河南大学

硕士学位论文

中文短文本分类的相关技术研究

姓名：崔争艳

申请学位级别：硕士

专业：应用数学

指导教师：胡小华

2011-05

摘 要

随着搜索引擎、电子邮件、微型博客和观点评论等短文本信息在互联网范围内的大量出现，有关短文本的相关研究逐步受到人们的关注。目前的文本分类技术多是针对长文本进行，虽然性能较好但由于短文本字数少、数量庞大，并多数依存于网络，并不一定适用。国内针对短文本的研究多集中在语义扩展、特征处理等方面，并没有特别深入系统的研究。

本文对短文本的涵盖范围、特点及研究领域进行了详细的分析研究，并对相关的研究现状和关键技术进行介绍。针对短文本特征稀疏等特点，考虑到传统的分词会因为词汇量稀少而丢掉重要的语义信息，本文采用“字”作为短文本的特征进行表示，并结合共现分析概念提出了基于字共现的特征提取方法。该方法在传统词频统计的基础上加入文本中字与字之间的共现信息量，使得特征字能够更全面地表达出短文本语义信息，通过实验证明该方法能够明显提高短文本的分类准确率。

有实验证明在诸多分类算法中，K近邻（KNN）和支持向量机（SVM）对短文本的分类效果最好。由于短文本数量庞大，本文采用KNN分类算法并加以改进。因KNN算法在分类前需要把所有训练文本存储起来与待测样本进行比对，计算量比较大，本文提出了一种改进的基于近似域KNN分类方法。该方法事先对训练集中各类别进行区域划分，确定类别中心域和近似域的范围，然后根据待测样本到各类中心向量之间的距离，判定样本在各个类别中的分布情况。只针对处于类别近似域内的样本再利用KNN算法进行分类，缩小KNN的搜索范围，进而提高分类的速度和准确率。同时，为减少样本误判率，对处于各类别边界区域的样本，在对其进行类别权重判断时设定边界参数，加大类别权重，这样使得边界模糊区域样本分类正确率有所提高。

关键词：短文本，分类，字共现，近似域

ABSTRACT

With search engines, e-mail, mini blog and view comments and other short text messages over the Internet within the scope of a large number of the emergence of research related to short texts gradually by the people's attention. The current text classification technology is for many a long text, although the performance is better but because of the short text as a small number of words, a huge number, and most dependent on the network., not necessarily applicable. Internal, short text studies were focused on semantic extension, feature processing, etc., and no special in-depth system.

This paper on the scope, characteristics and research of short text carried out a detailed analysis, and the current situation and related research and key technologies are introduced. For the short text feature sparse features and so on, taking into account the traditional segmentation of losing important semantic information because of few vocabulary, we use a "word" as a short text features that, combined with co-occurrence of the concept was proposed based on word co-occurrence feature extraction method. This method is based on the traditional word frequency statistics by adding the text between words in common is the amount of information, making the characteristics of the word to more fully express the semantics of short text messages, through the experiments show that the method can significantly improve the efficiency of the classification of short text.

Has proved in many classification algorithms, K-Nearest Neighbor(KNN) and Support Vector Machine(SVM) classification of the best short text. Because the large number of short text, we use KNN classification algorithm and improved. KNN algorithm needs to store up all the training text and test sample for comparison before the classification, has a large amount of computation, so positive and negative domain is proposed based on the KNN classification method. This method of training set in advance a regional breakdown of the categories to determine the type of center domain and the approximate domain of fields, and then according to the distance between the test sample to the center vector of each category, find out the distribution of the sample in each category. KNN algorithm is used only for classification of the sample which in the approximate area of a category to narrow the scope of KNN search , thus improving the speed and accuracy of classification. Meanwhile, to reduce the sample

error rate, for the sample on each category boundary region setting boundary parameters when category weight judged, increasing the category weight, which makes the sample classification accuracy of the samples of boundaries fuzzy region increased.

KEY WORDS：Short text, Classification, Word co-occurrence, Approximate domain

关于学位论文独创声明和学术诚信承诺

本人向河南大学提出硕士学位申请。本人郑重声明：所呈交的学位论文是本人在导师的指导下独立完成的，对所研究的课题有新的见解。据我所知，除文中特别加以说明、标注和致谢的地方外，论文中不包括其他人已经发表或撰写过的研究成果，也不包括其他人为获得任何教育、科研机构的学位或证书而使用过的材料。与我一同工作的同事对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

在此本人郑重承诺：所呈交的学位论文不存在舞弊作伪行为，文责自负。

学位申请人（学位论文作者）签名：_____

201 年 月 日

关于学位论文著作权使用授权书

本人经河南大学审核批准授予硕士学位。作为学位论文的作者，本人完全了解并同意河南大学有关保留、使用学位论文的要求，即河南大学有权向国家图书馆、科研信息机构、数据收集机构和本校图书馆等提供学位论文（纸质文本和电子文本）以供公众检索、查阅。本人授权河南大学出于宣扬、展览学校学术发展和进行学术交流等目的，可以采取影印、缩印、扫描和拷贝等复制手段保存、汇编学位论文（纸质文本和电子文本）。

（涉及保密内容的学位论文在解密后适用本授权书）

学位获得者（学位论文作者）签名：_____

201 年 月 日

学位论文指导教师签名：_____

201 年 月 日

1 绪论

1.1 研究背景及意义

截至2011年，互联网已经走过了40多年历程，相对人类上下五千年文明史来讲，40年是极为短暂的，但仅仅是这40多年，互联网为整个世界带来了翻天覆地的变化。正因为互联网的普及，人们的生活才变得更加便利，更加丰富多彩，人与人、与社会、与世界之间的距离也越拉越近。

据2011年1月19日中国互联网络信息中心（CNNIC）发布的《第27次中国互联网络发展状况统计报告》[1]数据显示，截至2010年12月，中国网民规模达4.57亿人，互联网普及率在稳步上升，稳居世界第一位，如图1-1所示。其中手机网民规模达到3.03亿人，占整体网民的66.2%。由此可以看出，移动网络、手机终端在中国互联网发展中起着更加重要的作用。

图 1-1 中国网民规模与普及率

搜索引擎、即时通信、网络视频、博客应用、论坛/BBS、电子邮件、网络购物等网络应用依然是广大网民首选，排在网民网络应用前十位。而社交网站和网络文学是2009年度新兴网络应用，微博客和团购是2010年新兴网络应用。微博客类应用反应了中国网络应用的新特点，使信息互通更加便利、快捷，也促进了各地区、各民族的沟通交流，推动了和谐社会的构建进程。

在网络应用中，主要的信息获取来源为：搜索引擎、网络新闻、网络视频；主要交流平台为：即时通信、博客/微博、论坛/BBS、电子邮件；主要电子商务为：网络购物

和团购等等。这些应用中，都产生了大量的文本信息数据，并且许多信息通常只是片断性描述说明或观点评论，具有很短的文字内容，我们称之为短文本，比如搜索页面片断、聊天信息、博客/微博信息、邮件主题、商品描述、论坛观点评论、图片/视频文字介绍等等。

由此可见，各种形式的短文本信息成为人们沟通交流及信息获取中不可缺少的方式，而且发挥着相当大的作用。最典型的代表就是微博[37]（即微型博客），它是一个基于用户关系的信息分享、传播以及获取平台，用户可以随时随地通过网络、手机甚至腾讯QQ签名等即时更新140字以内的文字信息。微博作为网民记录生活和发表评论的载体，其自身蕴含许多非常有价值的信息，话题涉及政治、经济、军事、娱乐等各个领域，在兴趣挖掘、热点话题跟踪与发现、流行语分析、舆情预警等领域都有着广泛的应用前景。微博虽然兴起较早，但真正进入主流网络应用是在2010年，以新浪为首的多家大型门户网站如搜狐、腾讯、百度等相继推出微博产品，微博的概念在2010年得到了空前的普及。截至2010年12月，微博的用户规模达6311万，网络使用率占13.8%。

面对互联网迅猛发展所产生的海量文本数据，如何准确有效的获取所需的资料和信息，更好的帮助我们进行工作和学习，文本分类技术在其中发挥着举足轻重的作用。特别是面对庞大的短文本数据，短文本分类的研究对于获取数据特征及进一步的数据挖掘工作，都具有重要的意义。

本文鉴于短文本存在的问题对其展开研究。由于短文本字数少、数量多，造成样本特征稀疏，特征维数较高，不能很好的抽取出文本关键特征。传统的采用关键词作为短文本的特征进行文本处理将会在一定程度上丢失文本语义信息，因而本文研究了用“ 字 ”作为短文本的特征进行处理，常用汉字数量不多会使得采用字为特征的文本表示方式能降低特征维数。但考虑到单个字表示能力弱的问题，文中结合共现分析概念充分挖掘字间相关性，使得特征字对于文本的表示能力增强。

短文本多数依附于网络，数量非常庞大，对其分类存在一定困难，有实验证明KNN方法和SVM方法在处理短文本分类问题上效果较好，KNN方法更适合大容量样本分类。但是使用KNN方法进行分类计算量较大，本文试图找寻方法来缩小样本分类的范围，因而在分类前对训练集文本类别进行了区域划分，并通过待测样本到类中心向量的距离判定样本在各类别中的分布，只对处于类别近似域的样本进行分类。

1.2 文本分类的研究现状

文本分类是文本挖掘的主要技术之一，它是指给定分类体系，将文本分类到某一个或者某几个类别中的过程。随着网络的迅速发展，电子信息文本不断扩充，文本分类作为一种有效的组织和管理方法，得到了广泛应用和快速发展。文本分类是信息处理的主要研究方向，有着极其重要的应用价值。由于其能快速、全面、准确地处理各种杂乱信息，在信息检索、信息过滤、搜索引擎、数字图书馆管理和文本管理等领域中有着广泛的应用。

文本分类的过程可以分为手工分类和自动分类。追溯到上世纪60年代，早期的文本分类研究是通过手工定义一些规则来文本进行分类，主要是基于知识工程。最著名的代表是yahoo的网页分类系统，它是由对某一领域有足够了解的专家定义分类系统，然后人工将网页分类。这种方法费时费力，难以保证一致性和准确性，现实中已经很少采用。

到上世纪90年代，随着网上在线文本的大量涌现和机器学习的兴起，大规模的文本（包括网页）分类和检索重新引起研究者的兴趣，产生了自动分类技术，分类速度和准确率都得到大幅度提升。文本自动分类 [2]大致可分为两种：知识工程方法和机器学习方法。知识工程方法指的是由专家为每个类别定义一些规则，这些规则代表了这个类别的特点，主动把符合规则的文档划分到相应的类别中。这方面最有名的体系是卡内基集团为路透社开发的Construe系统[3]。上个世纪90年代之后，机器学习方法成为主导。机器学习方法与知识工程方法相比，能够到达类似的准确度，而且减少了大批的人工参与，提高了效率。

现如今的文本分类技术正朝着智能性方向发展，特别是在搜索引擎方面，不再是单一的根据字符串匹配或模糊匹配进行搜索，而偏向于多元智能化，例如在搜索“ 中国现任国家主席是谁？ ”时，计算机会通过对网络搜索的网页进行自动研究分析，得出和“ 国家主席 ” 结合最密切的人物名字 “ 胡锦涛 ” 。

1.2.1 国外研究现状

国外文本分类的研究开始的比较早，始于1950年末。Luhn通过统计词汇在文摘中的标题或者文摘中的出现频率来选择文摘的标引类目，进行开创性研究[4]。1960年Maron提出了关键词自动分类技术，并在Journal of ASH上发表了有关文本自动分类的第一篇论文“ On relevance,probabilistic indexing and information retrieval"[5]。Good和Fairthome最早

认为自动分类有助于文献检索[6]。1971 年，Rocchio 提出了在用户查询中不断通过用户的反馈来修正类权重向量，来构成简单的线性分类器。1979 年，van Rijsbergen对信息检索领域的研究做了系统的总结。1992 年，Lewis 在论文中系统地介绍了文本分类系统实现方法的各个细节，是文本分类领域的经典之作。1995 年，Vipnik 基于统计理论提出了支持向量机(Support Vector Machine)方法，Thorsten Joachims第一次将线性核函数的支持向量机用于文本分类。

20世纪80年代末，在文本分类中占据主导地位的一直是基于知识工程的分类方法，即由专业人员手工编写分类的规则来指导分类。到20世纪90年代，随着网络的迅猛发展，网络上的网页、电子邮件、BBS以及博客等各种电子文本成几何级别数量增长，这些丰富的计算机可读的电子文本的出现使得基于机器学习的自动文本分类取代了基于知识工程的分类方法成为主流。

总的来说，国外对自动分类技术的研究大致分为三个阶段，分别是：自动分类的可行性研究、自动分类的实验研究和目前的自动分类实用化研究。在逐步的研究过程中，研究者们提出了多种分类模型和算法，如朴素贝叶斯(Naive Bayes)、K近邻(KNN)、支持向量机(SVM)、决策树、神经网络等等，并将这些技术引用到实际应用中，在信息检索、信息过滤、邮件分类等中有着广泛的应用。比较有代表性的是IBM的文本智能挖掘机，Autonomy公司的Concept Agents，还有一些自动分类系统，如麻省理工学院为白宫开发的邮件分类系统 、路透社的Construe系统，自动分类新闻稿件的文本分类系统，自动跟踪用户阅读兴趣的分类分析系统等。

自1995年后，随着世界范围内不断出现的数字图书馆研究热潮，国外计算机界和图书情报界陆续开展了对网络信息资源自动分类的研究，也有人将其称为自动分类技术的第四个发展阶段，如2000年美国ODLC的“ 蜗子计划 ”，用自动分类技术建立网络目录的研究[7-10]。

近几年来，新出现的文本分类方法主要有基于粗糙集理论的文本分类方法[11]、基于群的文本分类方法[12]、多分类器融合的方法[13]以及一些经典分类器的改进或者变形如KNNModel[14]、CB-SVM[15]等。

1.2.2 国内研究现状

国内对文本自动分类研究起步较晚，1981年侯汉清教授对计算机在文献分类工作中的应用做了探讨，并介绍了国外在计算机分类检索、计算机自动分类等方面的概况[16]。

随后，国内的研究单位和学者都开始了系统性的深入研究，最初国内的文本分类研究均是在英文文本分类研究的基础上进行，采用英文语料库对分类算法及技术进行相应的改进，后来研究人员逐步把分类技术引入到中文文本中，继而形成了中文文本自动分类技术研究体系。

1986年，上海交通大学研究所的朱兰娟、王永成等开发了中文科技文献(计算机类)实验性分类系统；1995年，清华大学的吴军开发了基于中文语料的文本自动分类系统；1998年，东北大学张月杰等提出通过计算预定义类别和文本特征项之间相关性来进行自动分类；1999年，邹涛等采用向量空间模型和基于统计的特征词提取技术，开发出中文技术文本分类系统。

此外，国内很多学者对中文文本分类算法也进行了深入的研究，黄萱菁等提出一种基于机器学习的、独立于语种的文本分类模型；周水庚等研究了隐含语义索引在中文文本处理中的应用；李荣陆等使用最大熵模型对中文文本分类进行了研究；张剑等提出一种在《知网》本体库基础上，建立文本的概念向量空间模型的特征提取方法；朱靖波等将领域知识引入文本分类，利用领域知识作为文本特征，提出一种基于知识的文本分类方法。目前，我国在中文文本自动分类领域已经取得了令人瞩目的研究成果，其中一些已被成功推广和应用。

相对于英文，对中文文本进行分类的一个关键性的因素就在于文本的预处理方面。英文文本中单词之间有空格来区分，而中文则需要进行对文本的分词处理。相当长的时间内，中文分类技术研究都没有公开的数据语料库，目前用的较多的是复旦大学建立的自然语言处理语料库，北京大学的人民日报语料库，清华大学现代汉语语料库，谭松波等人整理的文本分类语料库，以及搜狗实验室提供的网页新闻语料库。

1.3 短文本研究现状

短文本通常是由200字符以内的文字组成，内容少，文字短，具有稀疏性、实时性、不规范性和交互性等特点。短文本虽然内容少，但却包含大量有价值的隐含信息，因而随着短文本分类需求的日益增长，近年来人们对短文本的分类进行了一定的研究。

国外对短文本研究开始相对较早，主要集中在概念相似度计算方面[17-22]，有代表性的是Mehran Sahami等人提出的使用基于web语义核函数的方法和D.Metaler等人提出的基于相似性度量的方法；W.Yih等人通过扩展web语义核函数对上两种方法加以改进。

此外还有针对特征处理方法的研究，Xuan-Hieu等人提出了使用隐含主题建立一个通用框架，解决短文本稀疏性；J.Hynek提出一种基于Apriori的频繁词集分类方法来对数字图书馆中的文档摘要进行分类；D Song提出一种基于信息流的领域知识库进行短文分类。

国内对于短文本的研究起步较晚，目前中国科学院[23]、重庆邮电大学[24]和国防科技大学[25]等机构研究较多，主要集中在特征处理和分类算法上[26-29]。如龚才春通过构建短文本指纹网络实现了短文本语料的快速精确去重；胡佳妮等人提出使用潜在语义分析降低短文本的维数并去除噪声；樊迪提出一种利用标题与正文信息进行“ 联想 ”的方法，弥补短文本信息不足的缺点；覃张华在HNC理论框架下充分考虑短文本的领域、情境和背景，提取了对短文本主义影响较大的特征填充语境框架，有效处理短文本语义块的分类、句蜕和歧义等现象；闫瑞等人提出一种动态调整策略来训练组合分类器的短文本分类算法。

针对短文分类的研究大致可分为两类：基于规则的方法和基于语义的方法[30-35]。基于规则的方法是利用各类词汇相关联的规则进行分类，如吴薇提出了采用正则表达式作为规则生成工具，对大规模短文本进行过滤；王鹏利用依存关系抽取词对进行短文本的特征扩充；王细薇等提出了利用关联规则对短文本中的概念词语进行特征扩展；胡吉祥提出通过抽取消息文本中频繁模式(频繁出现的词或短语)来表征消息文本。而基于语义的方法主要应用通用知识库和领域知识库获取短文本中的语义信息，如宁亚辉等借助《知网》提出了基于领域词语本体的短文本分类方法 ;盛宇利用心理学中的‘ 熟悉原理 ”、“ 典型原理 ”等为模型建立特殊词库和典型案例词库，进行短文本的分类研究；王永恒提出了一种基于文本语义特征图对短文本进行分类。

除此之外，针对大量出现且又形式多样的网络短文本数据，许多应对于不同网络文本的研究也逐步展开[36-41]，如黄永文将产品评论中的产品特征、观点词作为语义内容，并将语义内容数量和评论文本长度等加入分类特征进行产品评论的挖掘；张卫等根据论坛帖子之间的回复关系构建了一棵回复关系树，有效改变原来帖子特征的稀疏性；何海江提出一种适合短文本的相关测试，用于衡量评论和文章语义相关程度；尹洪章等在分析聊天数据时序性的基础上，引入内容相似性信息，提出一种结合内容相似性和时序性的社会网络挖掘新方法；王乐利用短语消息间时间分布特征，设计了双时间窗口机制及其数据结构RMR，提出了短语消息流上的会话抽取算法；黄永光等人提出将不规范的短文本规范化，然后从规范的短文本中抽取特征串表征短文本。

单条短文本由于长度太短，很难挖掘出有效特征和有价值的信息，因此，对短文本的研究和处理一般都是针对整个短文本语料，从包含较大数量的短文本集合中挖掘出有用的信息，一般不是针对某条短文本进行处理。短文本语料独特的语言特征使得短文本处理方式及关键技术与常规文本有较大差异。

目前有关短文本的研究多数是以关键词作为文本特征进行表示，由于短文本内容少、词汇量小，关键词的提取容易遗失文本信息，同时采用关键词作为文本特征使得空间维数较大，如何更好的保有短文本为数不多的特征且降低维数？本文尝试用单个汉字作为短文本的特征进行表示和处理。基于单个字的表示能力差、关联性不强，文中结合共现分析概念来挖掘出字与字之间的关联，使得语义表达更充分。同时，针对用于短文本分类较好的KNN算法计算量大等特点，通过判定样本在训练集类别中的分布区域加以改进。

1.4 本文研究的主要内容

文本分类是提高信息检索及减少数据流量的一项关键技术，在信息过滤、信息检索及文本管理等领域有着广泛的应用，人们不断把文本分类的研究趋于实用化，特别是在新闻媒体、文献分类及网络文本分类和检索上，并逐步有针对性的开发出比较高效的分类系统。

现有的文本分类的研究与应用大多是针对长文本进行，长文本内容丰富，特征明显，表达充分。但是不管是在电子邮件、手机短信、文档/文献摘要，还是在网络文本等应用中，都出现了越来越多的短小文本，它们也发挥着相当重要的作用。特别是随着2010年微型博客在中国大陆的迅猛发展及日渐白热化，更加彰显短文本在信息领域及网络发展中起着不可忽视的作用。有专家预言，随着类似微型博客等短文本信息的便捷传播，可能会将互联网带入一个新的发展阶段。

短文本因为有其自身存在的一些特点，许多针对长文本的分类技术在其上并不能取得很好的效果。因此有不少专家学者对短文本展开研究，但随着短文本形式变化越来越多，并且缺少标准的语料库，目前的研究还处于不成熟阶段。

本文的研究内容如下：

1、本文对短文本的各种形式进行了详细介绍，对短文本的特点及其研究领域进行了概括，并简要介绍了文本分类的相关技术，包括文本预处理、特征提取和分类算法等。

2、短文本由于字数少、数量庞大使得样本特征稀疏，针对这一特点，本文采用以“字”作为短文本的特征，避免分词，更为简单有效。同时在此基础上结合共现分析概念，提出了基于字共现的特征提取方法，通过计算字与字之间的互信息量找出字的共现度量进行特征提取。该方法充分表达文本的语义信息，通过实验证明其有效性。文中通过实验还验证了不同的语料对短文本分类造成的影响，以及分字和分词对短文本造成的影响。

3、由于KNN分类算法计算量较大，文中通过事先对训练集中各类别进行区域划分，提出了一种改进的基于近似域KNN分类方法。以训练集中各类别的中心域和近似域为度量，判定待测样本的分布区域，只针对落入类别近似域的样本进行KNN分类，大大缩减了样本搜索比对的计算量，有效提高短文本分类的效率和准确率。同时，对处于各类别边界区域的样本，在对其进行类别权重判断时设定边界参数，加大类别权重，减少样本的误判率，使得准确率更高。

1.5 本文的组织结构

第一章　绪论。主要介绍本文的研究背景，对文本分类发展及国内外研究现状进行大体概括，并介绍了国内有关中文短文本的一些研究现状，同时概述本文的主要研究内容和组织结构。

第二章　短文本的特点及相关技术。对短文本涵盖范围、特点及有关研究领域进行详细分析介绍，并针对分类阶段各个过程的相关技术进行简要概述。

第三章　基于字共现的特征提取。根据短文本的特点，采用“字”作为短文本的特征项，并结合共现分析概念提出了基于字共现的特征提取方法，有效降低特征维数，通过实验证明其较好的性能。

第四章　改进的基于近似域KNN分类。通过事先对训练集中各类别进行区域划分，以此作为度量来判定样本在类别中的分布，提出一种改进的基于近似域KNN分类，并通过实验证明其高效性和准确性。

第五章　总结与展望。对本文的研究做出总结，对需要进一步研究和完善的工作进行展望。

2 短文本的特点及相关技术

互联网的发展使人们进入信息爆炸的时代，它为人们的工作和学习带来便利的同时，也不停为信息的获取提高难度，各种不同形式的信息充斥着人们的生活，不少人在获取信息的过程中要耗费大量的时间和精力。特别是近几年新兴的网络应用，如网络视频、网络购物、微型博客等，这些应用中不论是视频描述，商品描述还是微博信息，都产生了大量的短文本数据。

据2011年中国互联网络信息中心（CNNIC）发布的《第27次中国互联网络发展状况统计报告》数据显示，截至2010年12月，手机网民迅速扩大，规模达到3.03亿人，占整体网民的66.2%，移动网络、手机终端在中国互联网发展中起着更加重要的作用。移动网络和手机终端由于自身特点的限制，文本显示以短文本信息居多。

短文本通常是指长度比较短，一般不超过200字符的文本形式。普遍存在于网络文本、手机终端及文档文献中。如搜索页面片断、聊天信息、邮件主题、观点评论、商品描述、手机短信、文档/文献摘要等。短文本字数少，数量庞大，内容丰富，形式多样，通常具有稀疏性、实时性、不规范性和交互性等特点。

2.1 短文本涵盖范围

短文本字数少、内容丰富、形式多样且数量庞大，普遍存在于网络文本、手机终端及文档文献中，不但成为人们日常交流中不可缺少的信息形式，而且在文本发展和应用中发挥着相当大的作用。短文本信息中包含人们对社会各种现象的观点和立场，话题涉及政治、经济、体育、健康等众多领域，信息具有很大的研究价值。目前并没有权威的机构对短文本进行科学的定义，以下就短文本所涵盖的范围做出简要的介绍。

1、搜索页面片断。搜索页面片断是使用搜索引擎时返回的部分查询信息，这些信息包括题目和摘要两部分。题目通常是与查询内容相关联网页的题目，摘要部分是与查询内容紧密相关的描述信息片断，通常只有简短的几句话和一个链接，这部分信息为查询用户提供参考提示作用。如图2-1所示为百度返回的搜索页面片断。

图 2-1 百度返回的搜索页面片断

2、锚文本。锚文本是页面上用于链接的说明性文字，通常为一句话或几个字，用于点击进入下一个有关该话题的详细内容页面链接。对锚文本进行正确的分类具有重要意义。如图2-2所示为新浪新闻页面的部分锚文本。

图 2-2 新浪新闻页面的锚文本

3、聊天信息。随着互联网的出现和普及，各种即时通信软件如腾讯QQ、MSN、飞信等成为人们沟通交流的重要工具。聊天过程中产生了大量的短文本信息，对这些信息进行分类对进一步研究信息过滤、热点话题等问题具有重要意义，目前也有不少针对聊天记录中独特的奇异短语[42]的研究。

4、邮件主题。网络的普及使得电子邮件以方便快捷的方式逐步取代邮递信件，成为人们通信的主要工具，邮件主题通常为简短的说明性标题，对邮件的分类及过滤起着重要作用。如图2-3显示为邮件主题内容。

图 2-3 邮件主题

5、观点评论。观点评论通常为网络上新闻、博客、论坛等中有关主题的评论信息，大多数比较简短，但数量非常庞大，有时一个观点的评论信息能达到数千条。某一时段评论较多的主题，很有可能是该时段的热点话题。

6、商品描述信息。网络购物的兴起，使得商品信息铺天盖地，这些信息通常是针对该商品的一些特征性描述，内容较短，词汇量大。如图2-4为淘宝网上截取的个别商品简要信息。

图 2-4 淘宝网商品描述信息

7、图片/视频描述。图片和视频也大量充斥着网络，有关图片或视频的描述性文字也是比较简短的文本信息。如图2-5为百度视频截取的部分描述信息。

图 2-5 百度视频描述信息

8、微型博客。微型博客[43]（MicroBlog）是一个基于用户关系的信息分享、传播以及获取平台，用户可以随时随地通过网络、手机甚至腾讯QQ签名等即时更新140字以内的文字信息。微博信息是典型的短文本文件，对于热点话题跟踪与发现、流行语分析、舆情预警等研究起着重要作用。图2-6为新浪微博的部分信息。

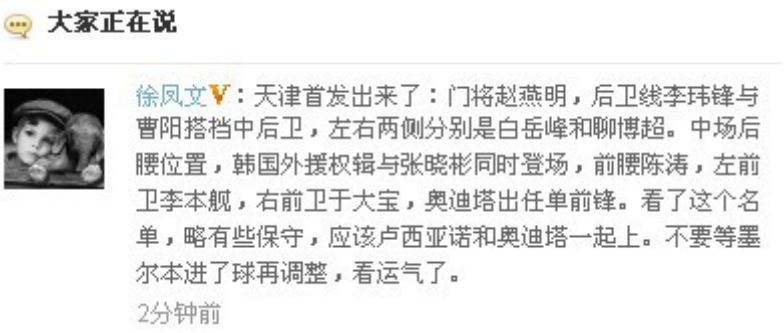

图 2-6 新浪微博信息

9、手机短信。手机短信通常只有简单的几个字，或几句话，也为标准的短文本文件，对研究信息安全、信息过滤[44]有重要意义。

10、手机网络。由于移动网络和手机终端自身限制，显示文本多以短文本文件居多。而且手机网络的迅速发展，使得对短文本文件的研究也越来越重要。

11、文档文献摘要。文档文献摘要为一个长文本的简要性概述，通常为二百字以内的短文本文件。由于海量文档文献的存在，使得对文档文献摘要的快速、准确分类研究要求也越来越高。

2.2 短文本特点

短文本由于内容少，数量庞大，使得其存在自身的一些特点，具体如下：

1、稀疏性。短文本长度相对较短，内容少，通常只用几句话甚至几个字进行描述。相对于长文本来说，它的关键词少，样本特征稀疏，很难抽取有效样本特征，更难以挖掘特征相互之间的关联性。

2、实时性。由于短文本内容少，保存价值不高，因此具有一定的实效性。特别是互联网上出现的短文本，都是动态出现的，刷新很快，像聊天信息、微博信息、评论信息等，都在以秒计时的速度不断的更新，难以采集，并且这部分动态文本数量非常庞大，这对短文本的计算速度和效率都有很高的要求。

3、不规范性。用语不规范和流行语多，是网络文本的最大特点，也为这类文本挖掘带来了诸多难点。特别是短文本，由于字数较少，甚至一些会有字数限制，因此用语表达更为简洁简练，如“88”代表“再见”，“94”代表“就是”，还会有一些奇异词汇，如“童鞋”代表“同学”，“长草”代表“很想要某个东西”等。随着网络的发展，

每年的流行语都不断出现，这些都具有一定的时代特色，对流行语的分析也可以帮助准确获取社会热点，如“三个代表”、“和谐社会”、“范跑跑”、“神马都是浮云”等；再加上一些错字和别字，这些都为文本特征提取带来了不小的难度。

4、交互性。即时通信和手机短信最先为人们在信息沟通上带来了便捷，同时也产生了短文本信息明显的交互性特征。用户的一次会话，通常由多条信息组成，这些信息之间彼此紧密关联，甚至每一条都无法独立描述会话的完整性。论坛和微博信息中也普遍存在用户信息交互性，因而想要了解一个时间段个体特征，必须关联本时间段里的每条信息。

2.3 短文本研究领域

1、相似度计算[45-49]

由于短文本字数少，样本特征稀疏，能抽取的信息有限，传统的文本相似度计算方法如余弦夹角方式等对其不能很好的适用，因此，研究者们开始转向利用外部信息特征来扩充短文本的语义信息，以弥补其特征不足的缺点，通常有以下几种方式：

（1）利用搜索引擎进行扩充。短文本在通过搜索引擎查询时，利用返回结果信息的排位、同现数目等不同特征来进行相似度的计算，以取得更好的语义表达。

（2）利用《知网》知识库进行关键词扩充。《知网》是以汉语和英语的词语所代表的概念为描述对象，以揭示概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库，在其中可以直接知道某些概念间的语义关联，包括同义、反义、下义、部分-整体关系等。针对短文本特征稀疏的缺点，将提取出的关键词通过《知网》中的概念进行语义扩展，使得相同主题、包含不同的同义词和近义词的文档通过相似度计算能更好地分为同类。

（3）利用维基百科或百度百科知识库进行扩充。维基百科和百度百科是当前存在的大型开放式网络百科知识库，每一个词条都有对应的语义关联和相关链接。将短文本关键词通过百科知识库进行扩展，进一步挖掘其所包含的更多的相关语义信息，使得相似度计算更加精确。

（4）通过源文档文件进行扩充。诸如文档摘要类短文本，可利用其相关源文件进行扩充，使得相似度计算更准确。

2、信息过滤

信息过滤是根据用户的需求对动态信息流或静态信息进行过滤，屏蔽对用户来说无意义的信息，在短文本的研究中起着重要作用，特别是广泛应用于手机短信、电子邮件以及网络聊天、论坛等。现有的文本过滤模型主要有：布尔模型、向量空间模型、潜在语义分析模型和神经网络模型。国内基于语义的文本过滤相关研究主要有三大主流过滤模型：东北大学姚天顺等构造了基于语义框架的中文文本过滤模型；复旦大学黄萱菁等构建了基于向量空间模型的文本过滤系统；中科院研究所晋耀红等人提出的基于语境框架的文本过滤方法[50-52]。

3、热点发现

热点话题检测与跟踪技术（TDT）是目前比较活跃的研究方向，它是以新闻、评论、视频等信息流为处理对象，实时监控发现热点话题报道，并将某时间段讨论比较多的话题以某种形式组织起来，全面的呈现给用户。例如中国知网根据学术论文关键词推出学术趋势，呈现某一时间段内学术研究热点及某个学术领域在不同时期研究趋势。针对热点发现的研究通常从五方面展开：新闻报道的切分、新事件的识别、事件关系识别、话题识别、话题跟踪[53-54]。金珠等提出一种基于信息检索和《知网》的话题跟踪及话题立场分类的方法；郑伟等提出对话题漂移现象的处理方法。

4、流行语分析

网络用语的不规范，使得流行语及奇异短语日异增多，如新流行语：“ 神马 ”、“ 给力 ”、“ 我爸是李刚 ”等；奇异短语：“ 偶稀饭 ”、“ 有木牛银在 ”等。流行语的分析对于了解时代特征、热点话题识别、舆情分析等都具有重要意义；奇异短语识别对于文本分类、聚类及相关研究也起着重要作用。黄永光等提出一种变异短文本聚类算法；清华大学夏云庆等提出基于中文网络非正规语言的处理方法和实践。

5、兴趣相关分析

兴趣相关分析研究广泛应用于各大综合网站、网络新闻、网络购物、论坛、博客及微博中。如Google推出的Google趋势，百度推出的百度指数，均是根据用户搜索主题词来寻找用户感兴趣和关注的热点趋势。博客和微博中通常会根据用户的日志内容，推荐一些相关文章，会根据用户资料推荐一些相关的朋友或圈子；腾讯QQ还会根据用户间具有相同的好友推荐他们是可能认识的朋友；淘宝网会根据用户关注过的商品推荐相关类似商品。兴趣相关研究使得网络应用更便捷、更人性化。Lau和Horvitz建立了但根据

查询和浏览信息来预测用户下一个查询的贝叶斯网络；Schechter等人构造用户访问路径树，采用最长匹配方法，寻找与当前用户访问路径匹配的历史路径，以此预测用户接下来的访问请求；唐灿等提出了一种面向新兴趣点发现的协作算法，建立了包括新兴趣点的多商品模糊兴趣模型；李村合等出了基于Web挖掘与相关反馈的多层次用户兴趣挖掘算法。

2.4 文本分类相关技术

文本分类的任务是：在给定的分类体系下，将文本自动地分类到某一个或者某几个相关联类别中的过程。换个角度来说，文本分类就是一个映射的过程，将未知类别的文本映射到已有的类别中。该映射可以是一一映射，也可以是一对多的映射，用数学公式表示如下：

$$f : A \rightarrow B \tag{2-1}$$

其中，A 为未知类别的待分类的文本集合，B 为分类体系中已有的类别集合，f是由A到B的映射。文本分类的映射是系统根据已经被标注类别的训练文档集合，找出文本特征和文本类别之间的关系模型，然后利用得到的关系模型对新的文档进行类别判断。

文本分类的过程通常分为文本预处理、特征表示、特征提取、分类算法和效果评估五部分。

2.4.1 文本预处理

由于文本分类处理的是大量半结构化或非结构化的，用自然语言描述的，计算机难以理解的文本数据，首先需要对这些数据进行相应的预处理。

文本的预处理是对文本进行文本去噪（去除标签、禁用词、停用词，还原词根等）、分词、词性标注、词义消歧等处理。由于中文文本是由连续的字符串组成，词和词之间没有像英文空格之类的明显的分隔符，因此需要对字符串进行分词处理，分词是文本预处理中一个重要的步骤。

2.4.2 特征表示

文本的特征表示就是将预处理后的文本数据转化为计算机可以识别的格式，如数值向量或符号向量。常用的文本特征有字、词和短语，另外还有概念和N元组等[55]。常见

的文本特征表示主要有四种模型：布尔模型、向量空间模型、潜在语义索引模型和概率模型。

(1) 布尔模型

布尔模型是最简单的模型，它定义了一个二值变量集合对文本进行标识，这些变量对应于文本中的特征项，若对应特征项在文本中出现，就赋值就为1，反之为0。布尔模型实现简单且速度快，但是对文本的表示能力差，无法区分特征项对文本的重要程度。

(2) 向量空间模型

向量空间模型在第三章中有详细介绍。

(3) 潜在语义索引模型

潜在语义索引模型(Latent Semantic Indexing Model)是向量空间模型的一种改进，它利用文本中特征项之间存在的潜在语义结构来表示特征项与文本之间的某种内在关系。潜在语义索引(LSI)方法，是由Scott Dectwester在1990年提出的，最早应用于信息检索。它能够加强相关文本之间的关联性，将高维的向量空间转化为低维的潜在语义空间，降低维数，简化算法复杂性。

(4) 概率模型

概率模型（Probabilistic Model）是考虑特征项之间及特征项与文本之间的相互关联性，如出现的频率、文本的长度等，更为准确的描述特征项与文本之间的相关性。把文本分为相关的和无关的，为特征项赋予一定的值用以表达其在相关和无关文本中出现的概率，系统通过计算这些概率来做出决定。概率模型需要事先确定相关概率，参数估计难度较大，因而并未广泛应用。

2.4.3 特征提取

特征提取是从特征总集中挑选出一部分对分类类别有贡献的特征项组成特征子集，根据某个特征评估函数对每个特征项进行评估，分别计算出它们的评分值，然后按评分值大小对各个特征项进行排序，提取出最高分的一些特征项作为文本的特征项子集。特征提取在不影响文本主题信息的情况下，尽量减少要处理的特征数据，由此可以降低向量空间维数，使计算变得简单，提高文本处理效率。常用的特征提取方法有：文档频率、信息增益、互信息、期望交叉熵等。

(1) 文档频率（Document Frequency，DF）

文档频率是计算量最小的一种特征提取方法，它是指在整个训练集中，出现特征项

的文本数。可表示为：

$$DF = \frac{\text{出现特征项的文本数}}{\text{训练集总文本数}} \tag{2-2}$$

文档频率方法的优点是计算量小，效率高，比较适合大规模文本集，是常用的降维方法，这种方法会忽略低频词表达的语义信息。

(2) 信息增益（Information Gain，IG）

信息增益通过词条t在类别c中出现或者不出现的次数来预测文档的类别，它根据训练数据计算出各个特征项的IG值，删除值最小的项，其余的按大小进行排序。特征项的IG值越大，说明贡献越大，被选取的可能性越大。计算公式为：

$$IG(w) = -\sum_{i=1}^{M} P(C_i)P(C_i) + P(w)\sum_{i=1}^{M} P(C_i \mid w)\log P(C_i \mid w) + P(\overline{w})\sum_{i=1}^{M} P(C_i \mid \overline{w})\log P(C_i \mid \overline{w}) \tag{2-3}$$

其中，M为类别数，w为特征项，$P(C_i)$ 表示 C_i 类文本在文本集中出现的概率，$P(w)$ 为文本集中包含w的文本的概率，$P(\overline{w})$ 为文本集中不包含w的文本的概率，$P(C_i \mid w)$ 为文本包含w时属于 C_i 类的条件概率，$P(C_i \mid \overline{w})$ 为文本不包含w时属于 C_i 类的条件概率。

信息增益反映了词条的类别区分能力，选择信息增益高于一定阈值的词条作为特征，可以有效地降低特征空间的维数。

(3) 互信息（Mutual Information，MI）

互信息度量特征项与类别之间的共现关系。特征项对于类别的互信息越大，它们之间的共现频率也越大。对于特征项w和某个类别 C_i，互信息可表达为：

$$MI(w, C_i) = \log\frac{P(x, y)}{P(x)P(y)} \approx \log\frac{F \times N}{(F + F_c) \times (F + F_w)} \tag{2-4}$$

其中，N为训练样本集中的文本总数，F代表w和 C_i 同时出现的次数，F_c 代表为 C_i 出现w不出现的次数，F_w 代表w出现 C_i 不出现的次数。当特征项属于某一类别时，互信息就大；不属于某一类别，两者相互独立，互信息为0；很少在某一类别中出现，互信息为负数。该方法充分考虑了文本中的低频词对文本重要性，但有些过于考虑低频词，有时会导致分类效果较差。

(4) 期望交叉熵（Expected Cross Entropy）

期望交叉熵反映了文本类别的概率分布和在出现了某个特定词汇条件下文本类别的概率分布之间的距离，值越大，对文本类别分布影响越大。公式为：

$$CE(w) = P(w)\sum P(C_i \mid w)\log\frac{P(C_i \mid w)}{P(C_i)} \quad (2\text{-}5)$$

其中，$P(C_i \mid w)$ 为包含w的文本在属于 C_i 类的概率。

以上特征提取方法均是基于评估函数的，此外还有基于语义的特征提取方法。如：基于语境框架的文本特征提取方法，能有效地处理语言中的褒贬倾向、同义、多义等现象[52]；基于本体论的文本提取方法，应用本体论 (Ontology) 模型可以有效地解决特定领域知识的描述问题，充分考虑特征词的位置以及相互之间关系；基于《知网》的概念特征提取方法，利用《知网》将语义相同的词汇映射到同一概念，合并为同义词，有效降低文档向量的维数，减少计算量，提高效率。

2.4.4 分类算法

文本分类是指给定分类体系，将文本分类到某一个或者某几个类别中的过程。具体分类过程描述是，用一个已知类别的文本训练集来训练分类器，再用训练好的分类器对未知类别的文本进行分类。分类算法在文本分类中是非常重要的一个环节，不同的数据集选用不同分类算法，会显示出不同的分类效率。常用的分类算法有决策树（DT）、人工神经网络（ANN）、支持向量机（SVM）、朴素贝叶斯（NB）、遗传算法（GA）和K近邻（KNN）等。有实验证明在诸多分类算法中，KNN和SVM对短文本的分类效果最好，本文采用了对大样本处理较好的KNN分类算法对短文本进行分类，第四章中有详细介绍。

2.4.5 效果评估

采取一定的评价指标和准则来评价文本分类器性能的好坏，在文本分类中起着非常重要的作用。国际上广泛采用微平均和宏平均相结合的评价准则，并采用查准率和查全率两个指标来衡量分类系统的性能。

在文本分类中，一般用准确率P（Precision）和召回率R（Recall）以及F1值来衡量分类系统的性能。对于第i个类别，其准确率和召回率分别定义如下：

$$P_i = \frac{l_i}{m_i}\times 100\%，\ \mathrm{R_i} = \frac{l_i}{n_i}\times 100\%，\ F1_i = \frac{P_i \times \mathrm{R_i} \times 2}{P_i + \mathrm{R_i}} \quad (2\text{-}6)$$

这里，l_i 表示分类的结果中被标记为第i类别且标记正确的文本个数，m_i 表示结果中表示被标记成第i个类的文本个数，n_i 表示被分类的文本中实际属于第i个类别的样本个

数。

微平均和宏平均是计算全局的查准率、查全率和F1测试值的两种方法。其中，微平均用mP、mR、mF1来表示；宏平均用MP、MR、MF1来表示，公式如下：

$$mP=\frac{\sum_{i=1}^{m} l_i}{\sum_{i=1}^{m} m_i} \qquad mR=\frac{\sum_{i=1}^{m} l_i}{\sum_{i=1}^{m} n_i} \qquad mF1=\frac{mP\times mR\times 2}{mP+mR} \tag{2-7}$$

$$MP=\sum_{j=1}^{n}\frac{N_i}{N}\cdot mP_j \qquad MR=\sum_{i=1}^{n}\frac{N_j}{N}\cdot R_i \qquad MF1=\sum_{j=1}^{n}\frac{N_j}{N}\cdot mF1_j \tag{2-8}$$

其中，N_j 为第j类测试文本数，N为测试文本总数。

2.5 小结

网络和通讯技术的迅速发展，使得短文本形式越来越丰富，本章对短文本的涵盖范围、特点及研究领域进行了详细的总结和分析，由于短文本的样本特征稀疏，且网络文本占据量大，目前针对短文本的研究领域多集中在信息过滤和热点发现上。本章还从文本预处理等五个方面对文本分类的相关技术进行了简要概述。

3 基于字共现的特征提取

3.1 基于字的文本特征

在文本分类中，样本特征的选定一般有字、词、短语等。经过实践证明在常见的文本集中，使用词作为文本特征能够取得较好的效果，因而多数分类均是以词为关键研究对象。如何准确分词，如何充分挖掘词间语义关系使其能更全面表达文本主题，成为人们研究的重点，有关对关键词特征扩展的方式方法层出不穷。以词作为特征项的方法在不断的研究中逐步趋于成熟，但是此方法对分词的依赖性很大，如果分词不当或者分词过程中不能很好的识别一些领域的专业词汇，反而会带来后续的一些麻烦，直接影响到分类的准确率。

基于字的文本特征是指以单个汉字作为特征项进行文本表示。人们通常认为，由于单个汉字孤立存在，不像词或短语一样结合其它汉字表示不同的语义，因而其对于文本的表示能力较差，在长篇文章中不能独立完整地表达语义信息，因此选取字作为文本特征的分类并不常见。

其实早期的中日韩语言处理就常以字作为文本的特征项，一方面是因为此方式较为简单。据国家语委和国家教委发布的《现代汉语常用字表》统计，汉语中常用字2500个和次用字1000个，仅3500个字就能概括平时使用的99%以上的汉字，相对于成千上万的词语来说，以字为特征的向量维数远小于以词为特征的向量维数，另外由于汉字的数量远远小于词语，使得在文本处理过程中运算速度和存储空间都相对较小，效率得到提高，同时有实验证明在效率提高的基础上，准确率并无明显下降；另一方面是以字作为文本的特征项避免了分词。目前分词还存在许多尚未解决的问题，如歧义判断、分词规范的确定、未登录词识别等等，这为后续的文本处理带来许多麻烦，分词不当会直接影响到分类的准确率。

当然，以字作为文本的特征也存在一些缺点，比如单个汉字的表示能力较差，不能进行概念扩展，同时可能会包含许多的无用信息（如介词、助词等），这些缺点也会影响到文本处理的准确率，在文本处理中还需要加以改进。

本文所研究的短文本，由于自身字数少、数量多，导致样本特征稀疏，使得常用的以关键词为特征的提取方法并不能取得很好的效果。短文本的字数少，提取出来的关键

词就会更少，这样在分类中容易忽略许多重要信息，表达能力相对较弱。因此，通过综合考虑，本文采用“字”作为短文本的特征。

3.2 共现分析

共现是指两个事物总是共同出现或两件事情总是同时发生。共现的概念最早出现在情报学中，通过对共现现象的分析可以更多的了解事物之间的关联性。随后共现关系被引用在文献计量学中，用来分析文献的内容、引文及关键词汇的相关度。词汇之间的共现关系可归纳为对立关系、重合关系、包含关系、上下义关系、相对无关关系、组合关系等。充分理解词汇之间的共现关系，可以帮助我们挖掘出许多词汇间相关联的语义关系及深层含义，同时对自然语言分析起着重要作用。目前，共现分析已被广泛地应用在文献检索、信息检索、文本分类、文本聚类等领域，利用共现分析挖掘语义信息能够取得较好的效果。

共现分析在文本分类中可定义为：在大规模语料库中，如果两个词汇经常共同出现在同一个窗口单元（如一定词语间隔、一句话、一篇文档等）中，则认为这两个词汇在语义上是相互关联的。而且，共现的频率越高，其相互间的关联越紧密。通过共现分析的这种关系，可以统计出词汇之间的相互关联性进而更准确地对文本进行处理。共现分析包含同引分析、共词分析、共篇分析三种类型，其中同引分析在文献计量学中有着广泛的应用。

在文本中，通常选取词作为基本特征，共现分析的概念也主要是针对词汇进行的，而汉字则是组成词汇的基本单位，字与字之间的关系和语义信息直接关联着词与词之间的关系及词义。在文本中，一篇文章所代表的含义是由句义及句与句之间的关系来表达的，一句话所代表的含义是由词义及词与词之间的关系来表达的，而一个词所代表的含义通常是由字义及字与字之间的关系来表达的。由此我们可假定，字与字之间的关系也能够充分表达出句子所代表的含义。特别是应用在短文本中，由于短文本往往只有一句或者几句话，词汇量较少，关键词的提取往往会忽略掉许多的文本的源信息，如果能充分挖掘出字与字之间的关系，则文本主题信息能够表达得更为精确。

如果两个字经常在同一个窗口单元中共同出现，则可以认为它们在语义上是相互关联的，在一定程度上反应了该本文的主题关联，从共现频率上也可表现出它们对文本的重要程度。例如在一篇短文本中，如果“四”、和“级”、“英”字经常同时出现，则

可以判定文本的内容与英语四级有关。

在此思想的基础上，本文提出了基于字共现的短文本特征提取方法。对于短文本信息，选取单个字作为文本的基本特征，通过计算字与字间的互信息并结合共现分析方法，使短文本的特征字具有更强的表达能力。同时本文还通过相关实验对该方法进行验证，结果显示其不但简单易于实现，而且避免了分词，减少了计算量，便得分类效率得到提高。

3.3 字共现模型构建

基于词共现模型[56]原理，结合短文本自身特点，本文选取以字为基本特征来构建共现模型，具体过程由数据处理、共现信息提取和构建共现特征向量三大部分组成，构建流程如图3-1所示。对训练文本集经过分字、去除停用字等预处理后，通过计算字间的相互信息量得到每个字的共现度量，采用TFIDF统计字的词频权重，由共现度量和词频权重相结合得到字的共现权重，以此来构建文本的共现特征向量。

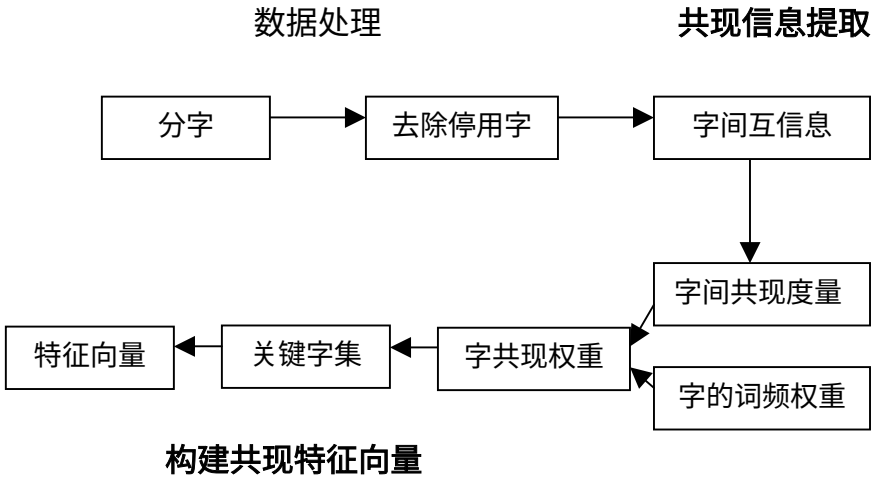

图 3-1 字共现模型构建流程图

3.3.1 数据处理

字共现模型的第一步是对数据进行处理。首先对文本数据进行分字处理，采用以字为特征的数据处理较为简单，并且省去分词的步骤，避免了后续因分词不合理而带来的一系列麻烦。其次是去除停用字，虽然短文本的字数少，但一些停用字如“ 了 ”、“ 的 ”、“ 在 ”等的存在，并无多大实际意义，而且它们通常出现的频率较高，反而会阻碍共现信息的计算，因此在数据处理阶段需要对组成短文本的字进行过滤，滤去高频的停用词，

此外，还要滤去同样无存在意义的代词和连词如“我”、“和”等。滤去高频词可以压缩空间维数，提高程序运行速度和效率，增强分类精度。

如下图3-2是给定一个新浪微博短文本信息（共11个字符），对此短文本文件进行分字后，得到由字符组成的集合，再对这些字符依照停用字表进行匹配，删除相匹配的停用字后得到特征集（共6个字符）。

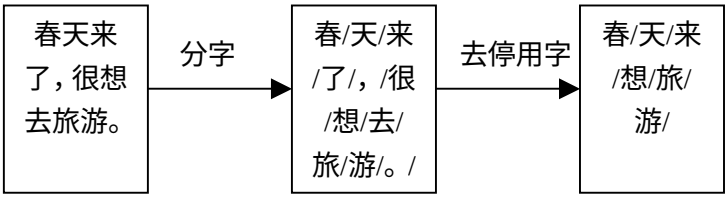

图 3-2 微博信息数据处理

图3-3是给定一个搜狗新闻网页的短文本（共有160个字符），经过分字去除停用词后得到的特征集（共101个字符）。

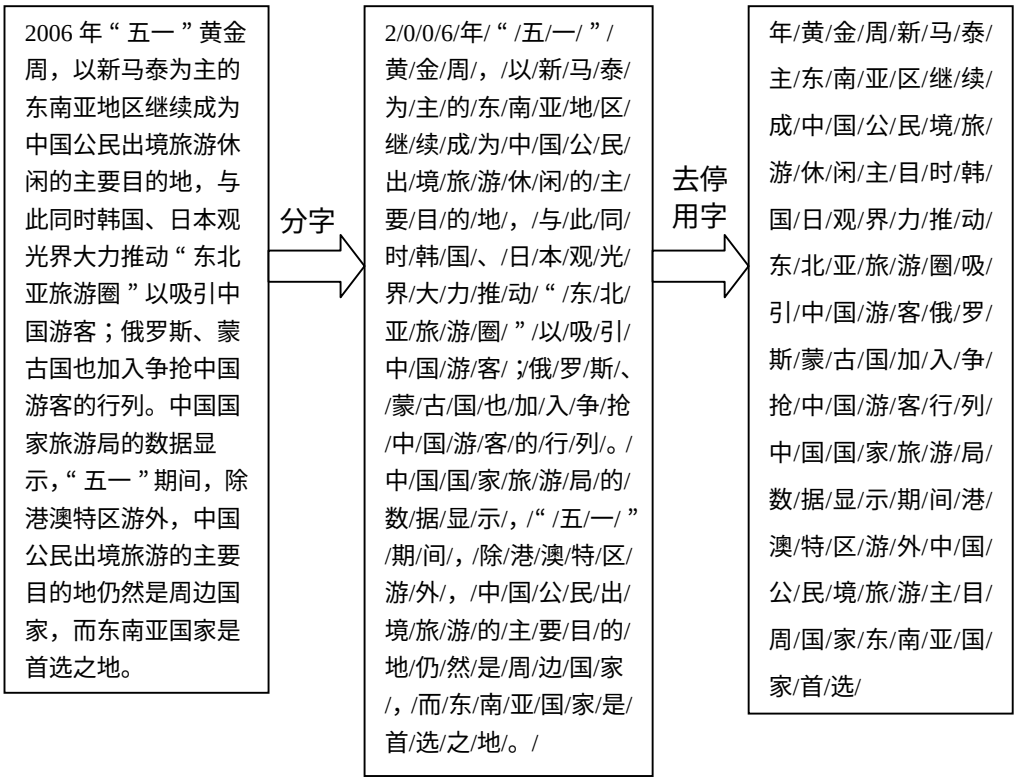

图 3-3 搜狗新闻短文本数据处理

本文选用的停用词表中含有包括符号和数字在内的1028个常见停用词，经过筛选得到含有362个停用字的停用字表，包含数字、符号、连词等。通过对短文本中特征字与停用字表的匹配，删除相匹配的停用词。

很多文本分类在分词的过程中，通常会删去一些低频词，特别是只出现一次的词。那是在大部分文本中，分词后得到的特征词可能就会达到数百上千个，往往比短文本的字数还要多，由于短文本本身字数就较少，在滤去停用词（字）、代词和连词等无意义词后，所保留的字数就更少了，因而不再对只出现一次的字进行删除，保留了文中全部具有存在意义的字。

3.3.2 共现信息提取

数据处理完后，就要根据提取出的文本特征字进行共现分析。共现分析概念中指明需要分析的是经常共同出现在同一个窗口单元(如一定词语间隔、一句话、一篇文档等)中的两个样本，一般的文本分类共现窗口的选择是以篇为级别。本文在进行共现分析过程中，选取一个短文本作（即是一篇文档）为一个窗口单元。由于短文本字数较少，有些可能就一句话，因而不选取句子作为窗口单元，省去分句的环节，并且一个文本通常描述一个主题，以整个文本作为窗口更具意义。

共现分析中分析特征之间的相关度主要Dice指数、余弦指数和Jaccard指数三种方法，其中Jaccard指数能够根据特征之间的共现频率更直观地反映两特征之间的相关度，应用更为广泛。并且，由于Jaccard指数对于高频词和低频词的共现并无过于明显的分界，因而，更适用于特征稀疏的短文本中，短文本由于字数少使得低频词可能涵盖更多意义。

Jaccard 指数计算公式如下：

$$J_{ij} = \frac{f_{ij}}{f_i + f_j - f_{ij}} \tag{3-1}$$

用W_i和W_j分别代表字i和字j，在（3-1）式中，f_{ij}是W_i和W_j共同出现的次数，f_i是W_i在文本中出现的次数，f_j是W_j在文本中出现的次数。换个角度，Jaccard指数J_{ij}也即为W_i和W_j的共现概率$P(i, j)$。

在短文本的特征字集中引入共现概念后，通过计算字与字之间的互信息就可以对字与字之间的关联度进行量化，得到字与字之间的共现信息。互信息来源于信息论，是一个基于熵的信息度量概念，它用来度量两个随机变量间的统计相关性。两个特征字i和j

之间的互信息的计算公式如下：

$$MI(W_i,W_j)=\log\frac{P(i,j)}{P(i)P(j)} \tag{3-2}$$

计算过程中以一个短文本作为一个窗口单元，其中，

$$P(i,j)=J_{ij}=\frac{f_{ij}}{f_i+f_j-f_{ij}} \qquad P(i)=\frac{f_i}{\sum f} \tag{3-3}$$

公式（3-3）中，f_{ij}是一个短文本（窗口单元）中W_i和W_j共同出现的次数，f_i是W_i在短文本中出现的次数，f_j是W_j在短文本中出现的次数，$\sum f$为一个短文本中的总字数。MI值越大，说明W_i和W_j的关联程度越强，MI值为零时，W_i和W_j无关联。

互信息充分考虑了文本中的低频词对文本重要性，使得低频词可能具有较大的信息值。过度考虑低频词在有关长文本的分类中是互信息存在的缺点，但在短文本中由于字数少，低频词大量存在，这使得互信息表现出了具有针对性的优势。

对于数据处理阶段提取出的特征字，通过统计字频及共现频率进行共现信息计算，并计算出$P(i)$、$P(j)$和$P(i,j)$的值，依据公式（3-2）得出字与字之间的互信息，以此进行字与字之间的相关性度量。

3.3.3 构建共现特征向量

(1) 特征项的向量空间表示

向量空间模型（Vector Space Model，简称VSM）是目前应用比较广泛且效果较好的特征表示模型之一，由Salton等人于60年代末提出，最早应用于信息检索领域。它是由一组规范化正交词条矢量所组成的n维向量空间（n为文本中特征项的数目），每个文本表示为向量空间中的一个点，而点与点之间的距离就代表了文本之间的相似度。在文本分类领域，VSM使用也最为广泛，它把一个文本表示为特征空间中的一个特征向量，该向量由维度和权值组成。特征向量中的每一个维度对应文本中的特征项集，维度的权值表示与其对应的特征项在该文本中的权重。这样，每个文本d表示为n维向量空间中的一个点（n表示文本中特征项的数目），特征项权重ω作为向量空间中的分量，表示形式为：

$$V(d_i)=\{\omega(t_1,d_i),\omega(t_2,d_i),\cdots\cdots,\omega(t_n,d_i)\} \tag{3-4}$$

其中，d_i代表第i个文本，$\omega(t_k,d_i)$（$1\le k\le n$）表示第k个特征项在d_i中的权重。权

重表示该特征项在文本中的重要程度，权重值越大说明该特征项的表示能力越强，反之权重越小表示能力就越弱。权重的计算主要依据两方面：一是，特征项在文本中出现的频率越高，它与文本表达主题就越相关；二是，特征项在文本集中出现的频率越高，它表示某个文本特征的能力就越弱。

向量空间模型简单易实现，使文本的可操作性、可计算性以及匹配效率都得到了较大提高，但是它的前提假设是特征项在文本中出现的次序是无关紧要的，提高计算效率的同时，会失去文本的部分结构和语义信息。

特征项权重计算通常采用TFIDF公式。TF是词频，即特征项在一个文本中出现的频率，如果一个特征项在某文本中经常出现，说明该特征项对此文本的表达具有重要意义，一般词频高的特征在文本中被赋予较高的权重，TF值越大，表示该特征项越重要。IDF是反比文本频率，是特征项在文本集中分布的量化，IDF越大，说明该特征项分布越集中，在区分文本内容类别的能力越强。

TFIDF权重计算公式归一化表示如下：

$$w(t,d)=\frac{tf(t,d)\times\log(N/n_t+0.01)}{\sqrt{\sum_{t\in d}[f(t,d)\times\log(N/n_t+0.01)]^2}} \tag{3-5}$$

其中，$w(t,d)$ 为特征项t在文本d中的权重，$tf(t,d)$ 为特征项t在文本d中的词频，N为训练集文本总数，n_t 为训练集中出现特征项t的文本数，分母为归一化因子。

(2) 共现度量

特征项之间的互信息不仅充分表达了字之间的相关性，也表明了共同出现的特征字对文本内容的重要性，因而在特征项权重计算中加入字的共现度量F使得特征项能够更充分的表达文本内容。

共现度量F的算法描述如下：

Step1：针对每一个短文本（窗口单元），对其中的每个特征字对 i 和 j，按照公式（3-2）和（3-3）计算出它们的互信息 $MI(i,j)$。

Step2：依据 $MI(i,j)$，分别对每一个相关的特征字 i 和 j 进行赋值，作为它们各自的共现信息量。

Step3：扫描文本中的所有特征字，分别统计出每个特征字的共现信息之和，其中特

征字 i 的共现信息之和为 $\sum_{j=1}^{n} MI(i,j)$，并计算出文本中所有特征共现信息总和 $\sum_{i=1}^{n}\sum_{j=1}^{n} MI(i,j)$。

Step4：依照公式 F＝ $\sum_{j=1}^{n} MI(i,j) / \sum_{i=1}^{n}\sum_{j=1}^{n} MI(i,j)$ 得到每个特征项的共现度量 F。

(3) 共现特征向量

由于共现度量充分体现了特征字之间相关联的重要信息，因而将所得到的每个特征字的共现度量与特征项的词频权重之和，用以表示特征字的共现权重，依据共现权重作为特征提取的衡量。综合考虑词频权重和特征字的共现度量信息，给出共现权重公式如下式所示：

$$W(t,d) = \lambda \cdot w(t,d) + (1-\lambda)F_t \qquad (3\text{-}6)$$

其中 $w(t,d)$ 为文本d中特征字t的词频权重，F_t 为特征字t的共现度量，λ 为共现度量参数，在本文中取 λ =0.4。

由此，加入共现分析概念，以特征项的共现权重W作为向量空间中分量的文档d的特征向量表示为：$V(d) = \{W(t_1,d), W(t_2,d), \cdots\cdots, W(t_n,d)\}$。将所有文本的特征向量组合成一个向量空间，依据共现权重按从大到小的顺序排列，取前m个作为文本的关键字集，以此得到的共现特征向量用于文本分类。

3.4 实验测试与结论

实验的主要目的是比较本章中提出的基于字共现的特征提取方法的性能，实验使用的是基于向量空间模型的文本表示，分类效果评估使用准确率P、召回率R和F1值来衡量。

3.4.1 实验说明

实验采用各具不同代表性的两部分数据集：一部分是搜狗实验室文本分类语料库 http://www.sogou.com/labs/dl/c.html，共下载1246个短文本语料，内容为网页新闻，文字数量均等，主题明确；另一部分是新浪微博信息http://t.sina.com.cn/，使用网页采集大师从新浪微博中采集出1058个短文本语料，文字数量不等，但都在200个字符以内，内容较为随意分散。

数据集共有2304个短文本，内容涵盖十个类别，分别是汽车、财经、IT、健康、体育、旅游、教育、招聘、文化和军事。抽取其中的1612个文本作为训练集，剩余692作为测试集，训练集类别分布如图3-4所示。

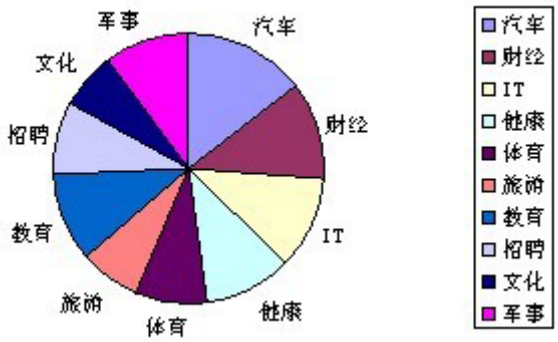

图 3-4 训练集类别分布

实验前先进行数据的预处理，对训练集文本分字并根据停用字表进行匹配，去除停用字，得到以字为特征的集合。为比较分字或分词对文本的影响，本文选用了中国科学院所研制的汉语词法分析系统ICTCLAS[60]对短文本进行分词处理，得到由关键词组成的文本集合。

有文献证实分类过程中并不是选择的维数越高分类效果就越好[57]，本文所用语料均为短文本，且采用以字作为文本特征，由于汉字数远小于词汇数，因而以字为特征的向量维数远小于以词为特征的向量维数。经过反复测定，在本文实验中，选取的特征维数为1000维。本实验重点在于分析文本特征，因而选用传统的KNN分类算法进行分类，K的取值为20。

3.4.2 实验分析

本章实验包括三个方面，分别是基于字共现的短文本分类比较；基于字共现的搜狗新闻和新浪微博两个数据集之间的分类比较；基于分字和分词的短文本分类比较。

（1）基于字共现的短文本分类比较

该实验对数据集分别进行加入字共现度量前后分类情况进行对比，分别针对十个类别分类情况通过准确率、召回率和F1值进行比较，结果如表3-1所示。其中，实验1为传统向量分类情况，实验2为基于字共现向量分类情况。另外针对十个类别F1值的柱状图对比如图3-5所示。

表 3-1 基于字共现的短文本分类比较

	准确率 P（%）		召回率 R（%）		F1 值（%）	
	实验 1	实验 2	实验 1	实验 2	实验 1	实验 2
汽车	88.31	92.46	80.46	85.35	84.20	88.76
财经	76.39	82.42	78.23	83.19	77.30	82.80
IT	80.52	88.35	75.14	80.81	77.74	84.41
健康	79.25	83.24	82.18	79.72	80.69	81.44
体育	84.46	89.46	72.62	83.39	78.09	86.32
招聘	64.35	74.21	77.54	79.23	70.33	76.64
旅游	82.01	86.19	70.21	78.21	75.65	82.01
教育	75.84	82.16	68.31	79.53	71.88	80.82
文化	70.16	78.38	80.92	77.72	75.16	78.05
军事	68.46	73.26	79.41	76.15	73.53	74.68

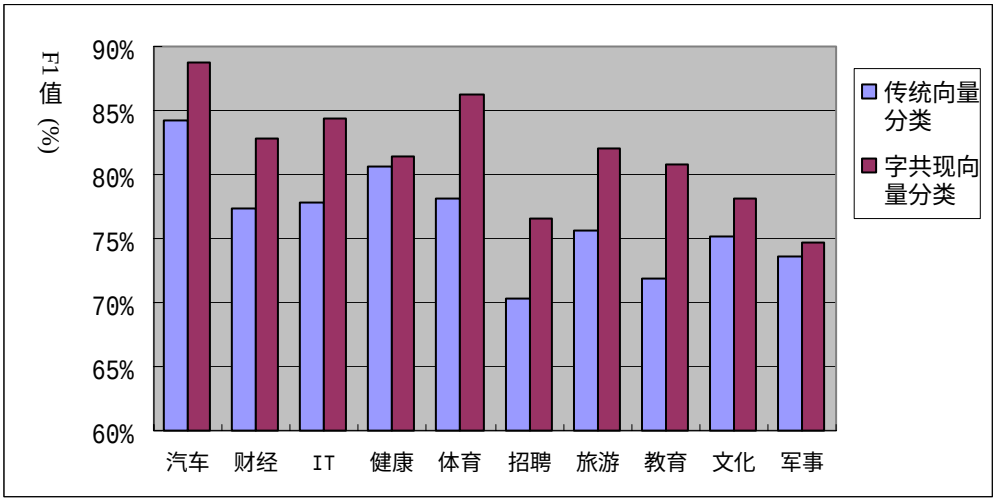

图 3-5 各类别的 F1 值对比

由表3-1和图3-5可以看出，基于字共现的特征提取方法在分类的效果评估中表现出较好的性能，字共现在原有特征基础上考虑了更多的语义信息，便得分类结果更为准确。通过表3-1还可以看出，汽车、体育、IT在分类中性能较好，而招聘和军事相对较差，各类别分类性能的差别除了与算法相关外，还会受到其它因素的影响，如类别特征不明显等。

（2）不同数据集间的比较

该实验分别对搜狗新闻数据和新浪微博数据集中的短文本进行分类比较，对两类数据集分别采用了基于传统向量的分类方法和基于字共现向量的分类方法进行对比，结果如图3-6所示。

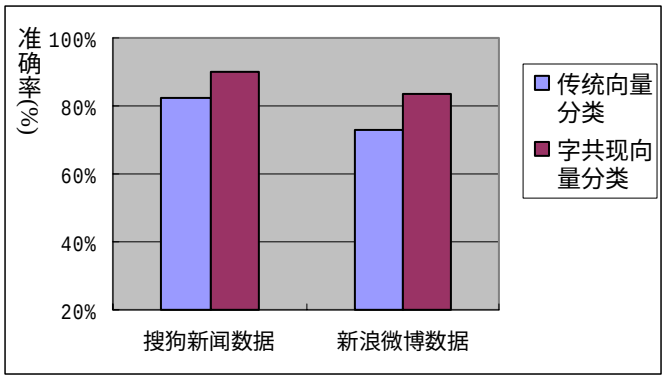

图 3-6 不同数据集分类比较

由图3-6可以看出，搜狗新闻数据的分类正确率要高于新浪微博数据。这是由于搜狗新闻数据文字数量均等，字词表达含义专业性较强，且主题内容明确集中，在分类中取得较好的效果；而新浪微博信息文字数量多少不等，有些甚至就几个字，且用语比较随意奇异短语多，表达内容较为随意分散，主题表现相对较弱，因而分类的准确率稍低，对于此类网络文本，可采取进一步的文本语义扩充或对奇异短语进行识别，以达到更好的分类结果。

（3）分字和分词对短文本分类的影响比较

为了比较选取短文本特征对分类的影响，在此做了基于分字和分词的文本分类比较。分别对搜狗新闻和新浪微博两个不同的数据集进行分字和分词处理，提取特征后进行分类比较，结果如图3-7所示。

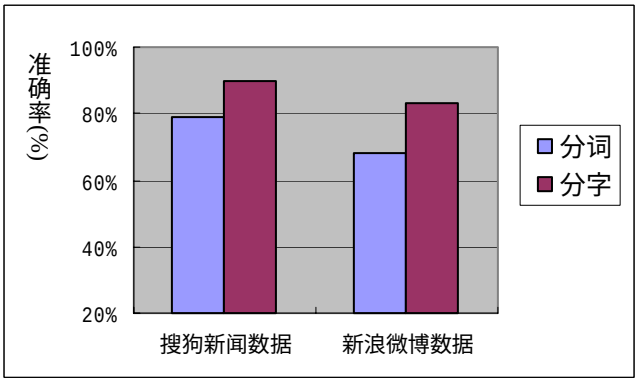

图 3-7 短文本分字分词准确率比较

由上图可以看出，对于短文本数据来讲，采用以字为特征进行分类，性能评估要高于以词作为特征的分类，这主要是由于短文本内容较少，样本比较稀疏，用词汇进行特征表示会失去过多的语义信息。同时从图中可以看出，对于新浪微博数据来讲，分别以字和词作为特征进行分类，结果差别比较大，这是由于微博数据内容较少且随意分散，使得以词汇作为特征的主题表达就更显困难；而搜狗新闻数据由于文本主题明确，文本形式比较标准，因而在分字和分字准确率上差别相对较小一些。

3.5 小结

本章从短文本独具的特点出发，采用“字”作为文本特征项，并结合共现分析概念提出了基于字共现的短文本特征提取方法。通过实验测试可以看出，选取“字”作为文本特征对于短文本的分类能够取得较好的效果，同时不管对于标准的新闻数据还是内容较为随意的微博数据，基于字共现的特征提取都能够挖掘出更多的文本语义相关信息，在文本分类中表现出较好的性能。

4 改进的基于近似域 KNN 分类

KNN算法简单直观，实现起来容易，分类准确率高，并且加入新的训练文本不需重新训练，从而减少了训练时间，但是使用KNN算法在对文本进行分类时，需要把训练文本集中所有的文本都存储起来，便于计算待测样本与各训练文本之间的相似度，这使得计算量和存储量开销较大，分类速度减慢。针对KNN算法的这一缺点，本文在KNN分类算法中通过对训练集中各类别进行区域划分，提出一种改进的基于近似域KNN分类方法。

此方法的基本思想是：通过计算测试集各样本与训练集中每个类别的中心域和近似域之间的距离，来判断待测样本的分布。处于某类别中心域内的样本必然属于此类，处于某类别近似域之外的样本一定不属于此类，而对处于某类别近似域内的样本再利用KNN算法进行分类。此方法减少了KNN算法的计算量和存储量，缩减了分类速度，通过实验证明其使得分类效率和准确率均得到提高。同时，为减少样本的误判率，对处于各类别边界区域的样本，在对其进行类别权重判断时设定边界参数，加大类别权重，这样使得分类准确率更高。

4.1 KNN 分类介绍

KNN（K-Nearest Neighbor）分类算法又称为K近邻分类算法，是由Cover和Hart于1968年提出，是基于实例学习的最基本算法之一，也是模式识别非参数的重要方法之一。基本思想是：对于一个待测样本，计算它与训练文本集中每个文本的相似度，依据文本相似度找出k个最相似的训练文本，通过计算每个训练文本的类别权重，将待测样本分到权重最大的那个类别中。

KNN类别权重计算公式为：

$$y(x,C_j)=\sum_{d_i\in KNN} sim(x,d_i)y(d_i,C_j) \tag{4-1}$$

$$y(d_i,C_j)=\begin{cases}1，\ d_i\text{属于类别}C_j\\0，\ d_i\text{不属于类别}C_j\end{cases} \tag{4-2}$$

其中， $sim(x,d_i)$ 表示测试文本 x 和训练文本 d_i 之间的相似度， d_i 是 x 的k个最近邻之一。

一般情况下，相似函数 $sim(x,d_i)$ 采用向量夹角的余弦值表示，如下式：

$$sim(x,d_i)=\frac{\sum_{k=1}^{m} w_k \times w_{ik}}{\sqrt{(\sum_{k=1}^{m} {w_k}^2)(\sum_{k=1}^{m} {w_{ik}}^2)}} \tag{4-3}$$

其中，m为特征向量的维数， w_k 为向量的第k维。

KNN算法直观且便于理解和应用，在实际应用中非常有效，是目前应用最为广泛的文本分类算法之一。但是应用KNN进行分类时，每个待测样本都要与所有的训练文本进行距离计算，计算量和开销都比较大。另外，采用KNN方法需要合理选择k，k的选择很大程度上决定了分类性能的好坏。

目前对于KNN分类算法的改进可分为两大类：一是通过快速搜索算法，在尽量短的时间内找到测试样本的最近邻；另一类是通过使用小量样本库来减少样本相似度的计算量，此类方法是通过对样本特征空间的降维和选择一些代表性样本来达到减少训练集的目的。本文主要讨论后一种方法，对于此方法传统的具有代表性的算法有Hart 的Condensing算法、Wilson的Editing算法和Devijver的Mul2tiEdit算法，Kuncheva使用遗传算法在这方面也进行了一些研究。但是这些方法在训练样本集中每增加或删除一个样本时，都要对样本进行一次测试，反复迭代直到样本集不再变化，这对于有成百上千的训练文档来说，其工作量也是非常大的。鉴于此，本文采用一种改进的KNN算法，根据样本在类别区域中的分布对其进行分类。

4.2 改进的 KNN 分类方法

改进的基于近似域KNN分类方法是对训练集中各类别进行区域划分，以此作为待测样本在类别中分布的度量，具体的分类流程由文本处理、判断样本分布和改进的KNN分类三大部分组成，如图4-1所示。

首先对短文本训练集进行数据的预处理和特征提取，得到由向量空间表示的特征项向量集合；在此基础上，以训练集中各类别中心为基点分别计算出每个类别的中心域和近似域范围，以此作为待测样本在类别中分布的度量；然后通过欧氏距离计算待测样本i与各类别中心向量之间的距离，以此距离和各类别区域半径的对比来判定待测样本i在向量空间中的分布；只对处于类别近似域内的样本利用KNN算法进行分类，分类过程中，为了减少样本的误判率，对处于各类别边界区域的样本，在对其进行类别权重判断时设

定边界参数，加大类别权重；最终输出分类结果。

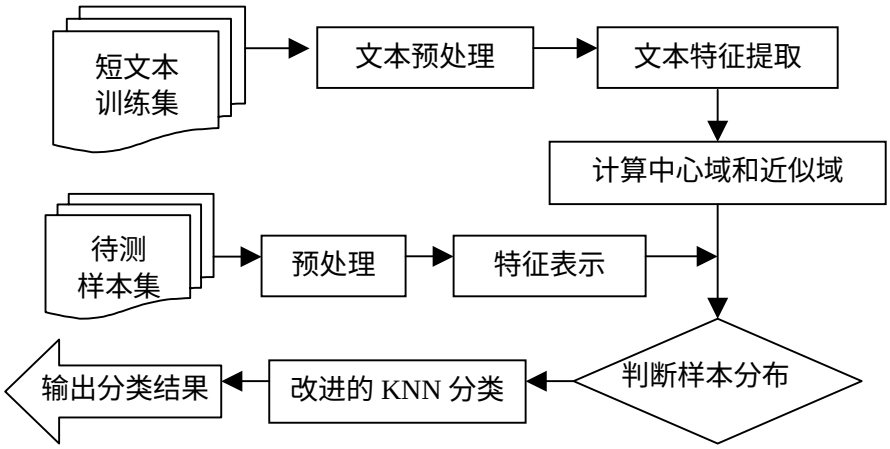

图 4-1 改进的基于近似域 KNN 分类流程图

4.2.1 文本处理

对训练集和测试集中的文本进行分字、去除停用词等预处理后，对特征项进行文本表示和特征提取，把特征项通过VSM表示为向量的形式。本章采用的向量集是上章实验得到的短文本共现特征向量集合。

4.2.2 判断样本分布

本章通过事先对训练集中各类别进行的区域划分，以此作为待测样本在各类别中分布的度量，寻找处于类别近似域的待测样本，使用KNN算法进行更进一步的细致分类。在模式识别中，为了能在某空间中进行分类，通常假设同一类的各个样本在空间中组成一个紧致集[58]，也即是属于同一类的样本从整体来说都分布在一个足够大的区域内，大部分样本都聚集在区域类别中心的周围。鉴于此，代表每个类的向量都可以通过类的中心向量来表示[59]。

在本文中，假设训练集中每个类别中的样本都聚集在类别中心的周围，距离类中心越近的区域，该类别的样本分布越密集。训练集中每个类别类中心向量的计算即是该类中所有文本向量的几何平均值：

$$v(\mathrm{C_k}) = \frac{V(d_1)+V(d_2)+\cdots+V(d_n)}{N} \tag{4-4}$$

判断样本分布的具体过程为：通过设定阈值，确定每个类别的中心域和近似域范围

及区域半径，将待测样本和各类别中心向量之间的距离d与每个类别的区域半径r进行对比，以此来确定待测样本在各类别区域中的分布。如果距离d小于最小的半径r，则样本落在类别的中心域内一定属于此类别；如果距离d大于最大的半径r，则样本落在类别的近似域之外一定不属于此类别；如果距离d在最大半径和最小半径之间，则样本落在类别的近似域内，对落入此区域的样本再进行KNN分类。

（1）类别区域

在文本向量中可以设定：若向量空间中存在某一区域，使得分布于该区域中的所有特征向量只属于某类别C，则称满足条件的区域为类别C的中心域，记为$O(C)$；若存在某一区域，使得类别C中的特征向量全部分布在其中，则称满足条件的区域为类别C的近似域，记为$A(C)$。

（2）判断样本分布

设训练集中文本向量为$V(d)=\{W(t_1,d),W(t_2,d),\cdots\cdots,W(t_n,d)\}$，其中W为文本$d$中特征项$t$的权重，类别$C=\{C_1,C_2,\cdots\cdots,C_n\}$，类$C_k$的中心向量用$v(C_k)$表示，r是以$v(C_k)$为中心的区域半径，随着区域半径r的缩小，属于$C_k$类的文本所占比重就会越高，反之就会降低。

在判断样本分布之前，首先计算出待测样本到训练集中每个类别中心的距离。在此，距离的计算采用欧氏距离法。欧氏距离（Euclidean distance）也称欧几里得距离，是广泛采用的距离定义，它是在m维空间中两个点之间的真实距离。点$A=(x_1,x_2,\cdots,x_n)$和点$B=(y_1,y_2,\cdots,y_n)$之间的欧氏距离定义如下式所示：

$$(A,B)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \qquad (4\text{-}5)$$

① 设定阈值

根据样本在空间中的分布，分别设定阈值θ_1和θ_2（$0<\theta_1<1$，$0<\theta_2<1$），以此来确定中心域$O(C)$和近似域$A(C)$的范围，并设r_1和r_2分别为$O(C)$和$A(C)$的区域半径。则当θ_1趋近于1时，以$v(C_k)$为中心，r_1为半径的区域U_1中的向量集内的大部分向量都属于C_k类，区域U_1接近于类C_k的中心域$O(C)$；当θ_2趋近于0时，C_k中的大部分向量都属于以$v(C_k)$为中心，r_2为半径的区域U_2中的向量集，则区域U_2接近于C_k的近似域$A(C)$。

② 判断样本分布

确定了待测样本和各类别中心向量之间的距离d以及每个类别的区域半径r之后，即可以通过它们之间的对比值来判断待测样本在各类别区域中的分布。判断样本分布流程图如下所示：

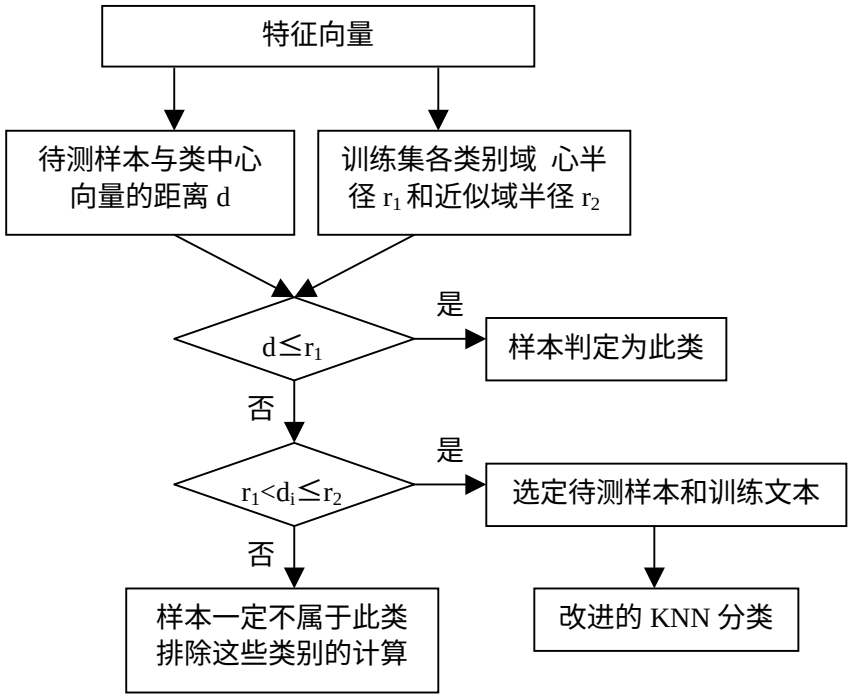

图 4-2 判断样本分布流程图

判断样本分布的算法描述如下：

Step1：对于训练文本集中的每个类别 C_k，根据公式（4-4）计算各类别的中心向量 $v(C_k)$。

Step2：通过欧氏距离公式（4-5），得到待测样本 i 与训练集中每个类中心向量 $v(C_k)$ 之间的距离 d_i。

Step3：设定阈值 θ_1 和 θ_2，确定中心域半径 r_1 和近似域半径 r_2。

Step4：比较距离 d_i 与半径 r_1 和 r_2 的值，判断样本的分布区域；

a、 如果 $d_i \le r_1$，则样本 i 落在类 C_k 的中心域内，一定属于 C_k 类；

b、 如果 $r_1 < d_i \le r_2$，则样本 i 落在类 C_k 的近似域内，可能属于 C_k 类；

c、 如果 $d_i > r_2$，则样本 i 落在类 C_k 的近似域之外，一定不属于 C_k 类。

如果形象的把类别中的各个区域表示为一个球形空间，则有关类别C_k的区域分布的空间切面图可直观的表示为图4-3所示，其中 $v(C_k)$ 为类C_k的中心向量，r_1 为中心域半径，r_2 为近似域半径，d_i 为待测样本i到类中心向量 $v(C_k)$ 的距离。由图中可以清晰地看出，以类中心向量 $v(C_k)$ 为中心，r_1 为半径的区域为类别C_k的中心域，在此区域内所有的样本都属于类别C_k；以 $v(C_k)$ 为中心，r_2 为半径的区域为类别C_k的近似域，在其间的样本可能属于类别C_k。

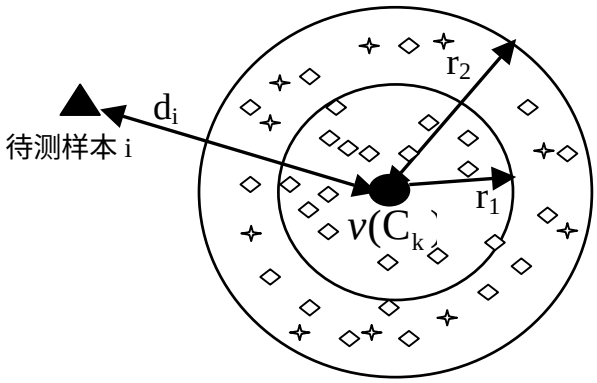

图 4-3 类 C_k 区域分布的空间切面图

4.2.3 改进的 KNN 分类

KNN作为文本分类最好的算法之一，虽简单易实现，但计算量却很大，上节经过事先对训练集中各类别进行了区域划分确定待测样本的分布区域后，只需要对分布在类别近似域内的样本再利用KNN算法进行分类即可，这样可以有效缩减计算量，很大程度上提高了KNN的分类效率。

处于类别近似域的样本，有一部分会因为处在区域的边界被误分，因而本文设定边界参数 λ 对类别权重计算公式加以改进，尽量避免样本误分的情况，提高分类的准确率。设定 λ 的值为待测样本到各类别中心向量之间的距离的倒数，即 $\lambda = 1/d$，则KNN类别权重计算公式可改为：

$$y(x, C_j) = \sum_{d_i \in KNN} sim(x, d_i) y(d_i, C_j) \lambda \qquad (4\text{-}6)$$

其中，$y(d_i, C_j) \in \{0,1\}$ 表示文本 d_i 是否属于类 C_j（是 y =1，否 y =0）；$sim(x, d_i)$ 表示测试文本 x 和训练文本 d_i 之间的相似度，d_i 是 x 的k个最近邻之一。

本文中利用KNN进行分类的算法描述如下：

Step1：依照判断样本分布算法所找出待分类样本 i 落入其近似域内的所有类别集合 C={C_1, C_2 ,……,C_n}，选定这些类别中所有文本组成的集合 D。

Step2：根据向量夹角余弦公式（4-3）计算待分类样本 i 与 D 中各文本向量之间的相似度 $sim(i, d_k)$，并按倒序排列。

Step3：依照 $sim(i, d_k)$ 的排列顺序，选出相似度最大的前 K 个文本。

Step4：根据类别权重计算公式（4-6），计算待测样本 i 的 K 个最近邻文本中每个类别的权重 $y(i, C_j)$。

Step5：比较各类别的权重，将样本 i 分到权重最大的类别中。

4.3 实验测试与结论

实验的主要目的是比较本章中提出的改进的基于近似域的 KNN 分类算法性能，实验中使用的是基于向量空间模型的文本表示，分类效果评估使用准确率 P、召回率 R 和 F1 值来衡量，计算公式依据（2-6）如下：

$$P_i = \frac{l_i}{m_i} \times 100\%，\mathrm{R_i} = \frac{l_i}{n_i} \times 100\%，F1_i = \frac{P_i \times \mathrm{R_i} \times 2}{P_i + \mathrm{R_i}}$$

l_i 表示分类的结果中被标记为第i类别且标记正确的文本个数，m_i 表示结果中表示被标记成第i个类的文本个数，n_i 表示被分类的文本中实际属于第i个类别的样本个数。

4.3.1 实验说明

本章的实验是在第三章实验的基础上进行的，所选用的数据集依然是搜狗实验室文本分类语料库和新浪微博信息，数据集共为 2304 个短文本，内容涵盖十个类别，分别是汽车、财经、IT、健康、体育、旅游、教育、招聘、文化和军事，抽取其中的 1612 个文本作为训练集，剩余 692 作为测试集。训练文本本集类别分布如表 4-1 所示。

表 4-1 训练文本集类别分布

类别	文本数	所占比率	类别	文本数	所占比率
汽车	237	14.7%	旅游	118	7.3%
财经	187	11.6%	教育	174	10.8%
IT	176	10.9%	招聘	143	8.9%
健康	164	10.2%	文化	106	6.6%
体育	139	8.6%	军事	168	10.4%

4.3.2 实验分析

本次实验包括两方面：改进前后KNN算法分类效率比较和改进前后KNN算法分类性能比较。

（1）改进前后KNN算法分类效率比较

由于KNN在分类时需要对所有的样本进行相似度比较，使得运行速度较慢，该实验主要分析改进前后两种算法的运行速度，对比图如4-4所示。

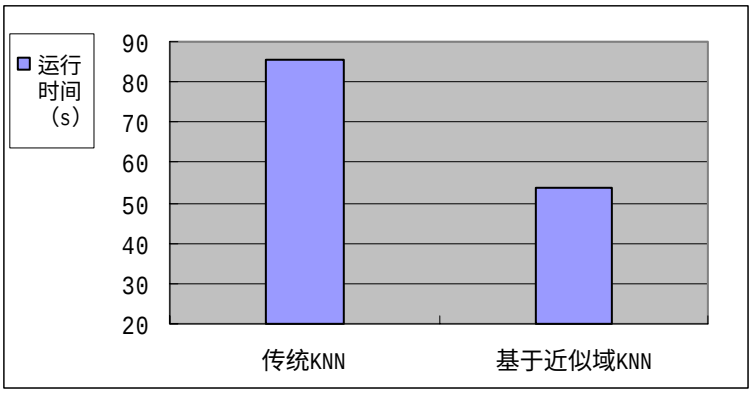

图 4-4 改进前后 KNN 算法分类效率比较

由图可以看出，由于改进的KNN分类算法事先对训练集各类别的文本区域进行了划分，通过判断待测样本的分布位置，只针对落在类别近似域内的各样本使用KNN进行分类，使得KNN在分类时相对缩小了搜索比对的范围，从而大大缩短了分类时间，使得分类速度更快，效率得到提高。

（2）改进前后KNN算法分类性能比较

该实验主要分析改进后算法的性能，因而对改进前后算法的分类结果进行对比，分别针对十个类别分类情况通过准确率、召回率和F1值进行比较，结果如表4-2所示。其中，实验1为传统KNN分类算法的分类情况，实验2为改进的基于近似域KNN分类的分类情况。另外针对十个类别F1值的柱状图对比如图4-5所示。

表 4-2 改进的基于近似域的短文本分类比较

	准确率 P（%）		召回率 R（%）		F1 值（%）	
	实验 1	实验 2	实验 1	实验 2	实验 1	实验 2
汽车	92.46	94.68	85.35	90.13	88.76	92.35
财经	82.42	86.43	83.19	82.69	82.80	84.52
IT	88.35	91.62	80.81	85.32	84.41	88.36
健康	83.24	85.54	79.72	75.13	81.44	80.00
体育	89.46	92.02	83.39	88.26	86.32	90.10
招聘	74.21	78.38	79.23	78.19	76.64	78.28
旅游	86.19	89.14	78.21	80.92	82.01	84.83
教育	82.16	84.32	79.53	81.11	80.82	82.68
文化	78.38	82.68	77.72	80.25	78.05	81.45
军事	73.26	76.55	76.15	79.31	74.68	77.91

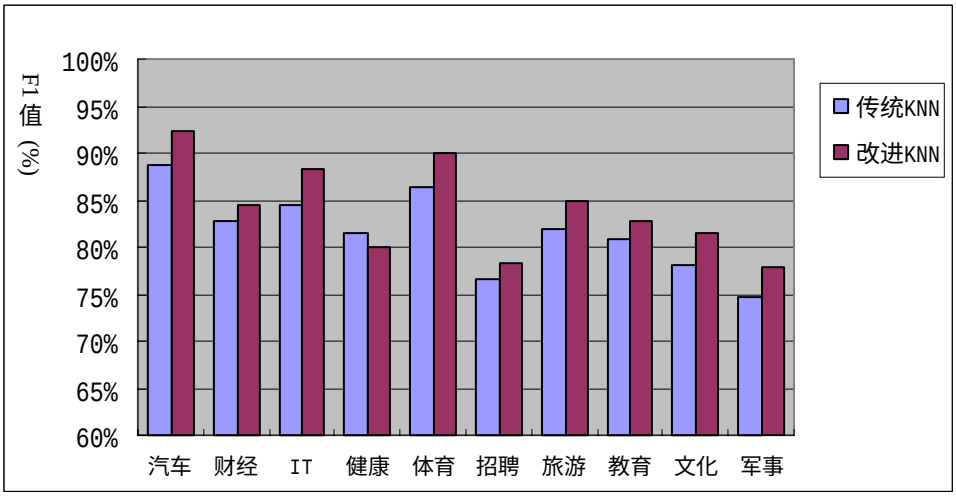

图 4-5 各类别的 F1 值对比

由图可以看出，改进的基于近似域KNN分类在分类的效果评估中表现出较好的性能，使得分类效率和准确率都得到了提高。该方法在分类先前进行了区域归属划分，使得一些特征相对比较明显或非常不明显的样本能够先被识别出来，而只对特征不是很明显的样本进行细致分类，很大程度上降低了分类的错误率。同时分类过程中设定了边界参数，降低了边界区域样本的误判率。图中还显示了出不同类别的准确率可能会有不同程度的差别，这是由于一些类别的主题刻画比较明显，特征表达较为集中使得分类效率更突出。

在实验中，阈值的选择起着重要作用，它影响着分类器的分类性能，只有选定适当的值才会使得在不影响效率的情况下提高分类的准确率。随着阈值的改变，虽然效率可能会提高，但同时准确率也会下降，因而阈值需要经过反复训练并通过各方面综合评定，以取得更好的效果。

4.4 小结

本章采用 KNN 算法对短文本进行分类，由于 KNN 计算量大，为此事先对训练集中各类别进行了区域划分，提出了改进的基于近似域 KNN 分类，该方法通过确定训练集中各类别的分布区域来断定待测样本的分布，只针落入各类别近似域的样本进行分类，实验证明该方法能够有效提高分类效率和准确率。

5 总结与展望

5.1 本文的工作总结

短文本随着互联网的迅速发展开始大量涌现，并且在信息过滤、热点发现、流行语分析及兴趣相关推荐等方面有着广泛的应用，本文对短文本分类的相关技术进行了研究，主要做了以下具体工作：

1、从搜索页面片断、微型博客及手机短信等方面，详细介绍了短文本的涵盖范围；针对稀疏性、实时性、不规范性和交互性来具体分析短文本自身存在的特点；从相似度计算、信息过滤等几大方面说明了短文本的研究领域。

2、经过研究和分析，采用“字”作为短文本的特征，进行特征表示和提取。由于汉字的数量远小于词汇数量，因而以字为特征的向量维数远小于以词为特征的向量维数，并且使得在文本处理过程中运算速度和存储空间都相对较小，效率得到提高。同时以字作为短文本特征还省去了分词，避免了因分词不当带来的麻烦。

3、结合共现分析的概念提出了基于字共现的短文本特征提取。通过分析字与字之间的共现信息，使特征字的语义表达更为明确，此方式在实验中表现出较好的性能。

4、通过实验分析了不同数据集对短文本分类的影响，以及分字和分词对短文本分类的影响。结果表明规范性文本在分类中能取得更好的准确率，且对短文本来说以字为特征比以词为特征分类性能要好。

5、针对KNN算法计算量大等问题，文中通过事先对训练集中各类别进行了区域划分提出了改进的基于近似域KNN分类，通过判定待测样本在训练集各类别区域的分布情况缩小选择区域，并针对处于各类别边界区域的样本，在类别权重判断时设定边界参数，加大类别权重，减少样本的误判率，实验证明该方法能有效提高分类的效率和准确率。

5.2 将来的工作展望

本文虽然完成了某些方面的测定，但是由于时间和条件限制，还存在一些问题，下面就文本所存在的问题和未完成工作提出进一步的展望。

1、文中只是考虑同一窗口内字间共现，并未进一步实验训练集中字间共现度量，

使得研究不够深入；

2、通过训练得到的阈值尚不稳定，需要进一步研究和改进。

3、对于不规范文本，有些特征并不明显使得类别界限不太明确，对分类性能造成一定影响，下一步可研究从不同角度对网络短文本信息进行扩充，使文本主题描述更为清晰。

4、本文的实验数据集较少且范围不够广泛，下一步需要搜集大量语料进行实验测试。

致 谢

时光飞逝，转眼间三年的研究生学习生涯即将结束，感谢母校河南大学提供如此安静舒适的学习和实验环境，感谢计算机与信息工程学院的培养与支持，使得我在三年时间无论从理论知识还是从技能技巧上都有很大收获，感谢在此期间帮助过我的各位老师、同学、朋友和家人们。

首先要感谢导师胡小华教授三年来的对我的精心培养和孜孜不倦的教诲，他渊博的知识、严谨的态度和认真的精神都对我产生了深远的影响，受益终身。胡老师在数据挖掘领域有着独道的见解和丰富的经验，几年来对我和师兄们进行不断的教导和关怀，他随和的性格和丰富的阅历都深深的影响着我们。论文从选题、修正到定稿，胡老师对我提出了很多宝贵的意见和建议，在此深表谢意。

感谢计算机学院的各位老师在这三年里的辛勤授课，他们深厚的专业功底、高超的水平和丰富的经验，对我的学习和研究提供很大帮助。

感谢师兄及各位同窗共读的同学们在学习和生活中对我的帮助，在与他们共同学习和讨论的过程中我得到了很大的进步。

感谢家人一直以来的理解和支持，他们是我坚强的后盾和精神支柱，在此对表示深深的感谢。

最后，感谢参加论文评审和答辩的各位老师的认真评阅，你们辛苦了！

参考文献

[1] 中国互联网络信息中心，第27次中国互联网络发展状况统计报告．http://www.cnnic.net.cn/dtygg/dtgg/201101/P020110119328960192287.pdf，2011-01-19．

[2] Joachims T．Text Categorization with Support Vector Machines Learning with Many Relevant Features[J]．Machine Learning，1998：137-142．

[3] Fuhr N，Hartmanna S，Lustig G,Schwantner M，Tzeras K．Air/X-Arule-based multi-stage indexing system for large subject fields[C]．Proceedings of Recherched'Information Assistee par Ordinateur(RIAO 1991)．1991：606—623．

[4] Luhn H P．Auto-encoding of documents for information retrieval systems[M]．Modem Trends in Documentation．New York：Pergamon Press，1959．

[5] Maron M E，Kuhn J L．On relevance，probabilistic indexing and information retrieval[J]．ACM，1960，7(3)：216-244．

[6] Good．I．J．，'Speculations concerning information retrieval'，Research Report PC-78，IBM Research Centre，Yorktown Heights，New York，1958．

[7] Lewis．D．D．Evaluating and optimizing autonomous text classification systems．In E．A．Fox，P．Ingwersen．&R．Fidel，Proceedings of the l 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval．New York：ACM，PP．246-254，1995．

[8] D．Merkl，Content-based document classification with highly compressed input data，Proc．Int．Conf．on Artilacial Neural Networks，Paris，France，l995．

[9] R．Ghani．S．Slattery and Yiming Yang．Hypertext categorization using hyperlink patterns and meta data The Eighteenth International Conference on Machine Learning(ICML'01)，pp．178-185．2001．

[10] Yiming Yang．Thomas AuR．Thomas Pierce and Charles W Lataimer．Improving text categorization methods for event tracking．Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'00)，PP．65-72．2000．

[11] Zhang Li Juan and Li Zhou-Jun．A Novel Rough Set Approach for Classification[C]．Granular Computing，2006 IEEE International Conference，2006，349-352．

[12] Sousa，T．et a1．Particle Swarm based Data Mining Algorithm for classification tasks[J]．Parallel Computing．2004，30(5-6)：767-783．

[13] Hothorn T．，Lausen B．．Douhle-bagging：Combining classifiers by bootstrap aggregation[J]．Pattern Recognition．2003，36(6)：1303-1309．

[14] Guo Gongde，，Wang Hui and Bell D A,et a1．A KNN Model-Based Approach and Its Application in Text Categorization[C]．In：CICLing2004．2004，559—570．

[15] Yu H．，Yang J．and Han J．．Classifying large datasets using SVMs with hierarchical clusters[C]．In：Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Washington D．C．，USA,2003，306-315．

[16] 侯汉清．分类法的发展趋势简论[M]．北京：中国人民大学出版社，1981．

[17] Mehran Sahami and Timothy D. Heilman. A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. www 2006, May 23-26, 2006, Edinburgh, Scotland ACM 1 59593 323 9/06/0005.

[18] D. Metaler, S. Dumais, and C. Meek. Similarity Measures for Short Segments of Text. proc. ECIR, 2007.

[19] W. Yih and C. Meek.Improving Similarity Measures for Short Segments of Text proc. AAAI, 2007.

[20] Xuan-Hieu Phan, Le-Minh Nguyen and Susumu Horiguchi. Learning to Classify Short and Sparse Text&Web with Hidden Topics from Large-scale Data Collections www2008/Refereed Track: Data Mining-learning April 21-25 2008. beijing, china.

[21] J Hynek, K Jezek, o Rohlik.Short Document Categorization-Itemsets Method[C]. In: PKDD 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, Workshop Machine Learning and Textual Information Access, Lyon, France, 2000: 14-19.

[22] D Song, P D Bruza, Z Huang et al. Classifying Document Titles Based on Information Inference [C] . In: proceedings of the 14th International Symposium on Methodologies for Intelligent Systems, Japan, 003: 297-306.

[23] 龚才春．短文本语言计算的关键技术研究[D]．北京：中国科学院计算技术研究所，2008．

[24] 樊兴华,王鹏.基于两步策略的中文短文本分类研究[J].大连海事大学学报,2008,34(3)：121-124.

[25] 王永恒．海量短语信息挖掘技术的研究与实现[D]．北京：国防科技大学，2006．

[26] 胡佳妮，郭军，邓伟洪等．基于短文本的独立语义特征抽取算法[J]．通信学报，2007，28(12) ：121—124．

[27] 樊迪．中文短文本自动分类技术研究[D]．北京：清华大学，2009．

[28] 覃张华.基于HNC理论的短文本语境框架提取实现[J].北京工商大学学报(自然科学版),25(5) ：49-52．

[29] 闫瑞．面向短文本的动态组合分类算法[J]．电子学报，2009．5：1019-1024．

[30] 吴薇．大规模短文本的分类过滤方法研究[D]．北京：北京邮电大学，2007．

[31] 王鹏．文本分类中利用依存关系的实验研究[J]．计算机工程，2010，46(3) ：131-133．

[32] 王细薇．基于特征扩展的中文短文本分类方法[J]．计算机应用，2009，29(3)：843-845．

[33] 胡吉祥．基于频繁模式的消息文本聚类研究[D]．北京：中科院研究生院，2006．

[34] 宁亚辉．基于领域词语本体的短文本分类[J]．计算机科学，2009，36(3) ：142-145．

[35] 盛宇利．自然语言理解心理学在短文本分类中的实证研究[J]．现代情报，2009．29(8) ：4-7．

[36] 黄永文．中文产品评论挖掘关键技术研究[D]．重庆：重庆大学，2009．

[37] 张卫．网络舆情分析中的特征提取研究[D]．北京：中国科学技术大学，2008．

[38] 何海江．一种适应短文本的相关测度及其应用[J]．计算机工程，2009．35(6) ：88-90．

[39] 尹洪章等．结合内容相似性和时序性的社会网络挖掘[J]．计算机工程，2008，34(1) ：83-85．

[40] 王乐利．短语消息聚类相关技术研究[D]．北京：国防科技大学，2008．

[41] 黄永光，刘挺，车万翔等．面向变异短文本的快速聚类算法[J]．中文信息学报，2007，21(2) ：63-68．

[42] Xia, Y., K. Wong, and W. Gao. NIL is not Nothing: Recognition of Chinese Network Informal Language Expressions. in 4th SIGHAN Workshop at IJCNLP. 2005.

[43] Java，A., et al., Why We Twitter: Understanding Microblogging Usage and Communities. Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007，2007.

[44] Paulson, L.D., Spam hits instant messaging. Computer, 2004. 37(4): 18.

[45] Chuang, S.L. and L.F. Chien. A practical web-based approach to generating topic hierarchy for text segments. in the 13th ACM conference on Information and knowledge management. 2004.

[46] Zelikovitz, S. and H. Hirsh. Improving short-text classification using unlabeled background knowledge to assess document similarity. in the 17th International Conference on Machine Learning. 2000.

[47] Fitzpatrick, L. and M. Dent. Automatic feedback using past queries: social searching? in Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval. 1997. Philadelphia，Pennsylvania， United States: ACM.

[48] Sahami, M. and T.D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. in the 15th international conference on World Wide Web. 2006.

[49] Fellbaum, C., Wordnet: an electronic lexical database. MIT Press；1998.

[50] 林鸿飞．中文文本过滤的逻辑模型[D]．沈阳：东北大学．2000.

[51] 黄萱菁，夏迎炬，吴立德．基于向量空间模型的文本过滤系统[J]．软件学报，2003，No．3：435-442．

[52] 黄曾阳．HNC(概念层次网络)理论[M]．北京：清华大学出版社，1998．

[53] Allan J，Carbonell J，Doddington G et a1．Topic Detection and Tracking pilot study：final report．Proceedings of the DRRPA Broadcast News Transcription and Understanding Workshop San Francisco：Morgan Kaufmann Publishers，1998：194—218．

[54] Wayne C．Multilingual Topic Detection and Tracking：successful research enabled by corpora and evaluation．Language Resources and Evaluation Conferance(LREC)，Greece，2000：1487-1494．

[55] Ikonomalkis M．，Sotos K．，TamPakas V．．Text classification using machine learning techniques[J]．WSEAS Transactions on Computem，2005，4(8)：966—974．

[56] saltonG.，McGnM.J.An introduction to modem information retrieval[M]，New York：McGraw-Hill，1983．

[57] Jingyang Li，Maosong Sun，and Xian Zhang．A Comparison and Semi-Quantitative Analysis of Words and Character-Bigrams as Features in Chinese Text Categorization．Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL，545–552．

[58] 边肇琪，张学工．模式识别[M]．北京：清华大学出版社，2000．

[59] 毛国君．数据挖掘原理与算法[M]．北京：清华大学出版社，2007．

[60] 张华平，刘群．中文自然语言处理开发平台[EB／OL]．http://www.nlp.org.cn．2002．

攻读硕士学位期间发表的学术论文及科研成果

崔争艳．基于语义的微博短信息分类[J]．现代计算机，2010，337 ：18-20．

中文短文本分类的相关技术研究

作者：崔争艳
学位授予单位：河南大学

本文链接：http://d.g.wanfangdata.com.cn/Thesis_D146009.aspx