用的机器学习分类技术和海量数据处理技术。

首先，本文研究与讨论了需求识别算法以及股票垂直类目词典挖掘的常用的相关技术，包括相关数据结构与算法、常用机器学习算法以及本文中使用的海量数据处理相关技术，包括 MapReduce 分布式编程模型。

其次，本文在前面介绍的相关技术的基础上，设计了基于模板匹配的需求识别算法，介绍了具体的设计思路，设计了相关数据结构和算法。在设计的基于模板匹配的需求识别算法的基础上，本文针对一个具体应用场景－股票类目 Query 需求识别，设计股票垂直类目相关词典挖掘方案，主要关注于特征的选择，并应用机器学习经典算法逻辑回归进行分类。

最后，本文基于前面的设计，针对具体的应用对实验环境、实验数据和实验过程进行了详细的介绍，并对本文挖掘出的股票类目模板词典、专名词典的召回率和准确率进行了评估。

实验结果表明，本文设计的需求识别算法可以很好地识别用户的搜索语句的具体需求，并且本文设计的股票需求识别挖掘方案具有很好的召回率和准确率。

关键词：需求识别；自然语言处理；海量数据处理

Abstract

The way of traditional search engine to search the full text retrieval is based on the inverted index. That is based on string matching retrieval methods, and not a good use of the search statement on behalf of the meaning of users' queries. This does not recognize the user's specific needs, and bounds to give users greater search costs. Introduction of vertical search to solve this shortcoming of traditional search engines. And to achieve vertical search is to identify the meaning of the first sentence of the user search, which is also the natural language processing problems to be solved.

In this dissertation, we design and implement algorithms for identifying the needs of the search queries. On the basis of this framework to identify the needs for a specific application - Vertical Category dictionary mining stocks, we design and implement this application. We design and implement related data structures and algorithms. In order to design and implement of the framework of identifying the needs and mine stocks vertical categories dictionary, we studied the dictionary to find the relevant technology, and introduced the machine learning classification techniques and massive data processing techniques used in this dissertation.

Firstly, this dissertation discusses research design and implementation of demand recognition framework, as well as mining stocks dictionary vertical categories related technologies including the commonly used data structures and algorithms, machine learning algorithms commonly used, as well as among the massive data processing technologies used in this dissertation, including distributed MapReduce programming model.

Secondly, basing on the related technologies described above, a framework for identifying needs based on template matching is proposed. The design ideas are briefly described, and the relevant data structures and algorithms are proposed and illustrated in detail. . In this dissemination, through a specific application - vertical categories stocks dictionary mining, focusing on the selection of features and

application of data mining and machine learning algorithms to classify related.

Lastly, basing on the previous design and implementation, a detailed experiment is carried out. Experimental environment, experimental data and experimental procedure are described in detail. Stock category template dictionaries and special names dictionaries are dig out. Finally, the recall and accuracy are evaluated.

Experimental results show that the proposed design framework to achieve recognition and demand can well identify the specific needs of the user's queries, and mining stocks demand recognition program designed in this dissertation has good recall and accuracy.

Key Words: Demand Recognition; Natural Language Processing; Massive Data Processing

目录

Contents

第一章　引言

1.1　研究背景

如今互联网已经渗透到各行各业，给人们生活的方方面面带了许多便利。比如电子商务允许网民足不出户就可以完成选购付款，等待购买的商品送货上门，省去了去超市的时间成本，而且可以购买到物美价廉的商品。再比如以前人们查找文档资料都是在浩如烟海的图书馆一本一本的翻查，直到找到自己想要的信息。自从有了搜索引擎，人们只需要在搜索框中输入一个查询语句，搜索引擎就可以帮助本文找到最满足人们搜索需求的信息。正是由于互联网给人们带来越来越多的便利，网名的数量也逐年增长。根据 CNNIC（中国互联网络信息中心）2015 年第 35 次中国互联网络发展状况统计报告[1]，2005 年到 2014 年中国网民规模和互联网普及率方展图如图 1.1 所示。

图 1.1 中国网民规模和互联网普及率发展图

正是由于互联网的蓬勃发展，大大小小的站点像雨后春笋一样不断的产生，

网页的数据量也是以惊人的速度在增长。为了帮助人们快捷地找到信息，搜索引擎技术应运而生。搜索引擎方便了用户从庞大的网页信息中找到自己需要的准确信息。截至 2014 年 12 月，从全球各大搜索引擎所占据的全球市场份额上看，Google 位居全球搜索引擎第一的位置，占领 66.4%的市场份额，全球最大的中文搜索引擎百度以 11.15%的市场份额位居第二，第三和第四分别是微软 Bing 和雅虎搜索引擎[2]。总的来说，经过了多年的研究，国内外搜索引擎技术已经比较成熟，性能和稳定性都能让人满意，并且给人们的日常生活带了很大的便利，在人们的日常信息获取中发挥巨大的作用。但是，搜索引擎的潜在价值开发远远不够。搜索引擎技术经过多年的发展，已经非常成熟，尤其是国内外各大搜索引擎公司在搜索引擎领域的研究引领潮流。

1.2 问题的提出

传统搜索引擎尽管满足了人们对查询信息的基本需求，但不能为用户提供更加丰富的搜索需求和用户体验。并且人们对于搜索信息的需求很大一定程度集中在特定的领域，并且在特定的查找需求时，使用搜索引擎的比例较高。根据 CNNIC 在 2014 年中国网民搜索行为研究报告[3]显示，目前用户对于搜索引擎使用场景偏娱乐化和休闲化，当用户有查找或下载游戏、电影、音乐等娱乐需求时，使用搜索引擎的比例高达 79.7%。另外，有 70%左右的用户在有购物需求时、在需要查找学习资料时、在寻找软件应用时、在查看新闻时，以及热点事件发生时会使用搜索引擎。正是由于用户使用搜索引擎的目的性往往是包括在特定领域的，传统搜索引擎那种基于倒排索引的全文检索就会存在如下三大问题：

1. 不能理解搜索语句

传统的搜索引擎的搜索方式是基于倒排索引的全文检索，也就是根据字符串匹配的检索方式，并没有很好的利用搜索语句所代表的含义，这样就不能识别出用户的具体需求，这样势必会给用户带来更大的搜索成本。

2. 结果页展现单一

传统的搜索引擎结果页都是千篇一律的 Title、URL 和摘要的格式，这个格式单一的样式对满足用户特定的需求具有局限性。

3. 与用户的交互性不够

传统的搜索引擎结果展示对于用户来说只是信息的被动接受者，很多时候用户的需求要想得到满足，往往需要进一步与搜索引擎进行交互，但传统的搜索引擎展示的结果限制了用户与搜索引擎的交互性。

正是由于传统搜索引擎的这些不足，垂直搜索引擎相关技术[4]被提出来了。为了实现垂直搜索，首先需要识别用户搜索 Query 的需求，因此本文设计了需求识别算法，并且针对一个具体的股票需求识别垂直类目挖掘进行设计与验证，并加以分析。

1.3　本文的主要工作

本文研究与讨论设计需求识别算法的相关技术，对设计的系统的相关数据结构，还有 MapReduce 技术、机器学习相关技术进行了介绍。本文设计了基于模板匹配的需求识别算法，提出了相关的数据结构和算法，并在这个需求识别算法的基础上针对一个具体的应用－股票垂直类目词典挖掘，进行了设计。

首先，本文研究与讨论了设计需求识别算法以及股票垂直类目词典挖掘所需的相关技术，包括相关的数据结构与算法、机器学习算法以及海量数据处理技术，包括 MapReduce 分布式编程模型。

其次，本文在前面介绍的相关技术的基础上，设计了基于模板匹配的需求识别算法，介绍了具体的设计思路，设计了相关数据结构和算法。本文还通过一个具体的应用－股票垂直类目词典挖掘，主要关注于特征的选取，并应用数据挖掘和机器学习相关的算法进行分类。

最后，本文基于前面的设计，对论文提出的需求识别算法对股票类目词典挖掘进行了详细的实验，对实验环境、实验数据和实验过程进行了详细的介绍，并对本文挖掘出的股票类目模板词典、专名词典的召回率和准确率进行了评估。

1.4　论文的结构安排

本文共分六章，各章内容如下：

第一章 引言。主要介绍了本文的研究背景、所研究的问题的提出，并简单叙述了本文的主要研究内容；

第二章 论文相关技术研究。介绍了设计需求识别算法时所用的词典查找技术，介绍了在股票专名挖掘是分类时使用到的相关机器学习算法，以及在处理数据时使用的海量数据处理技术。

第三章 基于模板匹配的需求识别算法设计。介绍了本文设计的需求识别算法，设计基于模板匹配的相关数据结构和算法。

第四章 基于模板匹配的股票类目需求识别词典挖掘。设计了股票垂直类目需求识别所需的模板与专名词典挖掘方案，详细设计了特征选取的方案，数据处理，并且使用逻辑回归算法对股票专名和非股票专名进行分类。

第五章 系统实验。基于前面的设计方案，针对股票类目进行了详细的实验，对实验环境、实验数据和实验过程进行了详细的介绍，并对本文设计并挖掘出的股票类目模板词典、专名词典的召回率和准确率进行了评估。

第六章 总结与展望。对本文的研究的内容与设计的需求识别方案进行总结，并对目前存在的问题提出接下来的研究方向。