代　号 10701　　学　号 1025121802

分　类 G354.2　　密　级 公开

题（中、英文）目 基于维基百科的中文短文本分类研究

Research on Chinese Short Text Classification

Based on Wikipedia

作 者 姓 名 范云杰　指导教师姓名、职称 刘怀亮 教授

学 科 门 类 管理学　学科、专业 情报学

提 交 论 文 日 期 二〇一三年三月

## 西安电子科技大学
## 学位论文独创性(或创新性)声明

秉承学校严谨的学风和优良的科学道德，本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中做了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切的法律责任。

本人签名：_____ 日期_____

## 西安电子科技大学
## 关于论文使用授权的说明

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属西安电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。同时本人保证，毕业后结合学位论文研究课题再攥写的文章一律署名单位为西安电子科技大学。

本人签名：_____ 日期_____

导师签名：_____ 日期_____

# 摘要

随着互联网的高速发展，快速准确地对文本进行分类作为信息处理的一个重要环节，受到了人们的高度重视。文本分类处理大多是针对长文本进行的，但短文本在现实世界中也是大量存在的，并呈现出爆炸式的增长趋势。短文本一般指160 字以内的文本，其稀疏性、实时性、海量性、不规范性的特点，使传统的分类模型对短文分类缺乏一定的适用性。目前，引入外部知识来扩展短文本特征是较为热点的研究方向，如何有效地获得丰富的语义知识资源，并构建与之相适的短文本分类模型，成为当前的短文本研究的一个重要课题。

针对上述问题并参考现有研究成果，本文引入特征扩展思想，将维基百科作为外部知识库，构建特征扩展词表对短文本特征进行扩充，在传统分类模型的基础上，提出了基于维基百科的中文短文本分类模型。

本文首先在研究中文短文本特点和传统文本分类模型的基础上，指出了传统分类模型在进行短文本分类时面临的缺陷，探讨了外部知识库维基百科运用于短文本分类的优势；其次，对维基百科知识库进行语义信息挖掘，在分析维基百科语义结构的基础上构建了基于维基百科的特征扩展词表，详细研究了相关概念获取方法、概念间相关度计算方法及相关概念集合的建立过程，并运用 JWPL 工具对维基百科数据进行了结构化处理；再次，对传统分类模型从短文本预处理、文本表示等步骤进行改进，将短文本表示为概念向量，依照维基百科特征扩展词表对向量空间的概念进行了扩充，并运用支持向量机算法构建分类器。最后采用 ICTCLAS 和 LIBSVM 搭建文本分类平台，将本文提出的基于维基百科的短文本分类方法和传统的分类方法进行对比，实验结果表明本文所提出的方法较传统方法更适合短文本分类，取得了更好的分类效果。

**关键词：短文本　中文文本分类　维基百科　特征扩展**

# Abstract

With the rapid development of the Internet, fast and accurate text classification as an important part of information processing is highly concerned by people. Text classification mostly processes the long texts. But the short texts also abound in the real world, and show a trend of explosive growth. The short text generally refers to the text of less than 160 words. The traditional classification model is lack of applicability for the short texts because of their characteristics of sparse, real-time, mass and non-standard. At present, the introduction of external knowledge to extend features of the short texts is a more hotspots direction. How to get rich semantic knowledge resources and build an appropriate classification model for the short text has become an important topic.

According to the problems above and referencing to the existing research, a method of feature extension is introduced to help text classification. Wikipedia was made as an external knowledge base and feature extension vocabulary was constructed to expand features of the short texts. On the basis of the traditional classification model, a Chinese short text classification model based on the Wikipedia was put forward in the paper.

Firstly, the defects of traditional classification model applied to short texts and the advantages of using Wikipedia in the short text classification were pointed out in the paper based on the research on the characteristics of Chinese short texts and the traditional text classification model. Secondly, the semantic information was mined from Wikipedia knowledge base and the feature extension vocabulary was built on the basis of the analysis of the semantic structure of Wikipedia. The methods which include of extracting the related concepts, calculating the relevance between concepts and building the semantic related concept sets were studied in detail. Thirdly，improved the traditional classification model from steps of short text preprocessing, text representation and so on. The short text was represented as the concept of vector and expanded with new features from the feature extension vocabulary. Then, the algorithm of support vector machine was used of building a classifier. Finally, the ICTCLAS and LIBSVM were used to building the text categorization platform. The comparative experiment between the method of short text classification based on Wikipedia and the traditional method showed that the proposed method in the paper is more suitable for short texts and achieved better classification results.

**Keywords: Short texts Chinese text classification Wikipedia Feature extension**

# 目录

# 第一章 绪论

## 1.1 研究背景及意义

2012 年 7 月 19 日，中国互联网络信息中心（CNNIC）发布了《第 30 次中国互联网络发展状况统计报告》。报告指出：我国网民的互联网应用习惯出现显著变化，包括新型即时通信、微博等在内的新兴互联网应用迅速扩散[1]。2012 年上半年，即时通信用户维持较高的增速，继续保持中国网民第一大应用的领先地位；有过半数网民在使用微博，比例达到 50.9%。此外，手机短信、新闻评论、交互问答等应用已经成为被人们普遍接受的沟通渠道和交流手段。这些应用都产生了大量的文本信息数据，具有数量大、内容短的特征，我们称之为短文本。可见，以中文短文本形式存在的信息呈现出爆炸式的增长趋势，快速渗透到社会和生活的各个领域，逐渐成为人们生活中不可或缺的信息传播方式。

这种海量的非结构化短文本信息虽然开拓了人们的视野，给生活带来了丰富的资源，但其庞大的数量所带来的“信息过载”和“信息迷航”也对使用者心理造成了一定的恐慌[2]。因此，如何有效处理海量文本，尤其是中文文本，从中挖掘有价值的信息并转化为知识，已经成为当今时代的一个重要课题，吸引了来自包括图书情报学、计算机科学等学科在内的学者以及 Google、百度等商业机构的关注。文本自动分类等文本挖掘技术就是在各种信息量异常庞大、信息载体纷繁复杂的形势下，应运而生的一套关键技能，其以无法比拟的优势顺应了时代的需求，迅速蓬勃发展起来[3]。

文本自动分类是在预定义的分类体系下，根据文本的特征（词条或短语），将给定文本分配到一个或多个特定类别的过程[4]。然而短文本信息独特的特征导致其分类过程不同于传统长文本的分类方式。单条短文本一般长度都非常短，约 160 字以内，导致样本特征非常稀疏，很难准确地抽取有效的语言特征；短文本实时性很强，且语言简洁，错误拼写、不规范用语和噪音比较多，给短文本信息分类带来了更大的挑战[5]。针对这些问题，相关领域开展了较多研究，引入外部知识来弥补短文本特征稀疏的现象是目前较为热点的研究方向。如何经济有效地获得丰富的语义知识资源，使之更好的为短文本分类服务，成为当前短文本研究的一个重要课题。

Web2.0 已经把万维网带入了一个新的时代，网站提供各种工具，旨在吸引用户贡献内容，一些学者提出了“群智”的概念来描述用户在 Web2.0 时代的贡献。如果能够将用户无偿贡献的“群体智慧”融入到知识库的建设过程中，将会有效弥补专家建立知识库的局限性以及真实文本的无序性。在众多的 Web2.0 应用中，

维基百科以其开放性和知识性的定位获得了巨大的成功[6]。维基百科（Wikipedia）是目前全世界最大的人人可以编纂的多语种在线百科全书[7]，到 2012 年底，中文维基百科已积累了超过 61 万个页面，其特殊的组织结构和实时演化的特点蕴含了丰富的语义资源。因此，如何挖掘维基百科中的语义知识，如何利用这些语义知识辅助短文本分类，是本文研究的主要问题。

基于以上分析，本文提出了一种基于维基百科的中文短文本分类方法。该方法在分析了短文本特点及中文短文本分类难点的基础上，引入特征扩展的思想补充短文本特征。首先利用维基百科建立特征扩展词表；其次对传统分类模型进行改进，使之适应中文短文本分类；最后分析了分类模型关键过程的实现方法，并进行了实验验证。实验表明，基于维基百科的短文本分类方法能够提高短文本分类的效果。在理论方面，该方法对语义知识挖掘、文本表示模型、文本分类算法都有着重要研究意义，对数据挖掘、自然语言处理、机器学习等学科相关理论完善和发展将产生积极影响。在应用方面，由于短文本数据表达了人们的各种情感色彩和情感倾向，涉及政治、经济、军事、娱乐、生活等各个话题，因此短文本分类技术在多领域有广泛的应用前景。

## 1.2 国内外研究现状

### 1.2.1 短文本分类研究现状

迄今为止，文本自动分类技术经过 20 多年的发展，已能较好的解决了现实生活中的部分问题，但在短文本分类领域，国内外只进行了少量的工作，且其效果并不理想。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，有代表性的是 Mehran Sahami 等人[8]提出的使用基于 web 语义核函数的方法和 D.Metaler 等人[9]提出的基于相似性度量的方法。国内对于短文本的研究起步较晚，目前主要集中在重庆邮电大学、中国科学院等机构，重点在特征处理环节和分类算法上。常用的短文分类算法基本可分为两类：一类是基于某种规则改进分类过程；另一类是基于外部语义信息扩充短文本的信息量，从而提高分类效果。

基于规则的方法主要是针对短文本特点，在特征提取、文本表示、分类器构建等多个环节提出创新的方法。J Hynek[10]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；Zelikovitz S[11]在短文本分类中使用了潜在语义索引（LSI），在创建简化向量空间时将训练数据和未标记的测试样本进行组合，使特征空间中包含了对短文本分类有帮助的语义关联；Qiang Pu 和 Hui He[12]使用基于字的 N-gram 模型抽取了中文短文本中的组块，反映出短文本的语义结构和特征间的依赖关系；王细薇[13]提出了一种基于特征扩展的中文短文本分

类方法，利用 FP-Growth 算法挖掘出训练集特征与测试集特征之间的共现关系，用特征共现集来扩展短文本特征，增强短文本特征的表述能力；郭泗辉[14]提出一种基于连接强度扩展特征的贝叶斯网络短文本分类算法，算法在考虑了连接强度的因素后，减少了本来不相关的两个节点被归类为父子关系的错误干扰，使每个节点找到的父节点更加准确，从而使文本的分类准确度得到了提升；高金勇[15]基于迭代的 TFIDF 算法对短文本向量进行了优化；吴薇[16]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；樊兴华[17]提出了基于组合朴素贝叶斯（NB）和 K-近邻（KNN）分类器的两步中文短文本分类方法；闫瑞[18]提出一种动态组合分类器的方法，每个树节点生成一个分类器，每个类别由多个节点分类器组合成的子树表示，最终得到一个树状组合分类器结构来支持分类；

基于外部语义信息的方法主要是应用通用知识库、领域词典、搜索引擎等补充短文本中的语义信息。D Song[19]提出一种基于信息流的领域知识库，并在此基础上进行短文本分类；X Phan[20]提出一种半监督学习方法进行短文本分类，通过外部的网络数据源扩展短文本的词条信息，这既解决了词特征的稀疏性问题，也使得训练得到的分类器覆盖的话题范围更加广泛；P Ferragina[21]利用 ODP（Open Directory Project）、WebKB 等手工标注知识库计算查询词、网页片段等短文本相似度；M Sahami[8]通过谷歌搜索引擎返回的结果来统计短文本片段相似度，从而丰富文本信息；宁亚辉[22]抽取领域高频词作为特征词，借助知网从语义方面将特征词扩展为概念和义元，通过计算不同概念所包含相同义元的信息量来衡量词的相似度，从而进行分类；王盛[23]利用知网的上下位关系有效补充了短文本语义信息量。

此外，很多学者对不同互联网应用中大量出现且形式多样的短文本数据的分类方法展开了研究。如黄永文[24]将产品评论文本中的产品特征、用户观点作为语义内容，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；崔争艳[25]结合知网本体库，将关键词映射到语义概念，并用 KNN 分类算法实现微博分类；王雅蕾[26]基于信息检索思想，提出一种基于类文档排名的分类算法，适用于处理交互问答系统中的大规模问题文本；刘金岭[27]利用对同义关系词汇归并和上下位词汇聚焦以及种子词汇的确定来实现对手机短信文本空间的降维，从而对中文短信文本进行快速的舆情预测。

### 1.2.2 维基百科研究现状

维基百科自诞生以来，获得了来自世界各地各个学科领域研究人员的关注。研究维基百科的论文已经不止一次地发表在 Nature 上。笔者使用 Wikipedia（维基百科）为主题，分别在 ISI Web of Knowledge 数据库和 CNKI 中检索包括所有

年限的文献并对检索结果进行初步的统计分析。如图 1.1 所示，从论文发表的年份分布来看，与维基百科相关的研究论文增长速度是飞快的，国外维基百科的研究已成为热点，而国内的研究尚处于起步阶段。如表 1.1 所示，维基百科的研究分布的类别较广，计算机类的论文占主要地位，为 792 篇，占总文献数的 82.3%。图书情报学排第三，有 83 篇文献，其中涉及到自然语言处理的文献约有 200 篇。在自然语言处理的基础研究和相关应用中，维基百科资源可以被用作一个大规模的语料库，也可以充当一个包含了世界语义资源的知识库，主要研究集中在语义相关度计算、词义消歧、关键词语义扩展、命名实体识别、文本分类聚类等方面。

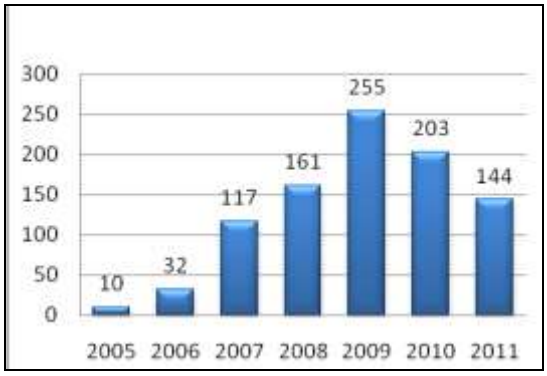



图 1.1(a) Wikipedia 在 ISI 的学术趋势

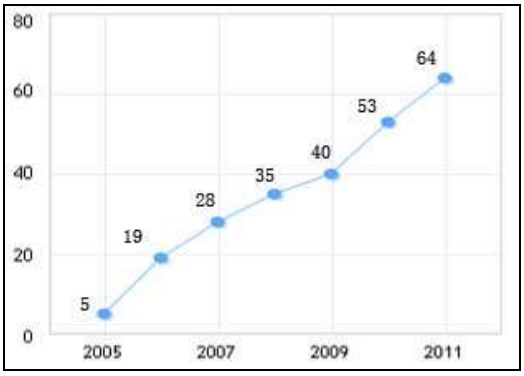



图 1.1(b) 维基百科在 CNKI 的学术趋势

表 1.1 维基百科研究论文在 ISI 的学科分布

| 学科类别 | 记录 计数 | %，共 962 |
|---|---|---|
| Computer Science | 792 | 82.33% |
| Engineering | 144 | 14.97% |
| Information Science Library Science | 83 | 8.63% |
| Business Economics | 36 | 3.74% |
| Telecommunications | 36 | 3.74% |
| Biochemistry Molecular Biology | 22 | 2.29% |
| Education Educational Research | 21 | 2.18% |
| Physics | 19 | 1.96% |
| Science Technology Other Topics | 16 | 1.66% |
| Operations Research Management Science | 14 | 1.46% |

其中在文本分类领域，维基百科往往作为外部知识库来改进文本分类模型。国外的相关研究是从 2005 年开始的，Gabrilovich 等人[28-29]通过计算文档与维基百科条目解释页面的相似度，来找到与原始文档最相似的维基百科条目，然后将这些条目与文档中原来的特征词语一起来表征文档，从而达到扩充文档表示向量的目的；Wang Pu 等[30]研究了维基百科对文本向量的语义扩展问题，将文档词向量中的每个词匹配到维基百科概念，利用同义词、上层概念、关联概念等实现向量语义相关性扩充；Peter S[31]将维基百科中的概念与文档词汇相匹配，利用维基

百科中的类别体系来决定文档的类别；Bawakid A[32]提出了一种基于质心的文本分类方法，利用维基百科抽取类别的候选概念并与文本的重要特征合并，从而丰富文本的特征向量空间。国内的苏小康[33]基于中文维基百科构建以语义标签、语义指纹表示知识单元的形式化知识库，进而利用该语义知识库对文本词条进行扩充从而提高文本分类精度；邱强[34]提出了通过关键词和维基百科知识建立分类器的新方法，通过关键词从维基百科中找出与之相关的维基百科文档，并标记为正例样本，基于抽取出的正例样本和未标记样本构造分类器。王锦[35]将文本中的词映射到维基百科的类别体系中，使用类别作为特征来对文本进行表示，提出了一种基于全局信息自学习维基百科类别的方法，最终构造基于维基百科类别为文本表示的分类系统。

### 1.2.3 研究现状分析

从国内外的研究情况可以看出，短文本分类和维基百科的研究在国内外都属于热点领域，虽然取得了一定成果，但仍然存在一定缺陷。

一方面，基于某种规则改进分类过程的方法在一定程度上提高了分类效果，但对于噪声大的短文本仍然难以获得可接受的准确率。而基于外部语义信息的分类方法虽然运用了结构化的知识库或语义词典，但由于现有领域知识库一般由专家进行编撰，只包含小范围的领域和有限的主题，词汇更新速度慢、可扩展性差，很难适应互联网用户用词新颖化、专业化的特点。

另一方面，维基百科作为“集群式”的网络知识库，相比专家编纂的语义词典，具有质量高、覆盖广、实时演化和半结构化的优点[36]，在自然语言处理领域有很多应用，但应用于文本分类的研究还很少，尤其是用于中文文本分类尚属于探索阶段。维基百科的规模庞大而且结构复杂、中文文本语法灵活、表达丰富的特点，都给中文维基百科的研究和应用带来了很大的困难。

## 1.3 本文的主要工作

本文在对研究现状进行分析的基础上，提出了一种基于维基百科的中文短文本分类方法，该方法从大规模中文网络知识库维基百科中挖掘词汇潜在的语义信息，扩展短文本的特征向量，能够在一定程度上弥补传统分类方法应用于短文本时的不足。对中文短文本分类效果的提升以及维基百科知识的挖掘都有一定的探索意义和研究价值。研究主要从以下几点展开：

(1) 对中文短文本的特点和传统文本分类模型进行了研究，指出了传统分类模型在进行短文本分类时面临的缺陷，探讨了外部知识库维基百科运用于短文本

分类的优势。

(2) 对维基百科知识库进行语义信息挖掘，在分析维基百科的语义结构的基础上构建了基于维基百科的特征扩展词表，详细研究了相关概念获取方法、概念间相关度计算方法及相关概念集合的建立过程，并运用 JWPL 工具对维基百科数据进行了结构化处理，为短文本分类奠定了基础。

(3) 引入维基百科对分类模型从短文本预处理、文本表示等步骤进行改进，将短文本表示为概念向量，依照维基百科特征扩展词表对向量空间的概念进行扩充，并运用支持向量机算法构建分类器。

(4) 为验证本文提出的短文本分类方法的可行性，采用了 ICTCLAS 和 LIBSVM 搭建文本分类平台，将本文提出的基于维基百科的短文本分类方法和传统的分类进行对比实验，结果表明本文所提出的方法较传统方法更适合短文本分类，取得了更好的分类效果。

## 1.4 论文的组织结构

本文共分为六章，文章结构及各章主要内容组织如下：

第一章 绪论。

本章介绍了短文本分类的研究背景及研究意义，分析了短文本分类和维基百科的国内外研究现状，并给出了本文的主要研究内容和本文的整体组织结构。

第二章 相关理论研究。

本章首先分析了中文短文本的含义及特点，介绍了中文短文本分类的概念，提出了传统分类模型运用于短文本分类时面临的难点，并探讨的短文本分类的应用领域。其次，对传统文本分类中的相关技术进行了探讨，包括中文文本预处理、文本表示、常用的文本分类算法以及分类结果的评价指标。最后对知识库作了介绍。

第三章 基于维基百科的特征扩展词表构建。

本章提出的特征扩展词表是为第四章短文本分类研究做准备的。首先概括了构建维基百科特征扩展词表的意义，其次详细介绍了维基百科的链接结构、类别体系、重定向以及消歧页面等语义结构。最后详细描述了扩展词表构建的关键过程。

第四章 基于维基百科的短文本分类模型。

本章首先分析了维基百科对短文本分类的作用。其次对传统文本分类模型进行了改进，提出了基于维基百科构建短文本分类模型的方案。最后详细描述了关键步骤的实现方法：在文本预处理阶段，本文将维基百科概念加入分词词典并设置了最长匹配的分词方法；在文本表示阶段，将短文本特征词匹配为维基百科概

念，建立了概念向量空间；在传统文本分类模型的基础上，增加了特征扩展的步骤，运用第三章提出的特征扩展词表对短文本概念向量进行了扩充；最后研究了SVM 分类器的构建过程。

第五章 实验设计与结果分析。

本章对第三章和第四章提出的基于维基百科的短文本分类方法进行了实验验证。首先利用 JWPL 工具将维基百科数据结构化为特征扩展词表，其次利用 ICTCLAS 和 LIBSVM 搭建实验平台，收集短文本数据集，将本文提出的基于维基百科的短文本分类方法与传统分类模型对比，实验结果证明本文提出的短文本分类方法较传统分类模型在短文本分类上有更好的效果。最后进行实验结果分析。

第六章 总结和展望。

总结了本文的工作，并指出了以后研究工作的方向。

# 第二章 相关理论研究

## 2.1 短文本分类概述

### 2.1.1 中文短文本的含义及特点

短文本是随着互联网和通讯行业新兴应用的产生和演化而催生出的一种新形式文本，一般字数较少、内容丰富、形式多样且数量庞大，目前并没有权威的机构对短文本进行科学的定义。短文本通常是指长度比较短，一般不超过 160 个字符的文本形式，普遍存在于网络文本、手机终端及文档文献中。如微博、聊天信息、新闻主题、观点评论、问题文本、手机短信、文献摘要等[24]。

短文本独特的语言特征给人们的日常生活带来了交互上的便利，却为计算机进行大规模自动文本分类带来了一定的挑战。一般来说短文本具有以下特点：

(1) 稀疏性：短文本的内容较短，通常只包含几个到十几个有实际意义的词语，难以抽取有效的特征词。

(2) 实时性：短文本更新速度快、易于扩散。如腾讯微博上关于刘翔在术后第一时间发布的“手术顺利”的报平安微博，仅隔 24 小时就有超过 200 万微博网友转播互动，送上关注和支持。

(3) 海量性：短文本大量存在于人们的生活中，由于短文本的及时更新和快速传播，使互联网中积累了海量的短文本数据，如百度知道从 2005 年 6 月上线以来，到 2011 年 5 月中旬的解决问题数已突破 1.16 亿条。这要求对于短文本的处理计算必须具有很高的速度。

(4) 不规范性：短文本表述简洁，简称、不规范用语以及网络流行用语被广泛使用，使文本噪音较大。如“天朝”、“五毛党”等网络用语，“杯具”、“河蟹”等谐音用法，“和谐社会”、“屌丝”等新词汇。

### 2.1.2 中文短文本分类的含义及难点

中文短文本是一种特殊形式的文本，其分类的原理与普通文本分类相同。根据 David 等人的定义，文本分类（Text Categorization or Text Classification）的主要任务是基于预先定义好的类别集合，对文本进行标注[37]。从数学角度来看，文本分类是一个映射过程，在给定的文本集合 D=$\{d_{1,}, d_2, ..., d_{|D|}\}$和预先确定的类别集合 C=$\{c_1, c_2, ..., c_{|C|}\}$之间存在着一个未知的理想映射 $f$，文本分类的目标是对每对$\{d_i, c_j\} \in Doc \times Cat$ 赋一个布尔值（True 或 False），即假设文档集合和类别存在一个未知的目标函数：

$$\Phi : Doc \times Cat \to \{True, False\} \quad \text{式(2-1)}$$

文本分类任务可以描述为寻找一个映射：

$$\hat{\Phi} : Doc \times Cat \to \{True, False\} \quad \text{式(2-2)}$$

最优近似逼近未知目标函数$\Phi$。在文本分类系统中，映射$\hat{\Phi}$被称为分类器或者分类模型。如果$\hat{\Phi}(d_i, c_j) = True$，表示文档$d_i$属于类别$c_j$，反之亦然[38]。

从数据挖掘的角度讲，自动分类是一个有监督的学习过程。根据给定的被人工处理过的训练文本集去发掘文本属性和类别之间的关系，然后依据这种关系，对新到来的测试文本集进行自动的类别判断。因此，文本分类可以分为训练和测试两个阶段，如图 2.1 所示。传统文本分类模型一般由数据集构建、文本预处理、文本表示、分类器分类、结果评价几个部分组成[39]。

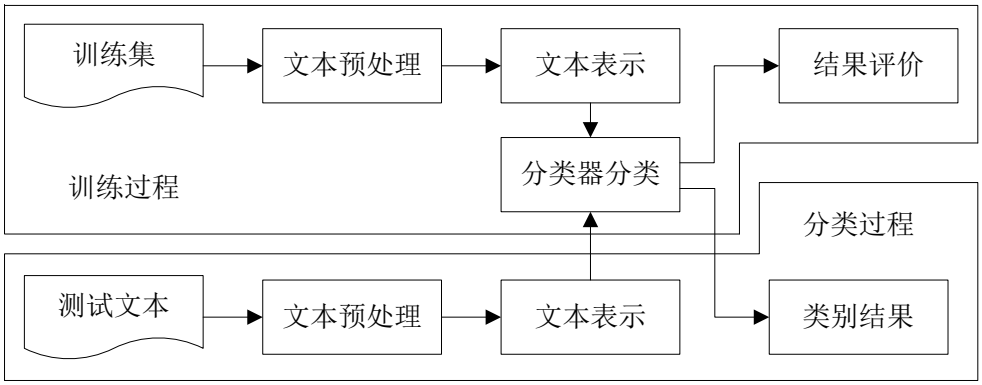



图 2.1 传统文本分类模型

相对于传统的文本分类问题，短文本分类的一般过程与之类似，但短文本的特点使文本分类面临以下难点：

(1) 短文本特征词少，用传统的基于词条的向量空间模型表示，会造成向量空间的稀疏。另外，词频、词共现频率等信息不能得到充分利用，会丢失掉了词语间潜在的语义关联关系。

(2) 短文本的不规范性，使文本中出现不规则特征词和分词词典无法识别的未登录词，这导致传统的文本预处理和文本表示方法不够准确。

(3) 短文本数据的规模巨大，在分类算法的选择上往往更倾向于非惰性的学习方法，避免造成过高的时间复杂度。

因此，短文本分类一般在预处理、文本表示、分类器的构建等环节中进行优化和改进，以提高分类效果和精度。

### 2.1.3 短文本分类的应用领域

随着短文本大量渗透到人们的生活中，短文本分类的应用也非常广泛，较为

流行的有以下几点：

(1) 舆情监控。短文本包含人们对社会各种现象的种种观点和立场，对一段时间内的大量短文本信息数据进行分类和分析，能够识别出群体的舆情趋向，达到监控的目的。

(2) 垃圾过滤。短文本是网络不良信息和社会安全问题的产生与传播的重要载体[64]，如垃圾短信、垃圾邮件、论坛的淫秽色情等不良信息。通过短文本分类技术，能够识别这些不良信息，有助于维护社会的稳定与和谐。

(3) 用户兴趣推荐。兴趣推荐研究广泛应用于各大综合网站、网络新闻、网络购物、论坛、交互问答系统及微博中。如微博中通常会根据用户的日志内容，推荐一些用户可能感兴趣的文章；交互式问答系统会根据用户提问提供相关已解答的问题，节约用户等待时间；购物网站会根据用户对商品的评论推荐相关类似商品。用户兴趣推荐的应用使得网络应用更便捷、更人性化。

(4) 热点话题跟踪。该应用是以新闻、评论等信息流为处理对象，实时监控发现热点话题报道，通过短文本分类技术，将某时间段讨论比较多的话题以某种形式组织起来，制作事件专题节目，全面的呈现给用户。针对热点话题跟踪通常从五方面展开：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[40]。

## 2.2 中文文本分类的关键技术

### 2.2.1 中文文本预处理

由于文本本身是半结构化或非结构化的数据，计算机难以理解和处理，所以需要对这些数据进行相应的预处理，将文本转化为原始特征空间中元素的序列。文本预处理一般分词、去停用词、去除标点符号等步骤。

1. 分词

对英文及类似语种来说，计算机能够较为容易的识别出每个单词，因为这些语种采用空格或标点将词隔开。与英文文本不同，中文文本的词与词之间没有明显的切分标志。而对中文文本而言，凡是涉及句法、语义等研究项目，都要以词作为基本单位[41]。因此中文文本中词的获取必须通过分词来实现。汉语自动分词算法一般可以分为四大类：基于词典匹配的分词方法、基于统计的分词方法、基于理解的分词方法以及组合分词方法[42]。

(1) 基于词典匹配的分词方法又叫做机械分词方法，它是按照一定的策略将待分词的汉字串与机器词典或词库中的词条进行匹配，若在词典中找到某个字符串，则识别出一个词[43]。根据文本扫描方向的不同，可以分为正向匹配和逆向匹

配两种；根据词条匹配不成功时采用的重新切分策略，可以分为增字词匹配和减字词匹配两种；根据长词优先匹配还是短词优先匹配，可以分为最大匹配和最小匹配两种[44]。最大匹配优先切分长度较长的词，而最小匹配优先切分长度较短的词。机械分词法实现简单，切分效率高，但会产生很多的无用词。

(2) 基于统计的分词方法[45]，即通过统计相邻字与字之间互现次数，来确定它们是否有可能构成一个词来切分文本。这种方法不需要词库，只需统计大规模语料库中相邻字共现的频率即可，具有较高的分词准确率，但是分词速度较慢。

(3) 基于理解的分词方法[46]，即在计算机理解句子的过程中加入了句法、语义分析，使计算机理解自然语言的含义，进而处理歧义现象。这种方法可以在一定程度上提高分词准确率，但语法的灵活多变性使该方法的实现存在一定困难。

(4) 组合方法，如字典与统计相结合的分词方法[47]，通常组合以上几种方法，利用各自优点，克服不足，以更好解决分词难题。

众多研究者在大量研究的基础上开发了汉语分词系统，为后续研究和应用提供了便利。早期的分词系统出现于 20 世纪 80 年代，典型的有：CDWS 分词系统[48]、汉语自动分词系统-NEWS[49]、书面汉语自动分词专家系统[50]等。现代分词系统即目前常用的分词系统，主要有：中国科学院计算所汉语词法分析系统 ICTCLAS[51]、海量智能分词系统[52]等。

2. 去停用词

停用词指的是那些出现频率很高但是对文本分类却没有太大作用的词，这些词分散了特征的权重分布，降低了与文本内容相关的词的权重，使得特征项集合不能准确地反映文本的内容。一般去除文本中的冠词、助词、代词、介词和连词等，如“的”、“即便”、“例如”等，只保留汉语句子的核心部分包括名词、动词、形容词、副词[53]。在实际应用中，去除停用词一般是根据建立停词表来完成的，这种方法简单而且有效。另外，对于一些稀有的词语，我们也可以把它们当作停用词来处理。

经过以上处理，文本被表示成一个个特征词，使文本数据的结构满足文本表示形式要求，便于计算机识别。

### 2.2.2 文本表示

由于自然语言的表示方式十分复杂，计算机不能直接对其进行识别，因此将文本转化为机器可以识别的统一的数据格式，但又不能损失过多的语义信息，即采用一种简化、统一的方式来实现对文本内容的描述，这个过程就称为文本的表示。文本表示首先要从文本中提取能够表示该文本的特征，常用的文本特征有字、词、短语、概念等。然后通过模型将特征处理为某种形式。

目前有代表性的文本表示模型主要包括布尔模型（Boolean Model）、向量空间模型（Vector Space Model）、概率模型（Probabilistic Model）和图空间模型（Graph Space Model）等等，分别从不同的角度表示文本数据[54]。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，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；概率模型是考虑特征项之间及特征项与文本之间的相互关联性，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，该模型需事先确定相关概率，参数估计难度较大，并未广泛应用；图空间模型是将文本表示是一种语义网络，其中节点表示特征，节点间的边表示特征间的关系，这种表示方法不仅反映了词汇的频率信息，也反映了文档的结构信息，是目前新出现的文本表示方法[55]。

向量空间模型是使用最广泛的一种模型，也是本文采用的文本表示方法，由 Salton 等人[56]在上世纪 70 年代提出的。该模型将一个文本空间 d 映射到一个特征词 $t_i$ 的集合$\{t_1,t_2,\ldots,t_n|t_i \in d\}$中，由于每个特征词对文本分类的重要程度不同，所以将每个特征值后加上相应的权值，则一个文本被表示成向量空间模型中的一个点，它由 N 维向量空间中的一个向量表示：

$$d_i = \left((t_{i1}, w_{i1}), (t_{i2}, w_{i2}), ... (t_{im}, w_{im})\right) \qquad \text{式(2-3)}$$

公式(2-3)中的$d_i$表示文本的向量，$t_{im}$表示特征词条，$w_{im}$表示特征词条$t_{im}$在文本$d_i$中的权重。

常用的权重计算方法有：布尔权重、词频权重、TF-IDF 权重、TFC 权重、LTC 权重和熵权重等，其中 TF-IDF 函数被广泛采用。该方法由 Salton[57]首次论证提出，其主要思想为：一个词在特定的文档中出现的频率越高，说明它对区分文档内容属性的标识能力越强；一个词在文档中出现的范围越广，说明它区分文档内容属性的标识能力越低。其经典计算公式为：

$$W_{ik} = tf_{ij} \cdot idf_j \qquad \text{式(2-4)}$$

其中$tf_{ij}$是指特征$t_j$在文档$d_i$中出现的频率；$idf_j$指出现特征$t_j$的文档的倒数。计算公式为[58]：

$$idf_j = \log(\frac{N}{n_j} + L) \qquad \text{式(2-5)}$$

L 的取值一般通过实验来确定，N 表示文档数，$n_j$指出现特征项$t_j$的文档数。考虑到文本长度对权重的影响，还应该对权重公式做归一化处理，将各项权重规范到[0,1]之间：

$$w_{tf-idf}(t_k)=\frac{tf(t_k)\cdot\log(\frac{N}{n_k}+L)}{\sqrt{\sum_{k=1}^{n}[f(t_k)\cdot\log(\frac{N}{n_k}+L)]^2}} \quad \text{式(2-6)}$$

### 2.2.3 文本分类算法

文本分类算法主要包括两种：一种是有指导的训练，用带有类标识的样本进行训练，也称为有监督分类；另一种是无指导的训练，训练文本无类别标识，也称无监督分类。本文主要研究有指导的文本自动分类算法。分类算法就是通过构造某种分类模型（也称为分类器），并以此来判断待分类文本所属的类别。常见的文本自动分类算法包括：类中心分类法（K-means）、K-近邻算法（KNN）、贝叶斯算法（Naive Bayes）、决策树（Decision Tree）、神经网络（NNS）和支持向量机算法（SVM）等。Yiming Yang 对几种常用的文本分类算法进行了分析和比较，结果表明，朴素贝叶斯、K-近邻、支持向量机是三种较好的文本分类算法[59]。

1. 朴素贝叶斯[60]

朴素贝叶斯（Naive Bayes，NB）分类器是一种基于贝叶斯分析的分类器。该算法是基于概率统计的算法，它以词在文本中出现的比率作为它属于某个类别的概率，并综合考虑诸特征词属于各个类别的概率来计算文本属于各类别的概率，最后将文本归入所属概率最大的类别。具体计算步骤如下：

(1)计算每个特征词属于各个类别的概率。在训练阶段，对于每个类别 j 就形成了向量$\left(w_{1j},...,w_{nj}\right)$，其中 $w_{ij}$ 表示第 i 个特征词属于第 j 类的概率，计算方法为：

$$\begin{aligned} w_{ij}=P\left(t_i\mid C_j\right)&=\frac{t_i\text{在}C_j\text{类别文档中出现的次数}}{\text{所有词在}C_j\text{类别文档中出现的次数}}\\ &\approx\frac{1+\sum_{k=1}^{|D|}N\left(t_i,d_k\right)}{|V|+\sum_{s=1}^{|V|}\sum_{k=1}^{|D|}N\left(t_s,d_k\right)} \end{aligned} \quad \text{式(2-7)}$$

其中$|V|$表示类 $C_j$ 中不同的特征词的个数，$|D|$表示类 $C_j$ 的训练文本的个数，$t_i$ 表示第 $i$ 个特征词。

(2)在分类阶段，根据新文本 $d_k$ 中每个特征词属于各类别的概率综合计算 $d_k$ 属于类 $C_j$ 的概率。计算公式如下：

$$P\left(C_j\mid d_k\right)=\frac{P\left(C_j\right)\prod_{i=1}^{n}P\left(t_i\mid C_j\right)^{N(t_i,d_k)}}{\sum_{r=1}^{|C|}P\left(C_r\right)\prod_{i=1}^{n}P\left(t_i\mid C_r\right)^{N(t_i,d_k)}} \quad \text{式(2-8)}$$

其中 $P(C_j)=\frac{\text{类}C_j\text{训练文档数}}{\text{总训练文档数}}$，表示 $C_j$ 类出现的概率。$\prod_{i=1}^{n}P(t_i\mid C_j)^{N(t_i,d_k)}$ 是类 $C_j$ 包含文本 $d_k$ 的概率 $P(d_i\mid C_j)$。$|C|$ 为类的总数，$n$ 为特征词总数，$N(t_i,d_k)$ 为特征词 $t_i$ 在 $d_k$ 中的词频。

(3)比较新文本属于各个类的概率，将文本分到所属概率最大的那个类别中。

2. K-近邻[61]

K-近邻（K-Nearest Neighbor，KNN）是最著名的模式识别统计学方法之一，是一种基于实例的文本分类方法。基本思想是：对给定的待分类文档 $d$，考虑训练文本集中与该待分类文档距离最近，也就是最相似的 $k$ 篇文档，根据这 $k$ 篇文档中大多数文档所属类别来判定待分类文档所归属的类别。具体步骤如下：

(1)将测试文本集合表示为特征向量空间，将待分类文档表示为文本的特征向量。

(2)计算待分类文本与训练文本集中每篇文本的相似度，相似性计算公式 $Sim(d,d_i)$ 可以采用余弦夹角、Dice 系数、Jaccard 系数、欧式（Euclid）距离、明氏（Minkovski）距离、曼氏（Manhattan）距离等度量方法。然后从训练文档集中选出与待分类文档最相似的 $k$ 个。$k$ 值是一个经验值，一般先确定一个初始值，然后根据实验测试的结果进行调整。

(3)在待分类文本的 $k$ 个近邻中，依次按下式计算对类的归属值 $p$：

$$p(d,C_j)=\sum Sim(d,d_i)\,y(d_i,C_j) \qquad \text{式(2-9)}$$

其中，$d$ 为待分类文本的特征向量，$Sim(d,d_i)$ 为 $d$ 与训练文档 $d_i$ 的相似度，$y(d_i,C_j)$ 是类作用函数，当 $d_i$ 属于类 $C_j$ 时，该函数值为 1，否则为 0。

(4)比较各类的 $p$ 值，将待分类文档分到 p 值最大的那个类别中。

3、支持向量机

支持向量机（Support Vector Machine, SVM）是建立在统计学习理论基础上发展而来的一种机器学习方法[62]，由 Vapnic 在 1995 年提出，它基于结构风险最小化原理，将原始数据集合压缩到支持向量集合，学习得到分类决策函数。支持向量机方法是本文所采用的分类算法，将在第四章中详细介绍。

### 2.2.4 分类评价指标

文本分类其本质上是一个映射过程，所以评估文本分类系统的标志是映射的准确程度。文本分类中常用的性能评估指标有：查准率 P（Precision）、查全率 R

（Recall）和 F 测试值（F-measure）[63]。

二值分类的性能评估一般使用列联表（Contingency Table），见表 2.1 所示。

表 2.1 二值分类列联表

| 原标记类别 / 新标记类别 | 真正属于该类的文档数 | 真正不属于该类的文档数 |
|---|---|---|
| 判断为属于该类的文档数 | *a* | *b* |
| 判断为不属于该类的文档数 | *c* | *d* |

1、精确率（Precision，p）

$$p = \frac{a}{a+b} \qquad \text{式(2-10)}$$

精确率是所有判断的文本中与人工分类结果吻合的文本所占的比率，p 描述了分类结果中的准确程度，即分类结果中有多少是正确的。

2.召回率（Recall，R）

$$r = \frac{a}{a+c} \qquad \text{式(2-11)}$$

召回率又称查全率，是人工分类结果应有的文本中与分类系统吻合的文本所占的比率，r 描述了正确分类的能力，即已知的文本中，有多少被正确分类。

3. F 测试值（F-measure）

对于一次测试，准确率和召回率一般是成反比的。一般而言，查准率会随着查全率的升高而降低，多数情况下需要将两者综合考虑，得到一种新的平均指标：F 测试值。计算公式如下：

$$F_\beta = \frac{(\beta^2+1)\cdot p\cdot r}{\beta^2\cdot p + r} \qquad \text{式(2-12)}$$

其中$\beta$是一个用来调节 p 和 r 在公式中权重大小的参数，即当$\beta$=1 时，查全率和查准率同等重要，此时得到转化公式为：

$$F_1 = \frac{2pr}{p+r} \qquad \text{式(2-13)}$$

$F_1$ 值越大，表示分类器的性能越好，文本分类的效果越理想。

## 2.3 知识库简介

### 2.3.1 知识库的构建

知识是抽象的、用来传达概念的一种形式。在人类活动中，人们利用文字、语言等符号作为沟通工具，利用自身的思维能力作为理解工具，产生了知识。知

识库是有序化、易利用、有组织的知识集群。可见，知识是人类沟通和思维的产物，而知识库获取这种产物的源泉。

常用的知识库大致可以分为两类[64]：一类是人工构建的语义知识库(如HowNet、WordNet)；另一类是大规模的真实文本，包括互联网上的海量文本、各种百科知识库(如维基百科等)。人工构建的语义知识库一般领域专家手工创建，其中蕴涵了细致的内容选择与定义过程，具有较高的质量和权威性。但更新和调整不易，内容具有静态性和有限性。并且人工构建语义知识库的过程需要投入大量的人力和时间，难以满足网络文本信息处理的需要。基于大规模真实文本的知识库来源于互联网上的真实文本，蕴涵丰富多彩的知识，是群体智慧的集合。随着语言的发展、语言生活的变化、社会进步而不断地更新和扩充，具有动态性和及时性。常用的人工构建的知识库列举如下：

(1) 知网

知网[65]（HowNet）由中科院董振东教授等人创建，是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与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。它以词为基本单位，每一个词可以表达为几个概念。概念通过义元来描述，目前 HowNet 存在上千个义原。义原之间存在复杂的关系，构成一个树状结构，常常作为进行语义相似度计算的基础。

(2) WordNet

WordNet 是一个大型的词汇数据库。在 WordNet 中，名词、动词、形容词和副词被组织成同义词集（Synset），每个同义词集中包含一个或多个词形，这些词形在特定的语境下是可以互换使用的。词形或者词义之间通过某些关系链接起来，这样就形成了 WordNet 中的词汇网络。在这个网络中，概念作为节点，关系作为边，这样 WordNet 的结构就可以看作是图结构。

### 2.3.2 维基百科知识库

维基百科（Wikipedia）是互联网中装载人类基础知识的百科全书，其采用群体在线合作编辑的 Wiki 机制，由非盈利的 Wikimedia 基金会维护，数据完全开放[66]。维基（Wiki）是一种“允许互联网上多个不同的用户，以浏览器作为客户端，直接编辑网页内容”的机制[36]。Wiki 系统允许用户在 Web 的基础上对 Wiki 文本进行浏览、创建、更改。Wiki 的写作者自然构成了一个社群，可以帮助我们在一个社群内共享某个领域的知识。

维基百科作为维基技术最著名的应用之一，由 Jimmy Wales 和 Larry Sanger 于 2001 年 1 月创建，在 272 种独立语言版本中，共有 6 万名以上的使用者贡献了超过 1000 万篇条目。目前，维基百科网站浏览量排名全球第 6[67]，已经成为众多

网民获取知识、查找研究资料的重要来源之一。中文维基百科于 2002 年 10 月 24 日开始运行，截至 2012 年 12 月 31 日，已拥有 612628 条条目[68]。作为 web2.0 技术与维基模式的成功典范，其独特的管理和编辑技术使维基百科成为了一个包含语义资源的知识库，具有以下优良特性：

(1)互动协作：维基百科大部分页面都可以由用户使用浏览器进行阅览、修改、创建主题及条目，通过不断地编写和修订，最终完善相应的词条与内容，使维基百科的词条具有较高的质量。文献[69]发现，尽管在微小错误（如拼写或漏词等）方面，维基百科要稍高一些，但在重大错误（如概念误解）的数量上，维基百科与大不列颠百科全书几乎相等。

(2)平等开放：来自全世界的贡献者，不论身份、背景、文化和年龄等差异，都可以自由地添加、编辑其中的页面，任何用户在维基百科网站上有同等的自由度，被要求本着以客观事实为依据的原则进行编写和修改，避免主观性的言语。可以说，维基百科是由全世界网民的群体智慧汇集而成。

(3)更新及时：与传统的百科全书相比，维基百科保证了知识的时效性，能够在第一时间补充社会科技文化的新概念、网络生活的新动态，所以维基百科具有较多新词、流行词、俚语、技术术语和新近事件等，可在很大程度上避免语义知识库的更新滞后并降低维护代价。

(4)覆盖面广：维基百科是众人协作编写的百科全书，其内容涵盖地理、历史、社会、科学及教育等各个领域。文献[70]使用美国国会图书馆 3000 篇随机文章与维基百科的随机词条进行对比，发现除了在法律和医学领域稍逊之外，其他领域维基百科几乎都很好地进行了覆盖。

可见，维基百科作为知识库在大部分领域保持相对较高的准确度和覆盖度，作为短文本分类的外部知识库，相比知网和 WordNet 而言，能够有效克服短文本的特点给文本分类带来的挑战。对维基百科知识库的处理过程将在第三章详述。

## 2.4 本章小结

本章对中文短文本分类的相关理论和技术进行了概括。首先分析了中文短文本的含义、特点以及中文短文本分类的概念，提出了传统分类模型运用于短文本分类时面临的难点。其次，对传统文本分类中的相关技术进行了探讨：在中文文本预处理阶段，详细介绍了分词方法和去停用词过程；文本表示阶段重点介绍了向量空间模型；文本分类算法主要介绍了朴素贝叶斯算法、k-近邻算法以及支持向量机算法；并列举了常用的三种分类结果的评价指标。最后探讨了常用知识库的构建方式及维基百科的特点。

# 第三章 基于维基百科的特征扩展词表构建

短文本词汇个数少并且描述信息弱，是其区别于普通文本的最大特点。针对这种现象，本文采用维基百科作为文本分类过程中的外部语义知识库，利用维基百科中数量巨大且在不断增长的概念来对短文本的词汇进行补充，并挖掘概念间的相关关系丰富短文本的语义表达。但维基百科作为用户编纂的网络知识库，与结构化的语义词典不同。所以，需要首先挖掘维基百科中的语义知识，建立起概念之间的联系，从而构建特征扩展词表，为短文本分类奠定基础。

## 3.1 构建特征扩展词表的意义

维基百科作为半结构化的语义知识库，不能定量化获取特征词之前的语义关系，因此需要对维基百科进行结构化处理，建立特征扩展词表，为短文本的特征扩展过程提供语义资源。本文所指的特征扩展可以描述为：根据现实世界中词语的关联关系，对文本某些特征词进行扩展，由某个特征词衍生为多个特征词，从而提高文本的描述能力。例如，短文本“姚明出征奥运会”，可以提取该文本的特征词{姚明，出征，奥运会}，“姚明”这个词，我们很容易根据对常识的掌握联想到“篮球”、“NBA”等词语，短文本被表示为{姚明，出征，奥运会，篮球，NBA……}。这样，文本有了更丰富的词汇量，根据我们的实验验证，这种方式有利于将文本分配到合适的类别，是避免短文本特征稀疏的有效方法。短文本特征提取过程将在第四章进行详述，本章主要研究如何构建扩展词表来模拟人类的联想思维。

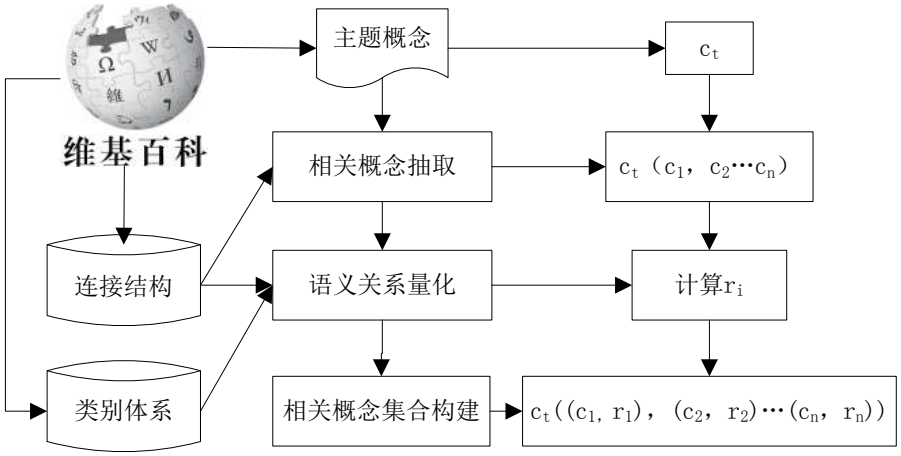



图 3.1 基于维基百科的扩展词表构建过程

扩展词表的构造来源通常有两种方式[71]：(1)机器自动构造的资源（例如未标记的测试数据和背景语料等）；(2)专家构造的资源（例如现存的各种语言知识库等）。本文采用第二种方法，将维基百科作为知识库，对链接结构和类别体系进行

处理，模仿人类联想的方式构建扩展词表，包括主题概念和相关概念两部分。对于短文本的某个特征词，通过相关概念抽取、语义关系量化，相关概念集合构建等过程，建立基于维基百科的特征扩展词表，过程如图 3.1 所示。

## 3.2 维基百科的语义结构

概念是维基百科的基本单位，也就是传统百科全书中被解释的一个对象、命名实体或事件，如“文本挖掘”、“刘德华”、“第二次世界大战”等，每一个概念的名称都是独一无二的，可通过维基百科的搜索功能直接定位至该概念的解释页面。维基百科中一个解释页面对应一个概念，如图 3.2 所示，页面的标题即为对应的概念名称。页面的第一段对该概念进行了基本的定义和解释，后续段落分别围绕概念的含义从各个角度展开具体阐述。这些解释文本的存储格式采用符合 Wiki 规范的文本源代码形式，由维基百科的贡献者合作编辑完成，自动格式转换后以网页形式发布[72]。解释文本中包含有许多指向维基百科内部或维基百科以外页面的超链接，点击可以直接跳转到相应的概念页面或者外部网页。另外，解释页中还包含了贡献者列出的该概念在维基百科中所属于的类别，每一个条目可以属于多个类别。


图 3.2 维基百科的解释页面

维基百科中存在一些特殊语义结构，如重定向、消歧页面、导航(分类索引、特色内容、新闻动态、最近更改、随机页面)，帮助(社区、询问处等)、工具箱等等其它服务。本文在构建维基百科特征扩展词表的过程中，主要运用到的语义信息包括概念解释页面中所包含的各类链接、类别间的体系结构、重定向、消歧页面等，介绍如下。

(1) 连接结构

解释页面文本设置链接的最初目的是为了方便用户在浏览时实现页面跳转，以获取更多的背景知识，对概念有更准确更全面的理解。客观上将维基百科中的解释页面有机结合起来，使页面和页面之间有了一定的关联，形成一个反映语义关系的链接图。图中每一个节点代表一个解释页面，页面和页面之间通过出链（outlink）和入链（inlink）相互连接，出链是指页面通过连接指向的目标页面，入链是指能够通过连接跳转到该页面的页面。

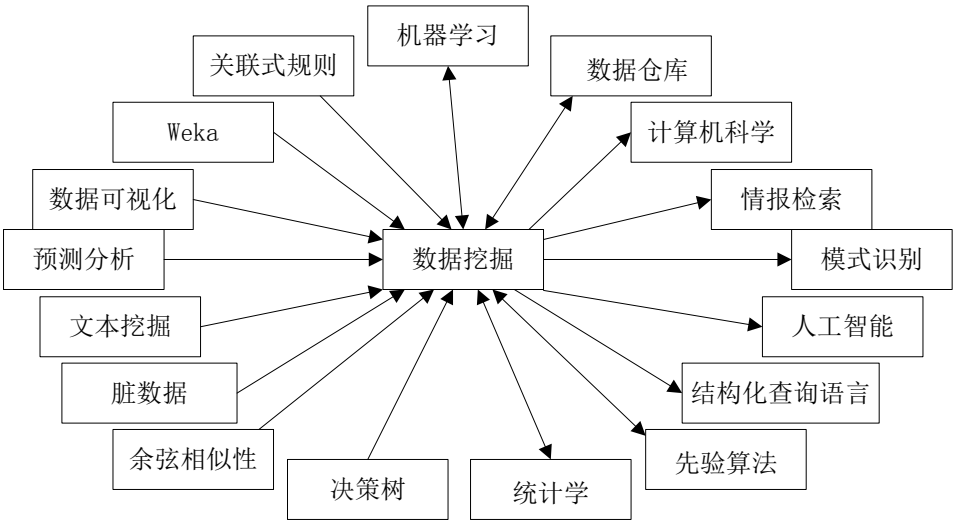



图 3.3 维基百科中的链接结构举例

如图 3.3 所示，其中页面“数据挖掘”的入链概念有：关联式规则、Weka、数据可视化、预测分析、文本挖掘、脏数据、余弦相似性、决策树……出链概念有：计算机科学、情报检索、模式识别、人工智能、结构化查询语言……既为入链又为出链的概念有：机器学习、数据仓库、统计学、先验算法。

(2) 类别体系

类别是维基百科中对概念页面信息进行组织的一种有效手段[73]。每一个概念页面通常归属于一个类别或多个类别。如图 3.2 中“文本挖掘”这个概念页面归属于“数据挖掘”、“人工智能应用”等多个类别。每个类别可以包含若干子类别，上下层类别之间不仅反映出继承的关系，也可能是实例、包含、属性等不同的语义关系。类别之间的这种关系构成一个巨大的分类体系。分类信息保存在分类页面中，如图 3.4 所示，与解释页面的结构不同的是，分类页面几乎不包含任何解释文本，只包含到所有子分类、上层分类以及相关解释页面节点的超级链接索引。


图 3.4 分类页面举例（软件工程）

(3) 重定向

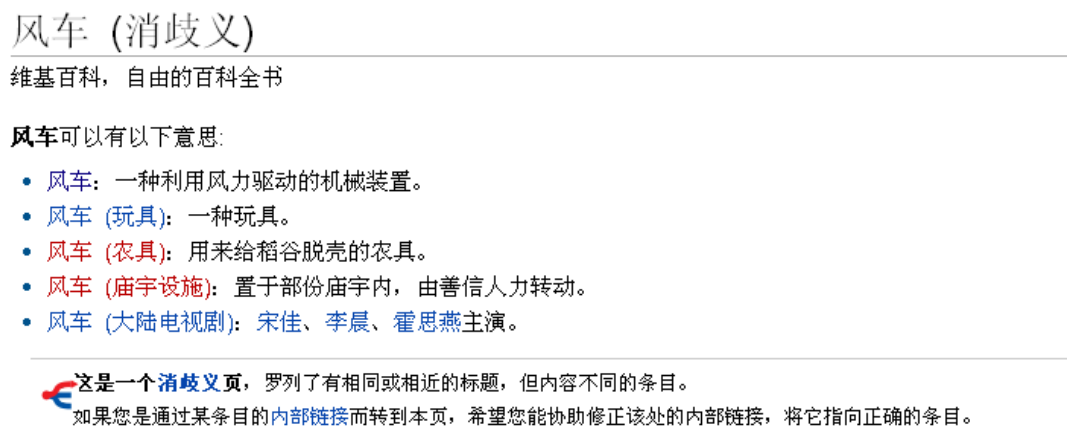
重定向页面是一种特殊结构的页面，不包含任何解释文本，仅含有重定向链接。重定向链接确保重定向页面和一个维基概念联系在一起，重定向页面标题属于目标概念的同义词，如“西红柿”被重定向到“番茄”。这种方法既避免了解释页面的重复性，又突出了二个概念之间的等价关系。同义性包括简称、缩写、错拼、俚语和译名等语言现象，如表 3.1 所示。访问维基百科站点重定向信息会触发页面跳转。

表3.1 维基百科重定向页面的原因和示例

| 重定向原因 | 示例 |
|---|---|
| 简称 | 奥运会 → 奥林匹克运动会<br>中情局 → 中央情报局 |
| 错别字 | 模版 → 模板 |
| 译名 | 铁达尼号 → 泰坦尼克号 |
| 缩写 | SARS → 严重急性呼吸综合征<br>DNA → 脱氧核糖核酸 |
| 别名、曾用名、绰号、同义词等 | 冀 → 河北省<br>七七事变 → 卢沟桥事变<br>周树人 → 鲁迅<br>统计论 → 统计学 |

(4) 消歧页面

消歧页面的作用是为了解决一词多义的问题。如图 3.5 所示的消歧义页面，词语“风车”，它既可以指生活中的一种玩具，也可以指用来给稻谷脱壳的农具或一种庙宇设施，还可以指 2011 年 9 月份播出的一部电视剧的名字等，甚至以后还会产生其他新的含义。消歧页面列举出了该歧义词的所有可能候选译词，并包含指向这些译词的解释页面的链接，对每个译词使用一句话做简单解释，为用户选择相应含义的概念提供了方便。


风车 (消歧义)
维基百科，自由的百科全书

**风车**可以有以下意思：

- 风车：一种利用风力驱动的机械装置。
- 风车 (玩具)：一种玩具。
- 风车 (农具)：用来给稻谷脱壳的农具。
- 风车 (庙宇设施)：置于部份庙宇内，由善信人力转动。
- 风车 (大陆电视剧)：宋佳、李晨、霍思燕主演。

**这是一个消歧义页**，罗列了有相同或相近的标题，但内容不同的条目。
如果您是通过某条目的内部链接而转到本页，希望您能协助修正该处的内部链接，将它指向正确的条目。

图 3.5 消歧义页面举例（风车）

重定向和消歧页面是维基百科语义挖掘需要重点关注的资源，可以用来解决中文语法现象中的同义词、多义词等现象。

## 3.3 构建特征扩展词表的关键过程

### 3.3.1 相关概念抽取

我们将维基百科中待扩展的概念称为主题概念，由主题概念扩展出的概念称为相关概念，确定某主题概念的候选相关概念需要遵循以下原则：

(1)相关概念必须在维基百科中有对应的解释页面，即该概念必须存在于维基百科中；

(2)相关概念应当与主题概念有一定语义上的关联，即它们是相关词或近义词；

(3)相关概念数量不宜过多。

结合之前对维基百科的介绍，内部链接是抽取相关概念最佳的语义资源。这是因为内部链接是维基百科贡献者根据自己对当前概念的理解以及页面解释文本中提到的其他概念与当前概念的相关性来添加的。例如图 3.1 是“文本挖掘”的解释页面，页面中有许多词语以链接的形式出现，包括“文字分析”、“信息”、“模糊识别”、“结构化数据”、“数据库”、“学科”……这些都是贡献者认为与当前概念有一定语义关联并在维基百科中存在的概念。

因此，我们将主题概念所在页面的内部链接确定为该主题概念的候选相关概念。但是，在候选相关概念中仍然存在着一些概念与主题概念并不太相关，如上述例子中的“学科”、“相关性”等概念与“文本挖掘”的关联很不明显。对于这种弱相关概念，人们可以通过思维理解能力去进行判断，但是机器是很难自动做出强弱性判断的，将所有链接的概念全部作为相关概念将大大增加扩展词表的冗

余性。

针对这种现象，本文只选取与主题概念具有双向链接关系的概念作为相关概念。如果概念 A 的页面出链和入链都包含概念 B，则称概念 A、B 具有双向链接关系。A 的解释页面利用链接引用了 B，B 页面也包含指向 A 的链接，这两个概念在解释文本中相互引用，必然存在更为直接、更为重要的语义关系或事实联系。

表 3.2 主题概念及其相关概念举例

| 主题概念 | 相关概念 |
|---|---|
| 基因 | DNA，蛋白质，染色体，体细胞，遗传，突变…… |
| 数据挖掘 | 机器学习，数据仓库，统计学，先验算法…… |
| 互联网 | 万维网，网民，TCP/IP 协议，IP 地址，电子商务，网站…… |

表 3.2 列举了几个主题概念及其相关概念的事例，可以看出具有双链接关系的概念基本上是相关联的。下一步将利用维基百科的语义结构将主题概念与相关概念间的语义关系量化，最终只保留相关度较高的概念作为扩展词表中相关概念，从而有效控制词表中相关概念的数量。

### 3.3.2 语义关系量化

扩展词表中的相关概念与主题概念的关联程度不同，对语义信息的补充能力就不同，所以要对这种关联关系进行量化，即计算主题概念 $C_t$ 与每个相关概念 $C_i$ 的相关度 $R_i$。计算相关度常用的方法包括两类：分布式方法和结构化的方法。前者是对语料库进行统计分析，以此来判断两个词的上下文相关程度；后者是在资源形成的具有联通结构的概念化体系中利用结构特点（如最短路径长度、局部网络密度、节点在层次中的深度及链接的类型等）或者信息量来计算[74]。由于维基百科是一个具有结构化特点的知识库，因此本文采用第二种相关度计算方法。

另外，需要特别说明的是，在研究相关度算法时常常会涉及相似度。语义相关性综合考查了概念间各种复杂的语义关系，包括意义上或事实上的联系，如包含关系、反义关系、功能联系关系和其他非典型关系等；相似性可以看成相关性的一个特例，如果两个概念非常相似那么它们必定相关，但相关不一定相似。如“汽车和汽油”看起来比“汽车和自行车”更加的相关，但是实际上后者的相似度更高。本文研究的概念间语义相关关系既包含了相关度，也包含了相似度。

目前比较经典的基于维基百科的语义相关度计算方法有 WikiRelate 算法（Struve M，2006）及 ESA 算法（Gabrilovich E，2007）。WikiRelate 算法[75]将 WordNet 中经典的相关度算法移植到维基百科分类体系中，在一些数据集上的效果相比 WordNet 上有所提高，但计算速度较慢，同类概念的计算结果很一般；ESA 算法[28]是一种基于维基百科的解释文本实现对文本向量进行改进的方法，由于用

到巨大的页面文本信息，ESA 算法的计算的开销较大。本文主要运用维基百科的链接结构和分类体系分别计算概念距离和类别距离，其次将这两个值进行线性组合计算概念间的相关度。

3.3.2.1 链接距离

链接距离是通过测定两个概念（主题概念 $C_t$ 与其某个相关概念 $C_i$）在维基百科中共同被添加链接的概率关系来衡量概念间的距离。本文计算链接距离的方法运用了 Milne D 等人在 2008 年提出的 WLM 算法的思想。WLM 算法[76]采用了修改了的 Google 距离方法，只是 WLM 是基于维基百科的链接而不是 Google 的检索结果。由于维基百科中的连接是贡献者人为添加的，所反映出的引用相关关系与人类对现实世界中概念之间关系的理解基本相符。因此维基百科的链接结构作为一个巨大的网络，与 Google 搜索引擎所在的互联网特性相符。

Google 距离（Normalized Google distance，NGD）[77]是在 Gongle 搜索引擎中输入词汇进行查询，利用 Google 返回的匹配记录数来计算两个概念问的语义距离。任意两个词 x 和 y 的 Google 距离计算方法为：

$$NGD(x,y)=\frac{\log(\max(f(x),f(y)))-\log f(x,y)}{\log N-\log(\min(f(x),f(y)))} \qquad \text{式(3-1)}$$

其中 f(x)表示在 Google 中搜索 x 时返回的匹配记录数；f(y)表示在 Google 中搜索 y 时返回的匹配记录数；f(x,y)表示在 Google 中搜索词组 x 并含 y 时返回的匹配记录数；N 表示 Google 能够索引到的 Web 页面的数目。如果 x 和 y 在 web 页面中从不共现，只单独出现，NGD 的值接近于无穷；如果 x 和 y 是同一个词，NGD 的值为 0。可见，NGD 是词 x 和 y 共现的对称的条件概率：假设给定一个页面含有 x（或 y），那么 NGD(x,y)就表示这个页面同时含有 y（或 x）的概率[78]。如果两个词语在同一个页面中同现的概率越大，这两个词语的 Google 距离越小。

WLM 算法正是运用了 Google 距离的思想，将链入某个概念的链接数目看成 Google 搜索这个概念时返回的页面数量，将两个概念在同一篇解释文档中都被添加链接看成这两个概念在 web 的同一个页面中出现。算法认为：同时存在链接指向概念 A、B 的概念页面越多，概念 A、B 的距离越小，相关度越高。计算两概念 $C_t$ 和 $C_i$ 的链接距离公式如下：

$$D_{link}(c_t,c_i)=\frac{\log(\max(|A|,|B|))-\log(|A\cap B|)}{\log(|W|)-\log(\min(|A|,|B|))} \qquad \text{式(3-2)}$$

其中，A 和 B 分别是维基百科中所有含有链接链向概念 $C_t$ 和 $C_i$ 的页面的集合，W 是维基百科所有解释页面的集合。符号“||”表示取集合的数量。由于单个概念的链接数量远远小于维基百科页面的总数量，所以 $D_{link}$ 的值一般在 0 和 1

之间。

3.3.2.2 类别距离

WLM 算法在英文维基百科上取得了不错的效果，但中文维基百科在规模上不如英文维基百科，主题页面之间的链接存在一定的稀疏性。因此，在中文维基百科上仅利用链接结构很难充分衡量概念的距离。因此，本文对此算法进行改进，通过计算概念所属的类别之间的距离，更准确衡量概念的相关度。

结合之前对维基百科的介绍，概念属于至少一个类别，类别之间的从属或上下位关系构成了类别体系。实际上维基百科的类别并不是严格的树型结构，而是一种接近图形的复杂结构[72]。如图 3.6 所示，“人工智能应用”分别为“人工智能”和“软件科学”的子类别，这两个类别分别被多个类别所包含。与传统分类体系和大众分类法相比，传统的叙词表存在用户参与不足、知识更新较慢的缺陷，而大众分类法存在缺乏层次结构、浏览不便的缺陷[79]，维基百科的类别体系比传统叙词表有更为灵活的分类关系，同时结合了大众分类系统中用户协作创造的优点，蕴含了更为丰富且贴近现实的语义信息。

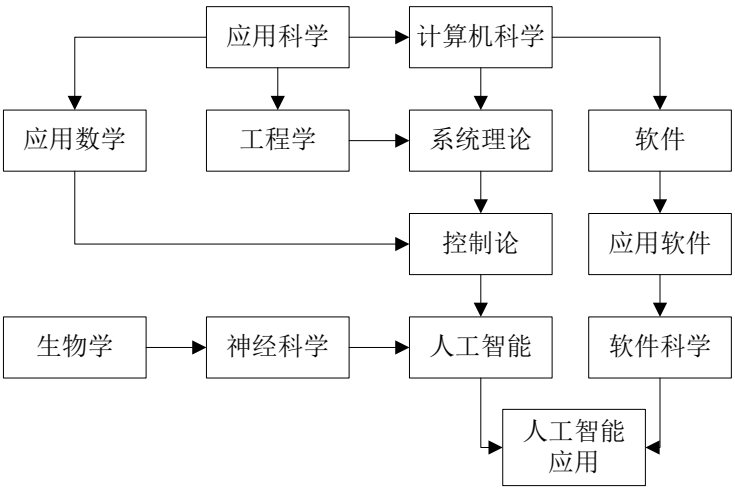



图3.6 维基百科的类别体系举例

在维基百科的类别体系中，一个分类节点可以包含任意多个上层分类节点和下层分类节点，因此两节点之间往往可以找到多条路径，其中必然存在一条最短路径 d。两节点的最短路径越小，节点的距离越近，这两个类别的相关程度也就越高。由于概念可能属于多个类别，那么两个概念间（主题概 $C_t$ 与其某个相关概念 $C_i$）就可能存在多种分类关系的组合，也就可能存在多个最短路径。我们将其中最小的最短路径值作为两概念之间的类别距离，则概念 $C_t$ 与 $C_i$ 之间的类别距离计算公式表示为：

$$D_{cat}(c_t, c_i) = \log(\min(d_{ti}) + 1) \qquad \text{式(3-3)}$$

其中，$d_{ti}$ 是两概念所属类别之间的最短距离，取 log 值是为了使 $d_{ti}$ 变化幅度

平均化，抑制类别距离与链接距离之间过大的差异。当两个概念属于同一类别时，它们的类别距离为 0。

3.3.2.3 相关度计算方法

对于主题概念 $C_t$ 与其某个相关概念 $C_i$，它们之间的概念距离形式上表现为链接距离 $D_{link}$ 和类别距离 $D_{cat}$ 的线性组合，表示为：

$$D(c_t,c_i)=\alpha D_{link}(c_t,c_i)+(1-\alpha)D_{cat}(c_t,c_i) \quad \text{式(3-4)}$$

其中 α（0≤α≤1）为调节参数。公式综合考虑了维基百科链接结构和类别体系中蕴含的概念之间的关系，能够对概念的相关度有较全面的衡量。由于概念与其本身的距离为 0，概念距离越大，它们之间的相关关系越不明显。当概念距离为 0 时，相关度为 1；语义距离为无穷大时，相关度为 0。因此，将概念之间的相关度计算公式定义为：

$$R(c_t,c_i)=\frac{1}{D(c_t,c_i)+1} \quad \text{式(3-5)}$$

3.3.3 相关概念集合构建

经过相关概念抽取、语义关系量化后，可以为每个主题概念构建形如 $C_t((C_1,R_1),(C_2,R_2),\ldots,(C_n,R_n))$的相关概念集合，其中 $C_i$ 是与主题概念 $C_t$ 具有双向链接关系的相关概念，$R_i$ 是集合中第 i 个相关概念与主题概念的相关度，由公式(3-5)得到。这样，对于短文本中的每个可以与维基百科相匹配的特征词，都可以构建它的相关概念集合。但是，集合中会出现一些相关概念，其相关度很小，可以认为它与主题概念的相关关系不明显；另外，集合的过于冗长会大大增加后续文本分类时特征空间的维度，造成维度灾难。因此，我们设立阈值 u=0.5，将相关度 $R_i<u$ 的相关概念去掉，只保留相关关系相对较强的概念在集合中。

综上各步骤，可以对每篇短文本建立扩展词表，词表中包括特征词在维基百科中的主题概念及其相关概念集合。特征扩展词表举例如表 3.3 所示。

表 3.3 短文本扩展词表举例

| 主题概念 | 相关概念 |
|---|---|
| 基因 | （DNA，0.93），（蛋白质，0.79），（染色体，0.61），（体细胞，0.55） |
| 数据挖掘 | （机器学习，0.93），（数据仓库，0.72），（统计学，0.58） |
| 互联网 | （万维网，0.87），（网站，0.81），（IP 地址，0.65），（电子商务，0.57） |

## 3.4 本章小结

为实现短文本的特征扩展，本章将维基百科结构化为特征扩展词表。本章首先概构建维基百科特征扩展词表的意义，其次详细介绍了维基百科的链接结构、类别体系、重定向以及消歧页面等语义结构。最后详细描述了扩展词表构建的关键过程：利用维基百科中概念间的双链接关系确定特征词的相关概念，并基于概念间的链接距离和类别距离计算主题概念与其相关概念的相关度，从而建立每个概念的相关概念集合。

# 第四章 基于维基百科的短文本分类模型

由于短文本固有的特点，传统的文本分类方法难以适用。本章对传统的文本分类模型加以改进，提出了一种基于维基百科的短文本分类模型。模型利用第三章构建的特征扩展词表获取短文本中词汇之间的关系，对原有的特征词进行扩展，采用维基百科概念作为特征项构造概念向量空间，并运用 SVM 分类器实现短文本分类。

## 4.1 维基百科对短文本分类的作用

和英语语言不同，汉语是一种表意型语言，注重词语的意思，而忽略词语的形式，句子结构复杂，用词灵活多变，存在着大量的同义、多义以及上下文依赖现象，使中文文本中蕴含着比其他语言更多的信息。目前最常用的文本表示方法——向量空间模型将文本向量利用词的权值来表示，具有简单易用、效率高、意思表达直观等优点。但向量空间模型假设词汇间是相互独立、互不影响的，从而忽略了词汇间的语义相关性，无法从语义层次表示文本，当文本中的多义和同义词很多时，相似度的误差将会很大。另外，短文本除了存在这些问题以外，又由于其本身的稀疏性、实时性、海量性、不规范性，给文本分类带来了如 2.1.2 小节分析的难点。

针对这些问题，本文引入了特征扩展的思想，利用维基百科建立扩展词表，对文本分类模型进行改进，具有以下作用：

(1) 特征扩展。

基于维基百科的扩展词表将某一主题概念扩展为多个相关概念，如果将短文本中的特征词匹配为维基百科中的概念，将相关概念作为与该特征词语义相关联的扩展特征词加入短文本的向量中，短文本的语义表达将更加丰富。

(2) 词义消歧。

针对自然语言中出现的多义词现象，维基百科有专门的消除歧义页面，对应于每个多义词，维基百科为了尽可能满足用户的不同理解，提供了不同的含义页面。依赖于维基百科这种结构，我们在构建扩展词表的时候，对具有多个含义的主题概念建立了多个相关概念集合。那么当对特征词进行扩充的时候，如果发现该词匹配的主题概念存在多个相关概念集合，我们可以考量特征词的上下文信息选择正确的相关概念进行扩充，这样就很好的解决了多义词消歧的问题。

(3) 同义词识别。

和多义词类似，维基百科同样有专门处理同义词的语义结构——重定向页面。

在维基百科中，每个概念都只被一篇文章来描述，所有和这个概念意义相同而表述不同的概念将会被重定向到这个页面上来。在短文本特征词与维基百科相匹配的过程中，将同义词匹配为维基百科中唯一对应的概念，从而用概念向量空间替代原先的基于词汇的向量空间，消除同义词影响的同时降低向量空间维度。

(4) 潜在语义信息挖掘。

传统的文本表示模型完全就是以词袋模式（Big of Word, BOW）对文本信息进行表示，没有考虑词汇间的相关关系。而基于维基百科的短文本分类过程，在进行文本表示时，不仅考虑了原始向量空间中特征词的权重，还考虑的扩展特征与原特征词的相关关系，从而达到挖掘潜在语义信息的目的。

(5) 流行词识别。

维基百科作为开放式的在线百科全书，采用群体在线合作编辑的 Wiki 机制，具有较广泛的词汇覆盖率和较快的词条更新速度。基于维基百科实时更新的特点，能够部分识别短文本中新颖的网络词汇、流行用语、热门事件等，提高文本预处理的质量。

## 4.2 短文本分类模型

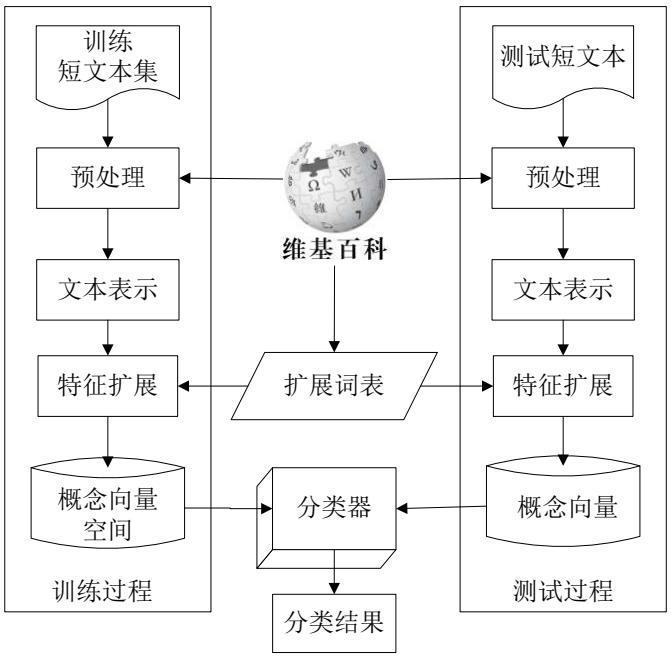
本文提出的短文本分类过程与传统的文本分类相比，区别主要在于：

(1) 传统的文本分类过程一般将文本表示成基于词汇的向量（BOW），而本文提出的文本分类方法将文本中的特征匹配为维基百科的概念，将短文本表示为基于概念的向量。

(2) 传统文本即长文本在分类的过程中，往往由于向量空间的高维和稀疏，需要采取相应方法降低向量空间的维度。而由于短文本的特征少，固需要对文本向量进行特征扩展。

(3) 为实现概念向量空间的建立和文本的特征扩展，本文提出的短文本分类算法对文本预处理步骤进行改进，使维基百科知识库能够更好的为短文本分类服务。

基于以上三个特点，本文设计的基于维基百科的短文本分类模型如图 4.1 所示。与传统长文本类似的是，短文本分类包括训练和测试两个过程。首先对训练短文本集进行处理，经过文本预处理、文本表示、特征扩展，构建概念向量空间；其次对测试文本进行相同的处理，将测试文本表示为概念向量；最后选择合适的分类算法构建分类器，输出测试文本所属的类别。




4.1 基于维基百科的短文本分类模型

## 4.3 短文本分类关键步骤设计

### 4.3.1 预处理过程

短文本预处理是分类模型的第一步，是对原始文档（txt、html、xml 等格式）的输入，进行相应的操作，包括中文分词、去除停用词、去除标点符号等，从而提取文档中的特征词。

中文文本分类的分词过程一般采用分词软件，将得到的词条作为文本特征提取的基础。一方面由于自然语言中不断有新词涌现，已有的分词软件并不能准确地识别出文本中所有的新词[34]，也不能对专业的领域词汇有很好的覆盖；另一方面，为了便于利用维基百科进行相关特征扩展时能够准确识别待扩展的主题概念，分词得到的特征词应该与维基百科中的概念相匹配。因此在进行切分处理的时候我们需要对分词软件进行改进，将维基百科所有概念词条加入分词词典，并采用最大匹配切分法。分词词典采用首字索引，同首字词条按序排列的词表结构方式进行组织。词表的每个地址空间包含两部分，一个存放词条，另外一个用来存储对应的词长，词长为词的字节数。同首词条首先按词长由大到小排列，这种排序方式满足最大匹配切分法的要求，保证优先切分出长词[80]。分词词典的结构举例如图 4.2 所示。

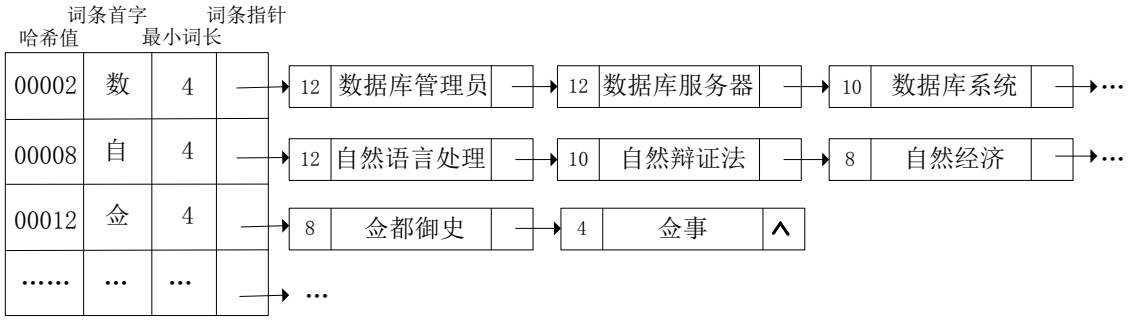



图 4.2 分词词典结构

最大匹配切分法是优先按字典中最长词的字符数对文本进行匹配的方法，该方法能够使分词结果尽量和维基百科的概念相符合。例如，“机器学习”这个词，原始的分词系统会分为“机器”和“学习”两个词，显然丢掉了原始的含义，而改进后的分词方法能够使词语具有更加精确的语义。本文采用正向匹配方法，过程可以描述为：将待切分的句子设为字串 S，词串 W 为最终的结果，词典中词的最大长度为 L 个字符，则用被处理文档里的当前字串中的前 L 个字符作为匹配字段查找字典。若字典中存在这个词，则匹配成功，将该词加入 W；如果词典中找不到这个词，则匹配失败，将 L-1 这个字串重新进行匹配，如此进行下去，直到匹配成功或剩余字串的长度为零为止。分词的过程如图 4.3 所示。

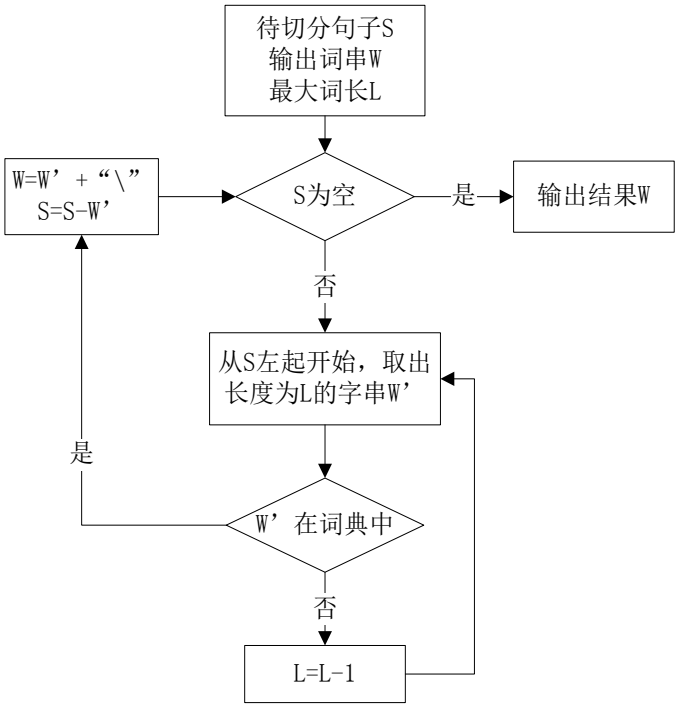



图 4.3 正向最大匹配分词方法流程图

对短文本进行分词后，需要消除文本中存在的大量高频但无意义的词语，如“什么”、“是”等，常用的方法是借助中文停用词表过滤这些噪音干扰。考虑到名词所携带的信息量较大，动词次之，而形容词和副词再次之的特性[53]，本文

只保留名词和动词作为特征词。

### 4.3.2 文本表示过程

由于文本属于非结构化的数据，不能直接被分类算法所使用，所以在对文本进行特征描述前，首先要建立文本表示模型。在预处理阶段，短文本被表示成了特征词集合。在文本表示阶段，将短文本 d 表示成由一个二元组 $d=((t_1,w_1),(t_2,w_2),\ldots,(t_n,w_n))$构成的向量，其中 $t_i$ 表示第 i 个特征词，$w_i$ 表示这个词的权重，n 表示经过预处理后短文本中特征词的个数。词的权重采用 tf-idf 的计算方法，详见公式(2-6)。

其次，将基于特征词的短文本向量转化为基于维基百科概念的向量表示。结合之前对维基百科的研究，概念是维基百科的基本单位，每个概念作为标题被一篇文章来解释描述。这种描述过程是由互联网用户完成的，在编写过程中实际上就包含到了人们对知识的感知，会考虑某个概念是否具备存在的意义，某个概念有没有同义词，某个概念是否会产生歧义等等。这样，概念合并了同义词、定义了近义词，使得特征词的选择更贴近语义。因此，我们利用维基百科的这种集体智能来构建短文本的语义表示模型。这样做的好处在于：一方面，将短文本特征词用维基百科的主题概念代替，才能利用扩展词表对短文本特征进行补充；另一方面，转换的过程就是消除同义词和多义词影响的过程，能够提高向量对文本表示的能力。

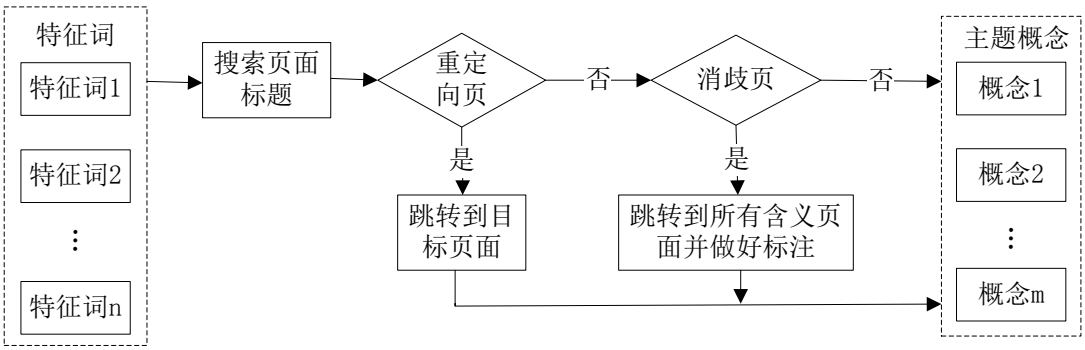



图 4.4 特征-概念的匹配过程

如图 4.4 描述了特征-概念的匹配过程。对于给定的某个特征词 T，在维基百科中搜索以该特征词为标题的页面。如果存在重定向，就顺着重定向一直找到最终的对应页面[11]。如果找到的对应的页面是一个歧义页面，即特征词 t 存在多个含义，分别找到所有候选含义页面，并用括号将类别标注，便于短文本分类时进行词义消歧。最终提取特征词 T 对应页面的概念 $C_t$ 作为特征词对应的主题概念。如表 4.1 的示例，列举了不同情况下特征词与概念的匹配情况。

表 4.1 特征-概念的匹配举例

| 特征词 | 主题概念 |
| --- | --- |
| 数据挖掘 | 数据挖掘 |
| 西红柿 | 番茄 |
| 风车 | 风车（农具）<br>风车（玩具）<br>风车（庙宇设施）<br>…… |

表 4.1 中，维基百科中本身存在“数据挖掘”这个概念页面，所以特征词“数据挖掘”被直接匹配为该概念；“西红柿”在维基百科中被重定向概念“番茄”，所以匹配结果为重定向的目标概念；“风车”是个多义词，匹配结果为多个含义的概念，并将不同含义用括号标注。

以上描述了单个特征词 T 在维基百科中的匹配过程，如果将短文本 d 中所有的特征词依次进行匹配会出现以下三种情况：

(1) 匹配成功，特征词只对应唯一主题概念，则用该主题概念 $Ct_i$ 代替特征词 $t_i$，其在概念向量中的权重 $Wc_i$ 仍为 $w_i$。如果有多个特征词同时对应一个主题概念，则将这些特征词的权重合并作为概念的权重。例如短文本特征向量中存在特征词（番茄，0.32）、（西红柿，0.21），经过匹配得到的概念和权重为（番茄，0.53），这是由于在维基百科中“西红柿”被重定向到“番茄”，这两个词表示同一含义，权重为这两个词各自权重的和。

(2) 匹配成功，特征词对应多个主题概念，即这个词有多个含义，需要进行词义消歧的处理。词义消歧的方法描述为：在维基百科中找到这个特征词的消歧义页面，将所有候选含义的页面与该特征词所在的特征向量进行语义比较，然后选择其中一个在语义比较过程中效果最好的页面作为当前特征词对应的维基百科的概念。语义比较采取频率统计的方法，计算该特征词 T 所在的特征向量中除这个词以外的其他特征词在不同含义页面中出现的频率，并累加得到每个页面的频率和 $F(t, p_k)$，计算公式如下：

$$F(t, p_k) = \sum_{t_j \in c_i \& t_j \neq t} T(t_j, p_k) \qquad \text{式(4-1)}$$

其中，$T(t_j, p_k)$ 表示特征向量中除 T 外的某个特征词 $t_j$ 在页面 $p_k$ 中出现的次数，$p_k$ 为 T 的第 k 个候选含义的描述页面。将 F 值最大的页面对应的概念作为该特征词在该短文本中的真实含义，将扩展词表中带语义标签的主题概念代替该特征词。

举例说明，对于短文本“风车将风能作为动力，不需要燃料”，经过预处理短文本中的特征词包括：“风车”、“风能”、“动力”、“燃料”，其中，“风车”是个多

义词，可能表示玩具、农具、一部电视剧的名字等。分别计算“风能”、“动力”、“燃料”这三个词在“风车”不同含义对应页面中出现的次数。结果如表 4.2 所示，可见 “风车”在这段文本中的真实含义为“风车（农具）”，则用该概念代替“风车”加入短文本概念向量。

表 4.2 “风车”消岐义方法举例

| 频次 概念 / 特征 | 风车（农具） | 风车（玩具） | 风车（庙宇设施） | 风车（电视剧） |
|---|---|---|---|---|
| 风能 | 1 | 1 | 0 | 0 |
| 动力 | 2 | 0 | 0 | 0 |
| 燃料 | 0 | 0 | 0 | 0 |
| 频次总和 | 3 | 1 | 0 | 0 |

(3) 匹配失败，则表示特征词不能对应到维基百科中任何概念，对于这样的特征词将其作为未登录概念来对待，保留到概念空间中，但不做扩展处理。

通过概念匹配，文本被表示成概念向量，形如 $d=((Ct_1, Wc_1),( Ct_2, Wc_2),\ldots, ( Ct_n, Wc_n),(t_1,w_1), (t_2,w_2),\ldots,(t_m,w_m))$的形式，其中 $Ct_i$ 为扩展词表中的主题概念，$Wc_1$ 为其权重, $t_i$ 为匹配失败的特征词，n 为向量中概念的个数，m 为未登录概念个数。

### 4.3.3 特征扩展过程

特征扩展是短文本分类过程的核心环节，是指利用扩展词表对短文本信息进行扩充，这里我们所说的文本不仅包含测试文本，还包含训练文本。对文本信息进行扩充的目的在于弥补短文本词汇个数少、描述信息弱的缺陷，使得文本信息尽可能逼近它所描述的真实语义信息。

在文本表示阶段，短文本被表示为概念向量，在本节将根据扩展词表对概念向量进行改进，加入新的概念并计算权重，特征扩展过程如图 4.5 所示，概括为以下四步：

Step1：判断是否为未登录概念。对于概念向量中的第 i 个概念 $Ct_i$，判断其是否存在于特征扩展词表的主题概念中，若为未登录概念则保留在概念向量中不做扩展，否则进行第二步。

Step2：加入扩展概念。在扩展词表中找到概念向量中第 i 个概念 $Ct_i$，根据语义相关概念集合 $C_t((C_1,R_1),(C_2,R_2),\ldots,(C_n,R_n))$，将相关概念 $C_{ij}$ 加入概念向量空间。

Step3：计算扩展特征概念的权重。权重代表了特征词表征文本的能力，而相关度代表了概念对另一个概念的解释能力，所以扩展特征词的权重既要考虑其主题概念在原文中的重要性，又要考虑其与主题概念的相关关系。扩展特征词的权

重值计算如下：

$$W_{ij} = Wc_i \cdot R_{ij} \qquad \text{式(4-2)}$$

其中，$Wc_i$ 为被扩展概念 $Ct_i$ 的权重，$R_{ij}$ 为 $Ct_i$ 的相关概念集合中第 j 个概念 $C_j$ 的相关度。

经过以上步骤，短文本被表示为了扩展后的概念向量，该向量由三部分组成：(1)原始概念及其权重 $((Ct_1, Wc_1),( Ct_2, Wc_2),\ldots, ( Ct_n, Wc_n))$；(2)扩展概念及其权重 $(C_{11},W_{11}),\ldots,(C_{1k},W_{1k}),\ldots,(C_{nk},W_{nk}))$，其中 k 为相关概念集合中相关概念的个数；(3)未登录概念及其权重$((t_1,w_1), (t_2,w_2),\ldots,(t_m,w_m))$。

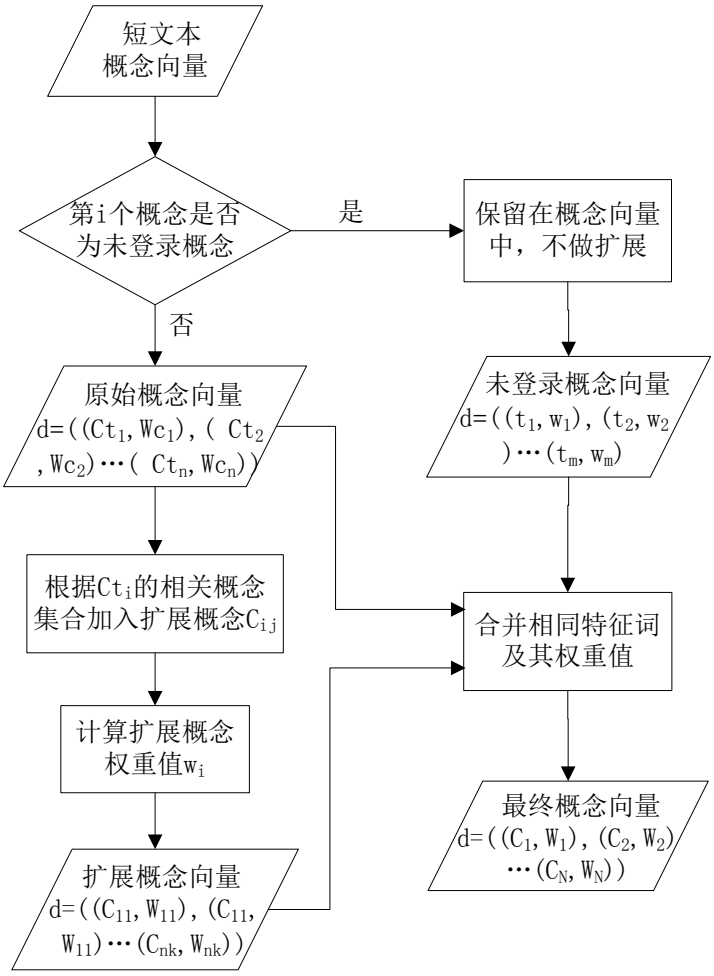



图 4.5 特征扩展过程

Step4：合并相同概念。由于特征词之间往往有丰富的语义联系，可能会出现扩展概念与原始概念相同或者扩展概念之间有重复的现象，这时需要合并相同的概念，并将它们的权重相加求和。对于训练短文本集，处理的对象是由训练集构成的概念向量空间；对于测试文本，处理的对象是单个短文本概念向量。最终语义特征向量表示为 $d=((C_1,W_1),(C_2,W_2)\ldots(C_N,W_N))$，N 为概念向量中合并后的概念

数量（包括未登陆概念）。

### 4.3.4 分类器构造过程

本文第二章介绍了三种常用的分类方法：贝叶斯、K-近邻、支持向量机，这三种方法的优缺点评述如下：

(1) 贝叶斯（NB）方法比较简单，优点在于分类速度快。但贝叶斯要求文本的特征词之间相互独立，这样的条件在实际文本中一般很难满足，因此该方法往往在效果上难以达到理论上的最大值。

(2) K-近邻（KNN）方法的优点在于算法实现简单、效果好。但 KNN 的以下 3 个不足之处使得它的实际应用受到了很大限制：①文本相似度计算量大；②分类性能受单个训练样本影响大；③KNN 是一种惰性学习方法，在分类任务执行前并没有预先建立模型[41]。

(3) 支持向量机（SVM）是目前分类性能最好的分类器之一，它专门针对有限样本，其目标是得到现有信息下的最优解而不仅仅是样本数趋于无穷大时的最优值[81]；算法将实际问题通过非线性变换转换到高维的特征空间，在高维空间中构造线性判别函数来实现原空间中的非线性判别函数，这种特殊性能保证分类器有较好的推广能力，其算法复杂度与样本维数无关。

本文构造的分类器针对短文本，由于短文本字数少，为提高分类效果，需要加大训练样本容量，因此 KNN 算法并不适合短文本分类，而贝叶斯算法忽略了特征词之间的关联关系，因此本文采用支持向量机作为文本分类算法。

SVM 是 AT&T 贝尔实验室的 Vladimir N. Vapnik 教授提出的一类基于统计学习理论（SLT）的新型机器学习方法。基本思想可概括为：首先通过非线性变换将输入空间变换到一个高维空间，然后在高维空间中求最优线性超平面，以该平面对两类问题进行分割。所谓最优超平面就是要求该平面不但能将两类别正确分开，而且要求其分类间隔最大。统计学习理论通过结构风险最小化原则，把最优超平面的构造转化为二次优化问题，从而求得全局最优解，这是支持向量机的核心内容[82]。

图 4.5 为样本只有两类时的分类例子，图中空心圈和实心圈分别代表两类样本，H 为最优超平面，$H_1$、$H_2$ 分别为通过各类中距离超平面最近的样本且平行于超平面的直线，它们之间的距离称为分类间隔（margin）。最优超平面方程为：

$$w \cdot x + b = 0 \qquad \text{式(4-3)}$$

其中 w 是超平面的法向量，b 是常数项。SVM 的主要思想是使分类间隔最大化，所以求最优分类面等价于求最大间隔。$H_1$、$H_2$ 上存在特殊点$(x_i,y_i)$，称为支持向量，如公式（4-4）所示。

$$w\cdot x_i+b=-1, y_i=-1 \quad or \quad w\cdot x_i+b=1, y_i=+1 \qquad 式(4\text{-}4)$$

这些支持向量是最靠近分类面的数据点，这样的数据点是最难分类的，因此它们和分类面的最优位置直接相关。

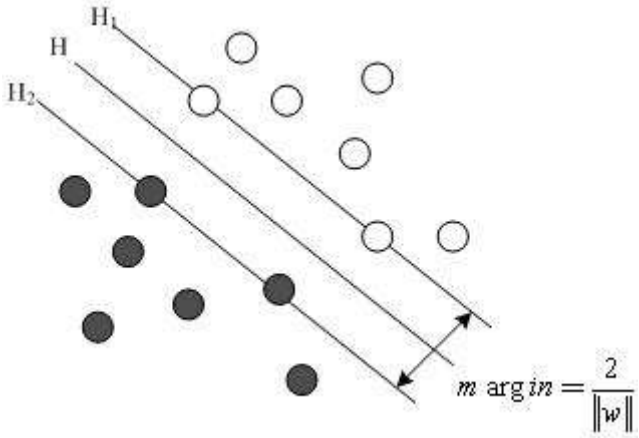



图 4.6 支持向量机算法的最优分类面

此时，分类间隔等于 $2/\|w\|$，使间隔最大等于使 $\|w\|^2$ 最小。

利用拉格朗日（Lagrange）优化方法可以把上述最优分类面问题转化为对偶问题：

$$\max_a Q(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_i a_j y_i y_j x_i^T x_j$$

$$\text{s.t} \quad \sum_{i=1}^{n}y_i a_i=0 \ （a_i\geq 0，\ i=1,...,n） \qquad 式(4\text{-}5)$$

其中，$a_i$ 为与每个样本对应的 Lagrange 乘子。这是一个在等式约束和不等式约束下的凸二次函数寻优的问题，存在惟一解，且解中将只有一部分（通常是少部分）$a_i$ 不为零，对应的样本就是支持向量。求解上述问题后得到的最优分类函数是：

$$f(x)=\text{sgn}(w\cdot x+b)=\text{sgn}(\sum_{i=1}^{n}a_i^* y_i x_i x+b^*) \qquad 式(4\text{-}6)$$

其中 x 代入测试集中的样本，即可对测试集进行分类。$b^*$ 是分类阈值，可以用任意一个支持向量求得，或通过两类中任意一对支持向量取中值求得。

最优超平面是在线性可分的前提下讨论的，如果数据有噪声，就会出现线性不可分的情况，可以增加一个松弛项 $\xi\geq 0$，成为：

$$y_i[((w\cdot x)+b)-1]+\xi_i\geq 0，\ i=1,...n \qquad 式(4\text{-}7)$$

将目标改为求解：

$$（w,\xi)=\frac{1}{2}\|w\|^2+C(\sum_{i=1}^{n}\xi_i) \qquad 式(4\text{-}8)$$

其中，$C>0$是一个常数，它控制对错分样本的惩罚程度。类似地，广义最优分类面的对偶问题与线性可分情况下几乎完全相同，只需要将约束条件变为：$0\le a_i\le C$，$i=1,...n$。通过把原问题转化为对偶问题，计算的复杂度不再取决于空间维数，而取决于样本数，尤其是样本中的支持向量数，这可有效地解决高维向量的分类问题。

SVM 中文文本分类的具体算法步骤描述为：

(1) 选择合适的核函数$K(x,x_i)$及核参数，作为高维特征空间向低维空间的映射函数。

(2) 从$k=0$开始，构造 k 个两类分类器的样本集合$L^k$：

$$(x_i^k,y_i^k),i=1,2,\ldots,n^k,x_i\in R^m,y_i\in\{+1,-1\},n^k=\sum_{j=k}^{n}n_j \qquad 式(4\text{-}9)$$

(3) 最大化下式以求解拉格朗日系数$a_i^k$，

$$Q(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jy_iy_jK(x_i,x_j) \qquad 式(4\text{-}10)$$

其中，$0\le a_i\le C$，$\sum_i a_iy_i=0$，$K(x_i,x_j)$是属性空间向量内积形式有：

$$K(x_i,x_j)=k(x_i)*k(x_j) \qquad 式(4\text{-}11)$$

(4) 得到支持向量$SV^k$，从而建立最优决策超平面：

$$f^k(x)=\mathrm{sgn}(\sum_{i=1}^{n}a_i^ky_i^kk(x_i^k,x)+b^k) \qquad 式(4\text{-}12)$$

(5) 建立 N-1 个支持向量机$f^k(x)$，k=1,2,3…,N-1。

(6) 根据式(4-12)计算输入特征数据 x 在每一个向量机$f^k(x)$，k=1,2,3…,N-1 中的决策输出值，并按下式作出分类决策：

$$f(x)\begin{cases}k & f^1(x)=f^2(x)=\ldots=f^{k-1}(x)=1,f^k(x)=-1\\ N & f^1(x)=f^2(x)=\ldots=f^{N-1}(x)=1\end{cases} \qquad 式(4\text{-}13)$$

(7) 输入待分类样本，利用分类决策函数得到分类结果。

## 4.4 本章小结

本章详细描述了基于维基百科的中文短文本分类方法，在传统文本分类模型的基础上，提出了改进后的短文本分类模型。本章首先分析了维基百科对短文本

分类在特征扩展、词义消歧、同义词识别、潜在语义信息挖掘、流行词识别等方面的作用。其次提出了基于维基百科的短文本分类模型，并对比了其与传统分类模型的区别。最后详细描述了关键步骤的实现方法：在文本预处理阶段，本文将维基百科概念加入分词词典并设置了最长匹配的分词方法；在文本表示阶段，将短文本特征词匹配为维基百科概念，建立了概念向量空间；在传统文本分类模型的基础上，增加了特征扩展的步骤，运用第三章提出的特征扩展词表对短文本概念向量进行了补充；最后研究了 SVM 分类器的构建过程。

# 第五章 实验设计与结果分析

## 5.1 维基百科处理

### 5.1.1 数据预处理

与其他百科网站不同，维基百科提供所有完整内容的电子档案给有兴趣的使用者，一般每隔十天左右就对当前版本的数据做一个备份。维基百科所有的数据都可以从维基媒体基金会提供的页面下载，数据包括：所有条目的历史修改记录、概念的解释页面的全部内容、概念页面的内部链接、分类的基本信息、概念到所属分类的链接、重定向页、外链接、图片链接信息等等。

本文下载了 2011-5-23 版本的维基百科中文数据备份，获取了如下 3 个数据文件，如表 5.1 所示。

表 5.1 本文所下载的数据资源列表

| 数据文件名 | 数据的意义 |
|---|---|
| pages-articles.xml.bz2 | 概念页面的解释文档内容 |
| pagelinks.sql.gz | 概念页面之间的内部链接 |
| categorylinks.sql.gz | 分类之间的所属关系 |

本文使用 JWPL 工具[83]对维基百科数据进行操作。JWPL（Java Wikipedia Library）是一个开源的访问 wikipedia 数据的 JAVA API 包，它的核心功能包括：(1)快速有效的访问 wikipedia 数据；(2)分析处理 wiki 格式数据；(3)可以处理任何语言。由于 JWPL 核心 API 访问的数据都是存储在 mysql 数据库中结构化数据，所以，首先要把最原始的 wikipedia 数据转换成 mysql 数据库的记录格式，并导入 mysql 中。本文主要利用的就是 JWPL 提供的数据转换工具—DataMachine 对表 5.1 所示的 3 个数据文件进行解析，解析后生成一个 output 目录，该目录里面有 11 个 txt 文件。

目前中文维基百科存在两种文字形式——繁体中文与简体中文。来自台湾、香港、澳门的贡献者一般使用繁体中文，来自中国大陆、新加坡、马来西亚则使用简体中文。作为一个全球华人共同创作的平台，中文维基百科发布的数据中，既有繁体形式的，也有简体形式的，甚至在同一个页面中繁简夹杂[84]。因此在将 output 目录中文件导入到 mysql 数据库前，先将其中的文字信息转换为简体形式，才能符合短文本分类时所选取的中文语料库的文字形式。本文利用维基提供的繁简对应词表和 MediaWiki 的繁简转换功能实现了文件的繁转简。

本文使用 Navicat 客户端程序访问 mysql 数据库，将 output 目录中有用的 9 个文件导入到 mysql 数据库内，生成的数据库（命名为 db_wikipedia）结构如表 5.2 所示。

表 5.2 数据库 db_wikipedia 结构

| 数据表名 | 规模 | 数据表意思 | 字段名 | 字段的意义 |
|---|---|---|---|---|
| category | 99213 | 类别信息 | pageId | 类别 ID，唯一 |
| | | | name | 类别名称 |
| category_inlinks | 248232 | 指向类别的链接信息 | id | 类别 ID，不唯一 |
| | | | inLinks | 该类别所属的父类 ID |
| category_outlinks | 248232 | 类别指向的链接信息 | id | 类别 ID，不唯一 |
| | | | outLinks | 该类别的子类 ID |
| pagemapline | 824011 | 所有概念页面的 title 信息 | id | 概念页面 ID |
| | | | name | 页面 title 名字 |
| page | 824011 | 概念页面信息 | pageId | 概念页面 ID |
| | | | text | 页面全文 |
| | | | isDisambiguation | 是否是消岐页面 |
| page_redirects | 56320 | 重定向信息 | id | 概念页面 ID，不唯一 |
| | | | redirects | 重定向到该概念页面的所有页面 ID |
| page_inlinks | 1569321 | 链入概念页面的链接信息 | id | 概念页面 ID，不唯一 |
| | | | inLinks | 指向该页面的页面 ID |
| page_outlinks | 1569321 | 概念页面链出的链接信息 | id | 概念页面 ID，不唯一 |
| | | | outlinks | 该页面指向的页面 ID |
| page_categories | 1726715 | 页面与类别的关系 | id | 概念页面 ID，不唯一 |
| | | | Category | 页面属于的类别 ID |

其中，表 pagemapline 记录的维基百科中所有概念的唯一 ID 和名称，由于维基百科是一部人人可以编纂的百科全书，其中很多标题概念对文本分类没有太大的意义或者说很多概念只是为了维护维基百科而建立的。所以，在对数据库访问之前，我们必须对表 pagemapline 所包含的概念名称进行筛选，去掉如下概念：(1)维基百科特有的冗余信息，如“维基索引”、“维基分类”、“相关人名列表”等；(2)年代名词，如“1980 年代”、“1975 年”等；(3)带特殊字符的概念，比如“@”、“$”、“￥”等。(4)长度大于 10 个字符的概念。经过处理，pagemapline 记录数由 120 多万条精简为 824011 条，再处理表 page、page_redirects、page_inlinks、page_outlinks、page_categories 中相应的数据，去除其中 ID 在表 pagemapline 中不

存在的记录，最终得到的数据表规模如表 5.2 所示。

### 5.1.2 特征扩展词表构建

在对中文维基百科的数据进行了整理之后，需要按照第三章介绍的方法构建基于维基百科的特征扩展词表。本文运用 JWPL 已定义好的 JAVA API 接口，调用 Page 类、CategoryGraph 类中已经定义好的一些方法，得到每个维基百科主题概念的相关概念以及相关度算法公式中的数据。如表 5.3 列出了构建扩展词表用到的 JWPL 抽象架构中的类和方法。

表 5.3 本文用到的 JWPL 中的类和方法

| 类名 | 方法名 | 类型 | 功能 |
|---|---|---|---|
| Page | getPageId() | Int | 概念的页面 ID |
| | getTitle() | String | 页面的概念名称 |
| | getCategories() | Set<Category> | 概念页面所属的类别集合 |
| | getInlinkIDs() | Set<Integer> | 概念的入链页面 ID 集合 |
| | getOutlinkIDs() | Set<Integer> | 概念的出链页面 ID 集合 |
| | getNumberOfInlinks() | int | 概念的入链页面个数 |
| | isDisambiguation() | boolean | 是否是消歧页面 |
| | isRedirect() | boolean | 是否是重定向页面 |
| CategoryGraph | getPathLengthInEdges() | double | 计算分类体系中两个类别节点的最短距离 |

由于维基百科的链接结构比类别体系更丰富完善，所以公式（3-4）中参数$\alpha$取 0.8，从而使概念间的链接距离对计算相关度的影响更高。将最终生成的维基百科特征扩展词表存储在数据库 db_ExWords 中，为短文本分类奠定基础。

## 5.2 短文本分类实验

### 5.2.1 实验设计

根据短文本分类的基本框架，本文在 windows 环境下，利用 JAVA 语言实现短文本分类过程。实验所用数据集来源于互联网 web 短文本。一般文本分类的实验数据有以下几个来源：TREC 会议网站、路透社的新闻稿、全美医学文献（MEDILINE）等，中文文本分类语料库一般使用搜狗新闻库。但中文短文本语料库目前还没有很成熟的。因此，本文使用 jspider 网页抓取工具，以新浪旗下的中文社区问答平台—“爱问知识人”作为数据集来源[85]，抓取用户提问的问题文

本作为语料库。实验选取 8 个类别共 80000 条问题文本，涉及商业经济、汽车、家庭与生活、健康与医学、买房装修、互联网、演艺娱乐、文学，每类 10000 条问题。取每个类别的 8000 条作为训练集，其余数据用作测试。

为验证基于维基百科的中文短文本分类方法的有效性，本文设置两组实验，第 1 组采用传统的文本分类方法，用 SVM 作为分类器，分类过程如第二章 2.1 节所示，主要是为了和本文的方法进行结果对比；第 2 组采用本文提出的基于维基百科的短文本分类方法，分类过程如第四章 4.2 节所示。对实验结果采用传统的信息检索的评估指标精准率和召回率以及 $F_1$ 测试值进行评估。

### 5.2.2 实验步骤

**实验 1：**

(1) 对测试文本和训练语料集合使用中科院的分词系统 ICTCLAS 进行分词、去停用词等预处理操作。

(2) 分别统计出测试文本和训练语料集的特征词对应的权重，采用 tf-idf 方法（见公式 2-6，L 取值为 0.1），将训练集和测试集表示为向量空间格式。

(3) 使用 LIBSVM 工具对空间向量表示的训练集进行训练，并对测试集进行分类，输出分类结果及评估指标。LIBSVM[86]是台湾大学林智仁（Lin Chih-Jen）副教授等开发设计的一个简单、易于使用和快速有效的 SVM 模式识别与回归的开源软件包，是 SVM 的一种实现工具。实验所使用的 LIBSVM 版本为 2.32。

**实验 2：**

(1) 改进 ICTCLAS 分词系统，将维基百科的概念词条（表 5.2 中数据表 pagemapline 中字段 title 的所有记录）加入后台分词词典，对测试文本和训练语料集合进行分词、去停用词等预处理操作。分词结果对比如图 5.1 所示。

2010年/m 百/m 度/qv 域名/n 被/pbei 劫持/v 事件/n 发生/v 于/p 2010年/m 1月/t 12日/t 。/wj 当天/t ，/wd 中国/ns 大陆/n 最/d 大/a 中文/nz 搜索引擎/n 公司/n 百/m 度/qv 被/pbei 自称/v 是/vshi 伊朗/nsf 网/n 军/n 的/ude1 黑客/n 组织/v 入侵/v ，/wd 导致/v 网/n 民/ng 无法/v 正常/ad 登/v 陆百/m 度/qv 网站/n 达/v 8/a 小时/n 。/wj 下午/t 2时/t ，/wd 百/m 度/qv 主页/n 完全/ad 恢复/v 正常/a ，/wd 但/c 百/m 度空间/n 、/wn 百/m 度/qv 贴/v 吧/y 等/udeng 仍/d 未/d 恢复/v 。/wj

图 5.1(a) 改进前的分词结果

2010年/m 百度/n 域名/n 被/pbei 劫持/v 事件/n 发生/v 于/p 2010年/m 1月/t 12日/t 。/wj 当天/t ，/wd 中国/ns 大陆/n 最/d 大/a 中文/nz 搜索引擎/n 公司/n 百度/n 被/pbei 自称/v 是/vshi 伊朗网军/n 的/ude1 黑客/n 组织/v 入侵/v ，/wd 导致/v 网民/n 无法/v 正常/ad 登陆/n 百度/n 网站/n 达/v 8/a 小时/n 。/wj 下午/t 2时/t ，/wd 百/m 度/qv 主页/n 完全/ad 恢复/v 正常/a ，/wd 但/c 百度空间/n 、/wn 百度贴吧/n 等/udeng 仍/d 未/d 恢复/v 。/wj

图 5.1(b) 改进后的分词结果

(2) 按照第四章 4.3.2 和 4.3.3 过程所述，基于 5.1 节构建的维基百科特征扩展词表，将训练集和测试集表示为扩充后的特征空间向量。

(3) 使用 LIBSVM 工具对空间向量表示的训练集进行训练，并对测试集进行分类，输出分类结果及评估指标。

### 5.2.3 实验结果

实验 l 与实验 2 的结果对比如表 5.4 所示：

表 5.4 实验结果对比

| | 实验一 | | | 实验二 | | |
|---|---|---|---|---|---|---|
| | 精确率(%) | 召回率(%) | $F_1$(%) | 精确率(%) | 召回率(%) | $F_1$(%) |
| 商业经济 | 63.00 | 53.00 | 57.57 | 66.51 | 54.2 | 59.73 |
| 汽车 | 54.70 | 36.32 | 43.65 | 48.84 | 41.73 | 45.01 |
| 健康与医学 | 76.94 | 82.04 | 79.41 | 88.34 | 86.2 | 87.26 |
| 家庭与生活 | 55.86 | 55.78 | 55.82 | 76.93 | 65.95 | 71.02 |
| 买房装修 | 38.01 | 35.93 | 36.94 | 42.44 | 41.79 | 42.11 |
| 互联网 | 73.13 | 59.21 | 65.44 | 74.91 | 62.07 | 67.89 |
| 演艺娱乐 | 48.23 | 58.88 | 53.03 | 66.19 | 79.02 | 72.04 |
| 文学 | 60.23 | 63.76 | 61.94 | 71.3 | 70.89 | 71.09 |
| 平均 | 58.76 | 55.62 | 57.15 | 66.93 | 62.73 | 64.76 |

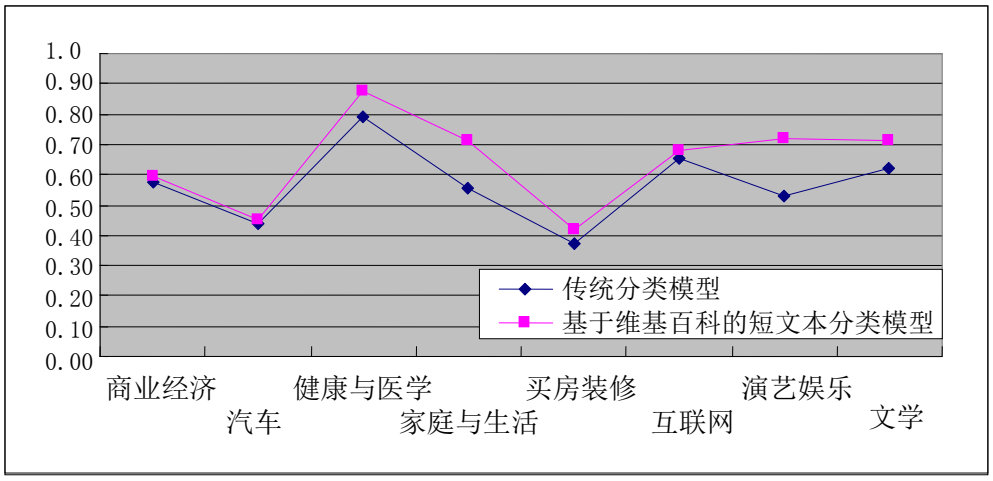



图 5.2 实验结果对比

## 5.3 实验分析

从表 5.4 的数据和图 5.2 的对比图可以观察得出，传统的文本分类模型在应用

于短文本分类时，平均 $F_1$ 值大约为 57%左右，而本文提出的基于维基百科的短文本分类方法，平均 $F_1$ 值约为 65%。整体分类精度有所提升，说明本文提出短文本分类方法，能够对文本的特征进行有效的扩充，并从语义层面考虑特征的关联关系，改善文本分类的性能。

但从表中可以看出实验数据整体偏低。一方面由于实验选取的短文本较短，尽管用本文的方法进行了特征补充，但与普通长文本相比仍难以达到较好的分类效果；另一方面，实验所用的数据来源于互联网，部分数据本身被标注了错误的类别，这种类间噪声会影响分类模型的泛化能力。

其中，汽车类和买房装修类分类效果相对不太好，主要原因可能是这两类的语料在文字使用上区分度不明显。健康与医学类、娱乐类、文学类获得了较高的 $F_1$ 测量值,这可能是由于本文所建立的维基百科特征扩展词表具备一定的偏好性。也就是说，词表中某一类的知识含量偏多，那么对应相同类别的特征扩展效果就比较好。另外，改进后的文本分类效果的提高还和改进前的精度有关系，如互联网类，实验 1 的 $F_1$ 测量值相对较高，那么扩充后的分类效果提高的空间就相对比较小了。

## 5.4 本章小结

本章对本文提出的基于维基百科的中文短文本分类方法进行了实验验证。首先利用 JWPL 工具将维基百科数据结构化并存储于 mysql 数据库中，编程调用 JWPL 中的类和函数，按照第三章所述过程实现维基百科特征扩展词表的构建。其次利用 ICTCLAS 和 LIBSVM 搭建实验平台，收集短文本数据集，将本文提出的基于维基百科的短文本分类方法与传统分类模型对比。最后进行实验结果分析，结果证明本文提出的短文本分类方法较传统分类模型在短文本分类上有更好的效果。

# 第六章 总结与展望

随着互联网和通讯技术的快速发展，涌现出海量的文本信息，其中包含了大量区别于长文本的内容长度相较较短的短文本。这些短文本广泛存在于互联网、电子通讯等各种媒介中，与人们的生活息息相关。但数据丰富、信息贫乏现象仍是当今信息社会所面临的主要问题。如何从堆积如山的数据中获得有用的、有价值的信息，单凭人工已经无法胜任，必须借助计算机强大的信息处理能力，并由此产生了文本分类问题，随之也诞生了多种文本分类技术。然而，现有的文本分类技术大多是针对长文本进行设计的，如何借鉴长文本的分类技术并针对短文本进行相应的改进，是研究人员所面临的一个艰巨任务，具有重要的研究意义。

## 6.1 总结

由于存在稀疏、实时、海量、不规范性的特点，短文本分类的有效途径是基于外部知识来辅助分类。因此本文引入特征扩展的思想，提出了一种基于维基百科的中文短文本分类方法。全文的工作简单总结如下：

本文首先介绍和分析了短文本分类和维基百科的研究现状，提出了中文短文本的特点和短文本分类的难点，从技术角度讨论了传统中文文本分类模型，如分词过程、文本表示过程、分类器构造、分类效果评价指标等。

其次在本文的核心章节，对基于维基百科的中文短文本分类方法进行了详细描述。第三章从介绍维基百科的语义结构入手，利用维基百科中概念间的链接关系确定特征词的相关概念，并基于概念间的链接距离和类别距离计算主题概念与其相关概念的相关度，从而建立维基百科特征扩展词表。第四章对传统中文文本分类模型进行改进，构建基于维基百科特征扩展词表的短文本分类模型。在文本预处理环节，对传统分类软件进行改进，加入维基百科中的概念；在本文表示环节，分情况讨论了文本特征与维基百科概念的匹配方法，在此基础上构建概念向量空间；加入特征扩展环节，利用维基百科特征扩展词表对文本向量进行补充并改进特征权重；最后利用 SVM 算法训练分类器。此模型在特征扩展、词义消歧、同义词识别、潜在语义信息挖掘、流行词识别等方面较传统模型更适应中文短文本分类。

最后对本文提出的分类方法进行实验验证。先将维基百科数据结构化为特征扩展词表，存储在数据库中；然后设置一组对比实验，将本文提出的短文本分类方法与改进前的传统分类模型对比。实验结果证明了基于维基百科的中文短文本分类方法能够提高短文本分类的效果。

## 6.2 进一步工作

本文对基于维基百科的短文本分类方法进行了初步的探索，将维基百科抽象为结构化数据并作为外部知识库对稀疏的短文本向量进行语义扩展、信息补充，实现了短文本分类，最终提高了短文本分类效果。但是由于时间和条件限制，仍然有一些问题是需要进一步探索和研究的。今后的工作重点将在以下几个方面展开：

(1) 维基百科作为一个开放式的大规模的在线百科全书，已经被越来越多的研究者认可和利用。本文虽然有效运用了维基百科中的链接结构和类别体系，但没有考虑到概念页面的内容、信息盒、外部链接等语义结构，因此需要对维基百科做进一步探索，优化概念间相关度的计算方法，更接近人对现实世界事物之间关系的理解。

(2) 本文在构建维基百科特征扩展词表时，人工设立阈值选取了主题概念的相关概念。阈值的选取参考了相关文献，但在后续工作中仍需要进行大量实验。因为阈值的确定关系到扩展词表的规模，关系到短文本向量空间的维度，从而影响最终的分类效果和效率。

(3) 短文本的特征扩展有两种思路，一种是扩展测试文本的特征向量，另一种是扩展训练文本特征向量。本文将测试和训练文本都进行了扩展，但采用哪种特征扩展方法更加适合也是需要进一步研究的问题。

(4) 由于短文本分类的研究尚处于起步阶段，还没有成熟和通用的短文本分类语料库。在后续工作中，建立质量高的语料库、尽量减小数据集合中的噪声对分类的影响是短文本研究工作的当务之急。

总之，本文还有很多不足和进一步完善的地方，无论是短文本分类的研究还是维基百科的应用，都具有更广阔的空间和更深远的意义。

# 致谢

随着毕业论文工作的结束，我在西安电子科技大学经济管理学院攻读硕士研究生学位的学习生活也接近尾声。在老师、家人和朋友的帮助和陪伴下，这两年多的时间让我成长了很多，收获了很多。

首先衷心感谢给予我悉心教导的导师刘怀亮教授，感谢他为我们提供了一个良好的学习、科研环境，使我们能够快速成长。刘老师是一个学识丰厚又不失诙谐亲切的人，我很感激他对我要求上的严格与认真以及能力上的肯定与信任。不管是在学习还是在生活中，刘老师都给予了我亲切的关怀、深深地教诲以及不懈的支持，我从中得到了进取的动力，也为他严谨的治学精神深深感动。回想起刘老师和刘姐对我的帮助、关心和鼓励，让我倍感温馨。

另外要感谢赵捧未老师、窦永香老师、刘成山老师、秦春秀老师及学院其他老师的帮助，他们的每一次指导都让我受益匪浅，每一门课程都让我印象深刻，从他们那学到的治学态度将使我受用一生。

感谢已毕业的师兄黄章益、胡涛在学习和科研上给予我的指导；感谢实验室的左晓飞同学以及师弟师妹们在生活和学习中对我的帮助；感谢我宿舍亲爱的姐妹们以及所有情报学同学，很荣幸有他们的陪伴，让我的研究生生活多姿多彩。

还要感谢含辛茹苦养育我的父母和一直支持照顾我的朋友们。父母的呵护和疼爱，才让我有了现在的生活；朋友的陪伴和支持，才让我的生活变得充实与快乐。

最后，向百忙之中抽出时间来评阅我论文的各位专家和教授致以衷心的感谢，向即将参加答辩评审的各位老师表示诚挚的敬意！

# 参考文献

[1] 第 30 次中国互联网络发展状况统计报告. http://www.cnnic.net.cn/hlwfzyj/［2012-07-19］.

[2] Fischer G, Stevens C. Information Access in Complex Poorly Structured Information Spaces[C]. In: Proc. of the CHI'91 Conference Proceedings. 1991:63-70.

[3] 马文娟. 文本特征降维与分类规则抽取方法研究与应用[D]. 大连: 大连理工大学, 2007.

[4] 庞观松，蒋盛益. 文本自动分类技术研究综述[J]. 情报理论与实践，2012, 35(2):123-128.

[5] 柴春梅. 互联网短文本信息分类关键技术研究[D]. 上海:上海交通大学, 2009.

[6] 赵浩镇. 中文维基百科语义信息抽取研究[D]. 武汉: 武汉大学, 2008.

[7] http//:www.wikipedia.org［EB/OL］.

[8] M. Sahami and Timothy D Heilman, A web-based kernel function for measuring the similarity of short text snippets[C]. In: Proceedings of the 15th international conference on World Wide Web, 2006:377-386.

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

[10] Hynek, K Jezek, O Rohlik. Short Document Categorization-Item sets 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.

[11] Zelikovitz S, Transductive M F. Learning for Short-Text Classification Problem using Latent Semantic Indexing International[J]. Journal of Pattern Recognition and Artificial Intelligence, 2005, 19(2):143-163.

[12] Hui He, Bo Chen, Weiran Xu, Jun Guo. Short Text Feature Extraction and Clustering for WebTopic Mining[C]. In: Third International Conference on Semantics, Knowledge and Grid. 2007: 382-385.

[13] 王细薇, 樊兴华, 赵军. 一种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.

[14] 郭泗辉，樊兴华. 一种改进的贝叶斯网络短文本分类算法[J]. 广西师范大学学报(自然科学版), 2010, 28(3): 140-143.

[15] 高金勇，徐朝军，冯奕競. 基于迭代的TFIDF在短文本分类中的应用[J]. 情报理论与实践. 2011, 34(6): 120-122.

[16] 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.

[17] 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.

[18] 闫瑞，曹先彬，李凯. 面向短文本的动态组合分类算法[J]. 电子学报，2009, 37(5): 1019-1024.

[19] D Song, P D Bruza, Z Huang. Classifying Document Titles Based on Information Inference[C]. In: proceedings of the 14th International Symposium on Methodologies for Intelligent Systems , Japan , 2003:297-306.

[20] X Phan, L Nguyen, S Horiguchi. Learning to classify short and sparse text & Web with hidden topics from large-scale data collections[C]. In: proceeding of the 17th International Conference on World Wide Web, 2008: 91-100.

[21] Ferragina, P. and A. Gulli. A personalized search engine based on web-snippet hierarchical clustering[C]. In: Special Interest Tracks and Poster Proceedings of WWW-05, International Conference on the World Wide Web. 2005.

[22] 宁亚辉，樊兴华，吴渝. 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.

[23] 王盛，樊兴华，陈现麟. 利用上下位关系的中文短文本分类[J]. 计算机应用，2010,30(3): 603-606.

[24] 黄永文，何中市，伍星. 用户评论的分类获取[J]. 计算机应用，2009，29(3): 846-848, 857.

[25] 崔争艳. 基于语义的微博短信息分类[J]. 现代计算机(专业版). 2010，(8): 18-20.

[26] 王雅蕾，王君泽，王国华等. 问答服务中的基于类文档排名的问题分类算法[J]. 情报科学. 2012, 30(2): 296-301.

[27] 刘金岭. 基于降维的短信文本语义分类及主题提取[J]. 计算机工程与应用，2010, 46(23): 159-161.

[28] Gabrilovich E, Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis[C]. In: Morgan Kaufmann Publishers Inc. Hyderabad, India, 2007.

[29] Gabrilovich E. Overcoming the brittleness bottleneck using Wikipedia Enhancing text categorization with encyclopedic knowledge[C]. In: Proceedings of The 21st National Conference on Artificial Intelligence(AAAI), Boston, 2006: 1301-1306.

[30] Wang, P. and C. Domeniconi. Building semantic kernels for text classification

using Wikipedia[C]. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008. Las Vegas, Nevada, USA: ACM.

[31] Schonhofen P. Identifying Document Topics Using the Wikipedia Category Network[C]. IEEE Computer Society, 2006.

[32] Bawakid A, Oussalah M. Centroid-based Classification Enhanced with Wikipedia[C]. IEEE Computer Society, 2010.

[33] 苏小康. 基于维基百科构建语义知识库及其在文本分类领域的应用研究[D]. 武汉：华中师范大学, 2010.

[34] 邱强. 基于关键词的文本分类研究[D]. 西安:西北农林科技大学, 2010.

[35] 王锦, 王会珍, 张俐. 基于维基百科类别的文本特征表示[J]. 中文信息学报, 2011, 25(2):27-31.

[36] 张海粟, 马大明, 邓智龙. 基于维基百科的语义知识库及其构建方法研究[J]. 计算机应用研究, 2011, 28(8): 2807-2811.

[37] Philip J H, David D L. Guest Editors' Introduction to the Special Issue on Text Categorization[J]. ACM Transactions on Information Systems, 1994, 12(3): 231-241.

[38] Sebastiani F. Machine learning in automated text categorization[J]. ACM Computing Surveys. 2002, 34(1): 1-47.

[39] 陈雅芳. 中文文本分类方法研究[D]. 浙江: 浙江大学, 2010.

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

[41] 冯志伟. 中文信息处理与汉语研究[M]. 北京:商务印书馆, 1992, 12.

[42] 奉国和, 郑伟. 国内中文自动分词技术研究综述[J]. 图书情报工作. 2011, 55(2): 41-45.

[43] 郭凯. 面向Web文本的数据清洗关键技术的研究与实现[D]. 西安: 西安电子科技大学, 2009.

[44] 郭辉, 苏中义, 王文. 一种改进的MM中文分词算法[J]. 微型电脑应用, 2002, 18(l): 13-15.

[45] 吴雅娟, 柳培林, 丁子睿. 基于统计分词的中文文本分类系统[J]. 电脑知识与技术, 2005, (4): 71-74.

[46] 李燕, 孟庆昌, 李宝安. 中文信息处理技术—原理与应用[M]. 清华大学出版社, 2005.

[47] 翟凤文, 赫枫龄, 左万利. 字典与统计相结合的中文分词方法[J]. 小型微型计

算机系统, 2006, 27(9): 1766-1771.

[48] 梁南元. 书面汉语自动分词系统-CDWS[J]. 中文信息学报, 1987, 1(2): 44-52.

[49] 张永奎，李国臣. 新闻语料自动分词系统[J]. 山西大学学报，1993，16(3): 280-284.

[50] 何克抗，徐辉，孙波. 书面汉语自动分词专家系统设计原理[J]. 中文信息学报, 1992, 5(2): 1-14.

[51] http://ictclas.org/［2012-12-31］.

[52] http: / /www.hylanda.com/［2012 -12-31］.

[53] 王元珍，钱铁云，冯小年. 基于关联规则挖掘的中文文本自动分类[J]. 小型微型计算机系统, 2005, 26(8):1380-1383.

[54] 周城. 面向中文Web评论的情感分析技术研究[J]. 长沙：国防科学技术大学, 2011.

[55] 黄云平，孙乐，李文波. 基于上下文图模型文本表示的文本分类研究[C]. 第四届全国信息检索与内容安全学术会议论文集（上）, 2008.

[56] Salton G, Yang C S. On the Specification of Term Values in Automatic Indexing[J]. Journal of Documentation, 1973, 29(4):351-372.

[57] Salton G,Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing Letters . 1988.

[58] 施聪莺，徐朝军，杨晓江. TFIDF算法研究综述[J]. 计算机应用，2009，29(6): 167-170.

[59] Yang Y M，Liu X. A re-examination of text categorization methods[C]. In：Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, BerkeIey, USA. August, 1999.

[60] Kononenko I, Simec E. Induction of Decision Trees Using RELIEFF[J]. Mathematical and statistical methods in artificial intelligence. Springer Verlag, 1995:1-19.

[61] D Koller, M Sahami. Hierarchically Classifying Documents Using Very Few Words[C]. In: Proceedings of the Fourteenth International Conference on Machine Learning, 1997: 170-178.

[62] Vapnik V. The Nature of Statistical Learning Theory[M]. Springer-Verlag, New York: 2000.

[63] 谭金波，李艺，杨晓江. 文本自动分类的测评研究进展[J]. 现代图书情报技术, 2005, (5): 46-49.

[64] 苏小康，何婷婷，涂新辉. 一种基于维基百科知识库的中文文本分类方法研究

[C]. 第十届全国计算语言学学术会议, 中国山东烟台, 2009:514-520.
[65] 董振东. HowNet. http://www.keenage.com/html/c-index.html. [EB/OL]
[66] http://www.wikipedia.org/download[EB/OL].
[67] http://www.alexa.com/topsites［2012-08-11］.
[68] http://zh.wikipedia.org/wiki/Wikipedia:About[2012-12-31].
[69] Thelwall M, Buckley K, Paltoglou G. Sentiment in Twitter events［J］. Journal of the American Society for Information Science and Technology, 2011, 62(2): 406-418.
[70] Kittur, Chieh, Suhbw. What's in Wikipedia? Mapping topics and Conflict Using Socially Annotated Category Structure[C]. In: Proc of the 27th International Conference on Human Factors in Computing Systems. Boston, 2009:1509-1512.
[71] 王细薇, 沈云琴. 中文短文本分类方法研究[J]. 现代计算机(专业版), 2010, (7): 28-31.
[72] 李赟, 黄开妍, 任福继等. 维基百科的中文语义相关词获取及相关度分析计算[J]. 北京邮电大学学报, 2009, 32(3): 109-112.
[73] 涂新辉, 张红春, 周琨峰等. 中文维基百科的结构化信息抽取及词语相关度计算方法[J]. 中文信息学报, 2012, 26(3): 109-115.
[74] 裘江南, 秦璇, 仲秋雁. 异质知识网络相关度算法研究[J]. 情报学报, 2011, 30(5): 495-502.
[75] Strube M, Ponzetto S P. WikiRelate! Computing Semantic Relatedness Using Wikipedia[C]. In: Proc. of Association for the Advancement of Artificial Intelligence. Boston, USA: IEEE Press, 2006: 1419-1424.
[76] Milne D, Witten I. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links[C]. In: The Workshop on Wikipedia and Artificial Intelligence at AAAI, Chicago, 2008: 25-30.
[77] Rudi L, Paul M B. The Google Similarity Distance[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3):370-383.
[78] 张玉芳, 艾东梅, 黄涛等. 结合编辑距离和Google距离的语义标注方法[J]. 计算机应用研究, 2010, 27(2): 555-557.
[79] 罗志成, 马费成, 吴晓东等. 从维基分类系统构建中文语义词典研究[J]. 信息系统学报, 2008, 2(2): 68-77.
[80] 朱恒民, 马静, 黄卫东. 基于领域本体的中文web文本主题特征抽取方法[J]. 情报理论与实践, 2008, 31(2):34-41.
[81] 马宏伟. 基于SVM的中文文本分类系统的建模与实现[D]. 辽宁: 大连理工大学, 2005.

[82] Guyon I, Boser B, Vapnik V. Automatic capacity tuning of very large VC-dimension classifiers[J]. Advances in Neural Information Processing Systems. San Mateo, 1993,5:147-155.

[83] http://code.google.com/p/jwpl/ [2012-05-01].

[84] 张淑君，张勤，宋倩倩．维基分类系统动态网络演化研究[J]．情报杂志，2009, 28(z2):69-77.

[85] http://iask.sina.com.cn/［2012-09-18］.

[86] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

# 硕士期间科研成果

**发表论文:**

[1] 范云杰, 刘怀亮. 基于维基百科的中文短文本分类研究[J]. 现代图书情报技术, 2012, 28(3): 47-52.

[2] 范云杰, 刘怀亮, 左晓飞等. 社区问答中基于维基百科的问题分类方法[J]. 情报科学, 已录用.

[3] Fan Yunjie, Liu Huailiang. Research on Technology Anti-corruption Platform Based on SOA[C]. The Second International Conference on Management Science and Engineering. Chengdu, China. 2011:326-331.