学校编码：10384　　分类号_____密级_____

学号：15420131152019　　UDC_____

硕 士 学 位 论 文

文本数据挖掘中的中文分词算法改进及其在图书馆流通数据的应用

Improvement in Chinese Word Segment Algorithm of TDM and Its Application in the Library Circulation Data

陈星晶

指导教师姓名：张志强教授

专　业　名　称：统计学

论文提交日期：2016 年 2 月

论文答辩时间：2016 年 4 月

学位授予日期：2016 年 6 月

答辩委员会主席：_____

评阅人：_____

2016 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

高校图书馆是高校中重要的机构，是高校师生学习和科研的重要场所。高校图书馆数据库每天都会产生大量的图书流通数据，积累了海量流通数据。这些数据潜藏着很多尚未被挖掘的有价值的信息，所以不少学者将数据挖掘的技术应用于挖掘图书馆流通数据，以进一步完善图书馆管理工作。图书馆中的每一本图书都拥有其特定的图书名和索书号，其中索书号包含了中图分类号。图书名是以文本数据储存的，是用户了解并借阅该书的第一途径，但直接对文本数据进行处理有很大的难度。所以，以往的大多数研究将索书号或中图分类号作为图书的代表变量进行数据挖掘，但会导致分析结果过于简单，应用范围有所局限。

为了直接对图书名这类文本数据进行分析，本文尝试将文本数据挖掘技术引入到图书馆流通数据。由于图书作者起名习惯的差异和中文存在大量的同义词会导致相同主题的图书在图书名上存在一字之差或者细微区别，且并非图书名中每一个词都对研究有意义，所以直接使用图书名会存在较大的问题，我们有必要对图书名进行中文分词然后提取关键词。在中文文本数据挖掘技术中，中文自动分词被公认为瓶颈问题。而高校图书馆中的图书大部分是直接以学科名作为关键词进行命名的，而学科名基本是由词和词组成的，所以基本的分词算法无法直接满足这个需求。本文在传统的隐 Markov 模型的基础上，打破独立性假设对模型进行了改进，进一步利用已知的学科信息进行词元与词元之间的匹配连接，使得在书名方面的分词准确率有所提高。在实证部分，构建用户借阅网络，应用了关联规则等技术进行进一步的挖掘，还进一步讨论了不同学科的用户借阅行为的相关性和不同学科图书的流通率差异，可以明显看出使用文本数据挖掘技术可以为优化图书馆管理得到更多有用的信息。此外，运用文本数据挖掘技术能够为完善学生用户课程体系相关的教育评估系统和评估高校学科建设和学科交叉提供可能性，因此对高校图书馆流通数据进行文本数据挖掘有一定的理论意义。

关键词：文本数据挖掘；中文分词；隐 Markov 模型；关联规则

Abstract

University Library is an important institution, which is an important place for learning and research. University library databases produce large amounts of data every day, and have accumulated a massive flow of data. These data hide a lot of valuable information not yet been excavated, so many scholars applying data mining technology to mining library circulation data in order to further improve the work of library management. Every book in the library books has its specific name and call number, and call number contained in the CLC. Title text is data storage of book name, which is the way users understand and borrow the book first, but directly to process text data has a lot of difficulty. Therefore, most studies in the past used call number or CLC as a representative variables of the book in data mining,which lead to simple result and limited range of applications. This paper attempts to apply text data mining technology to the library circulation data. Based on the traditional hidden Markov model, we break the independence assumption on the model, using known words of the subject information connected with the term match between, finally improve word accuracy of title. In the empirical part, build user lending network, using techniques such as association rules for further excavations, also discussed further borrowing behavior relevance of users from different disciplines and circulation rate of books on different subjects. Text data mining techniques can obtain more useful information for the optimization of library management. In addition, text data mining technologies offer the possibility to improve education evaluation system related to student curriculum and evaluation of subject construction in colleges, so TDM for the the circulation data has some significance.

Keywords: Text Data Mining, Chinese Word Segmentation, Hidden Markov Model, Association Rules

目 录

Contents

第一章 引言

1.1 研究背景与意义

随 Internet 技术的广泛普及，互联网已发展为巨大的信息空间，成为用户一个极富价值的信息源。移动互联、社交网络和电子商务等领域的发展拓宽了互联网的界限和应用领域。互联网每天都会产生数以亿计的新网页，并以爆炸式的增长速度不断产生着海量新数据。

互联网可视为异常巨大的分布式数据库，与我们常见的存放规则数据关系数据库不同，互联网大部分存放的是海量的非结构化的文本数据。研究表明，80%左右的电子化信息是以无结构化文本的方式存在的，其中 Web 中 99%的可分析信息是以文本形式存在的；机构内 90%的信息以文本形式存在的，例如数字化图书馆。

海量的文本数据对人来说是很难在短时间内完全理解并从中提取出有用的知识，而且传统的数据挖掘和信息检索技术在处理海量的文本数据时也不尽人意。文本数据和传统非文本数据相比，是一种无规则、复杂且抽象的数据。当前的数据挖掘技术并不能完全正确地理解文本数据所包含的信息和知识（特别是语义方面的）。因此，如何从海量的文本数据中挖掘出人们感兴趣而且有潜在价值的信息，已成为当前需要迫切解决也颇具挑战性的一大难题。

文本数据挖掘技术因其在文本数据尤其是自然语言方面有更加优秀的表现，引起了大家的广泛关注。社交媒体时代应运而生的舆情分析，是通过分析用户的评论来客观反映用户态度。在电子商务领域，淘宝上的商家可以根据用户对某商品的评论预测用户的需求，提供个性化服务，从而实现精准营销。舆情分析和用户个性化推荐就是典型的将文本数据挖掘算法应用于实际场景中并为企业产生价值的范例。

在信息化时代还未到来之前，文本数据大部分是以纸质的形式被储存在各类图书馆中，不同人群都通过图书馆这个庞大的文本资料库获取各类文本知识。虽然现在我们正处于信息化时代，互联网有取代传统图书馆之势，但是图书馆对于

各大高等教育学府等知识密集型机构来说，图书馆的核心角色难以被替代。

高等教育是为我国培养高素质人才的最主要的途径，高等教育为国家培养大批复合素质的精英，所以高等教育是我国教育事业的重点。而高校图书馆是高校中重要的机构[1]，主要面向高校师生提供教学和科研必需的纸质资料和电子化资料，并采购各类数据库为高校师生提供各类文献的检索与借阅。高校图书馆数据库每天都会产生大量的图书流通数据，随着数据量的增加，潜在的数据价值未被充分地得到挖掘。高校图书馆传统管理方式无法满足这种需求，陷入了“数据丰富，知识贫乏”的境地。为了探索丰富数据中隐含的有用知识，不少学者与图书馆馆员通过应用传统的数据挖掘的技术提取出这些流通数据中隐含的有用规律，为图书馆优化管理工作和服务工作提供数据支持。

高校图书馆的图书在数据库中一般是以索书号和图书名这两个字段共同储存的。图书名是基本都是以文本数据储存的，是用户了解该书的主要内容的第一途径，但直接对文本数据进行处理有很大的难度。每一本图书都有独一无二的索书号，其中包含了中图分类号和其他信息，中图分类号反映了图书的学科信息，所以大多数学者会将这个变量作为图书的代表变量进行研究。利用中图分类号作为图书的代表字段的这种处理方式，相比直接分析图书名来说可以大大降低问题的难度，但是信息缺失会影响结论的准确性。之前针对图书馆流通数据的数据挖掘的相关研究往往仅利用中图分类号作为图书的代表变量，鲜有将图书名作为代表变量进行研究。

高校图书馆的书名大多数是中文文本数据。相比于利用索书号或中图分类号进行数据挖掘，直接利用图书名对图书馆数据进行文本数据挖掘的优势主要表现于：

（1）用户借阅图书时，是直接使用书名信息及其书名关键词在图书馆查询系统来进行借阅的，更能传达用户的直接需求。而索书号只是一个图书定位媒介，对于用户来说是间接的，而且是一次性的。

（2）中图分类号是事先对图书进行人为的分类，相对比较粗糙且标准使用时间较长无法很好适应快速变化的学科发展，所以这种先验设定往往会对数据挖掘的最终结果产生比较大的影响。而图书名更能体现用户对于具体学科知识的确

切需求，所以能够从数据中获取更加精准的信息和知识。

（3）中图分类号的分类更为宽泛，所反映的学科分类无法和现行的国家学科分类标准无法很好地进行匹配，而使用图书名进行文本数据挖掘能够进一步匹配学科分类信息做更为深入的研究。

（4）传统的数据挖掘鉴于问题难度的考虑一般都是根据中图分类号进行分析的，分析结果较单一，一般是常识性的结论。使用文本数据挖掘可以结合其他文本数据使流通数据能够得到更加充分的挖掘，所以能够获得更加精准更深层次的知识。

对高校图书馆进行文本挖掘的意义在于能够得到更加精准的挖掘结果，为图书馆图书采购计划和图书馆馆藏图书的调整提供数据支持。图书馆进行采购是并不是以索书号和中图分类号作为依据的，而是在最近新出版的图书库里面根据具体学科或者图书关键词进行综合考虑然后制定采购计划。图书馆的馆藏空间是有限的，图书馆馆员需要定期调整书架，下架流通率较低的书籍以腾出空间摆放新采购的书籍，根据学科发展情况进行合理调整才能更好服务于高校用户。所以，文本数据挖掘更能胜任以上这些工作。

由于高校图书馆在高校中地位的特殊性，所以图书馆流通数据中除了能够为图书馆提供完善图书馆管理和更人性化服务的有用信息外，还记录了全体高校师生的学习和科研中的大部分信息，并能以此窥探高校学科发展情况。进一步，结合不同学生的课程计划（其也为文本数据），对图书馆借阅数据进行文本数据挖掘，可以用于完善现有的教学评估体系。若以图书馆流通数据去剖析用户借阅关键词词频用于考量高校用户的知识储备或科研兴趣，可以进一步去比较不同用户的知识储备或科研兴趣，进而为高校用户之间寻求学术伙伴提供数据支持。因此，将文本数据挖掘运用于高校图书馆流通数据有非常广阔的应用范围。

1.2 文献综述

国外最早将数据挖掘引入到图书馆领域的论文出现于1997年[2]。从那以后，国外许多研究者越来越多地将目光聚焦于数据挖掘在图书馆领域的应用。Nicholson[3]利用数据挖掘的技术分析数字图书馆的 Web 访问学术研究著作的记

录数据来对相似性学术研究进行分类。Chan C.C.H. [4]等利用关联分析中的 Aprior 算法用于图书馆数据中试图寻找学科之间的关联。Chang C.C. [5]等针对高校大学图书目录难以满足日常使用的问题，运用了数据挖掘的技术解决图书归类问题。Nicholson S. [6]将数据挖掘的技术和文献计量学进行结合应用于改善数字图书馆服务。

2000 年，国内研究者将数据挖掘技术引入图书馆界，相对的课题较少，主要引入概念为主。经过十几年的时间，国内图书情报学家们逐渐针对数据挖掘在高校图书馆的应用展开了研究。赵卫军[7]重点讨论了数据挖掘在高校图书馆资源优化、智能化服务、个性化服务等方面的应用；王慧敏等[8]以西安工程大学图书馆自动化管理系统的馆藏数据、读者信息、借阅数据作为基本数据源，在图书分类、入库比例以及各学院借阅量排名方面进行对比细分，探讨数据挖掘技术的应用可以更加细致地了解读者分布情况及图书需求情况，实现科学化管理和资源优化配置。随着进一步的发展，数据挖掘的更多新技术也被应用于图书馆中。陈定权等根据现有书目推荐系统的不足，研究如何将数据挖掘中的关联规则运用于图书馆的推荐系统。林和平等[9]运用数据挖掘中的关联规则技术处理图书馆产生的大量图书流通数据，预测分析读者的个性化需求。燕飞等[10]利用北京大学图书馆的流通数据构建起从用户到图书的“图书借阅网络”，从中发现共同借阅关系。刘晓亮[11]针对大数据时代背景下的图书馆数据挖掘技术进行了简要探讨，指出了图书馆数据挖掘所面临的挑战。这些文献在处理数据时基本都是将图书的中图分类号作为图书的代表变量，但这样处理存在很大的信息损失，而且大部分文献基本都定位于图书馆数据挖掘优化图书馆建设的思想，然而未将数据价值拓展至高校的学科建设和教育评估等方面中去。

为了使得图书馆流通数据能够得到更加充分的挖掘，我们需处理图书名这类文本数据，所以需要借助文本数据挖掘技术。然而，在进行中文文本数据挖掘时，会遇到一个特殊性性的问题——中文分词。其关键原因在于包括中文在内的亚洲语言大多数不存在英文那样以空格作为词与词之间的天然分隔符。

中文分词的准确与否，往往决定后续的中文文本数据挖掘的准确性，是基础和关键环节。由于中文分词本身的特殊性且汉语是一个非常复杂的语言体系，所

以中文自动分词是极具挑战的一个课题。前期的中文分词主要是利用字典匹配的方法，但是这类方法有一个比较大的缺陷，即过分依赖词典，词典的系统性直接影响算法切割的准确性。鉴于此因素，更多学者将统计方法引入到中文分词问题中，而最初只是一些基础的统计指标的引入。Sproat R. and Shih C.L.[12]提出将信息论中的“互信息”用于分词领域，量化了任意两个汉字之间的结合力。Maosong S. [13]等沿着这个思路进一步提出了汉字间 t-测试差的概念作为互信息的有益补充。韩冬煦[14]等提出使用卡方统计量以及边界熵提升未登录词的处理能力。

随着研究的不断深入，各类统计模型也不断地引入中文分词中。刘群[15]等利用提出了一种基于层叠隐 Markov 模型的汉语词法分析方法，将分词与词性标注、切分排歧和未登录词识别统一到一个系统内。其中分词部分是在经典隐 Markov 模型基础上引入类的概念，提高了分词效果。陶非凡[16]针对信息过滤中屏蔽关键词的技术进行进一步的研究，其主要的分词算法的核心部分是基于经典二元隐 Markov 模型的分词算法。田思虑[17]等在经典二元隐 Markov 模型的基础上加入最大逆向匹配的思想，考虑了词长和词序对分词结果的影响。

中文分词真正发展也就十几年左右，而且统计方法在中文分词的应用时间也不长，所以还有很多值得研究的地方。

1.3 论文章节安排

本文主要介绍的是文本数据挖掘的中文分词技术的统计模型的改进以及将其运用在图书馆的流通数据中，论文各章节的内容安排如下：

第一章介绍了论文的研究背景和相关的文献综述。

第二章对文本数据挖掘做了简要概述。

第三章具体介绍了中文分词的定义和常用算法，并详细介绍了中文分词中常用的传统隐 Markov 模型，在此基础上进行了算法上的改进，最后对针对图书名这一类特殊的文本数据做了进一步的算法改进。

第四章基于高校的图书馆流通数据，利用改进的中文分词算法、关联规则算法和文本数据挖掘技术进行了实证分析。

最后是总结和展望。

第二章 文本数据挖掘概述

处在互联网时代，文本是最为重要的信息媒介，有其所特有的优点。文本与图片和声音文件相比，所占的存储空间更小，储存更为便捷且表达的信息更为广泛。随着社交网络和电子商务的日益壮大，不断拓宽了互联网的领域，加速了数据的产生，所以文本的数量越来庞大，也越来越多样化。

文本数据挖掘所要面对的文本数据的特殊之处在于文本数据比传统数据挖掘面对的数据更为复杂且结构性不明显，所以并不能直接使用传统的数据挖掘算法解决文本数据分析问题。虽然技术上存在较大的难度，但这并不妨碍文本数据挖掘技术在各个领域的深入应用。当我们打开浏览器使用搜索引擎（如百度等）输入关键词进行搜索，或者打开淘宝首页时弹出的各类商品推荐，无不是应用了文本数据挖掘的技术在为我们提供高效便捷的服务。

2.1 基本概念

定义 1. 文本数据

文本数据（Textual Data，TD），是海量自然语言文本的集合，可以被人部分理解，但无法被人充分利用。与储存在各类关系型数据库的数据相比较而言，它是非结构化的，具有自然语言固有的模糊性与歧义性，并且存在着大量噪声。

文本数据在计算机（各类编程语言）中是以字符串的形式进行储存的，无法对其进行直接的数值运算。这就使得对文本数据进行研究比其他数据进行研究的难度大得多。

定义 2. 文本信息

文本信息（Textual Information，TI），是通过一定手段从文本数据中提取出来的，计算机可识别的，具有一定结构的关系。它是面向计算机的，无歧义的、是显性关系的集合。

文本信息是通过对文本数据进行预处理后整理得到的结构化关系，其优劣直接影响后续文本知识的获取质量。

定义 3. 文本知识

文本知识（Texual Knowledge，TK），是针对文本信息进行一定处理获得的有意义的模式，对人来说是可理解的且有用的。

定义 4. 文本数据挖掘

文本数据挖掘(Text Data Mining，TDM)，又称为文本知识发现（Knowledge Discovery in Texts，KDT），是指从海量文本数据中抽取出事先未知且有价值的信息和知识的过程。

2.2 基本流程

文本数据挖掘和传统的数据挖掘存在较大差异，主要体现在：

（1）文本数据挖掘面对的无规则的文本数据，所以算法上并不能直接使用传统的数据挖掘的方法进行直接使用。

（2）文本数据挖掘处理的文本数据是必须是大规模的，数据量大的好处是文本固有的噪声对最终的结果会小很多，这样文本数据挖掘的结果才是可靠的。

文本数据挖掘的目的是处理非结构化文本数据，从中提取有意义的数字索引，并且可以借助传统的数据挖掘（统计和机器学习）算法进行分析。虽然面对的数据存在差异，但文本数据挖掘和传统数据挖掘的整体思想和流程基本上一致。文本数据挖掘过程一般包括文本数据的预处理（文本数据的选择、清洗、特征提取等）、挖掘算法的应用、后处理（可视化）等步骤。其大致过程可以用图2.1来表示。

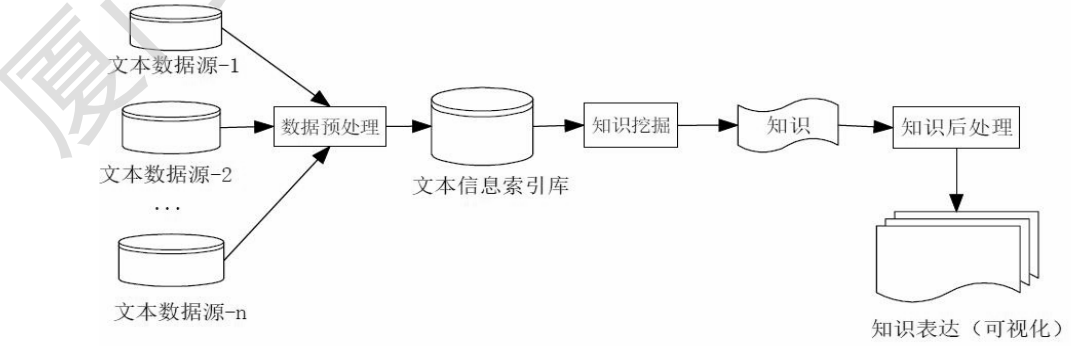

图 2.1 文本数据挖掘流程