**微博短文本分类系统的设计与实现开题综述**

1. 研究背景及意义

信息化时代的到来，使得人们对互联网越发依赖。文本是互联网上主要的信息载体，近年来，随着计算机和通讯技术的飞速发展，短文本的数量激增。所谓的短文本，是指那些长度较短的文本形式，其特点是长度短，一般不超过200个字符，内容精炼内聚，用很少的语言便能够表达出对人和事的看法，典型的载体包括手机信息，新闻快讯，网友评论，微博信息等。

根据中国互联网络信息中心（CNNIC）于2013年7月发布的《第32次中国互联网发展状况统计报告显示》，截止2013年6月底，中国网民规模达5.91亿，互联网普及率为44.1%，其中微博用户的网民已具备33077万规模，在社会化媒体中占据重要地位。

微博短文本在短文本中占据着半壁江山，其快速灵活的方式已经基本改变了人们交流的方式。人们在微博上畅所欲言，互相交换信息或表达看法。因此，对微博短文本分类技术及应用进行研究，在热点事件发现，个性化推荐，社会化媒体营销等领域都有较大的研究价值。

本课题研究工作是以实习公司的社会化媒体管理平台项目为背景，设计并实现一个微博短文本的分类系统，分类识别出海量微博文本中的广告、活动、新闻类等类别文本。为社会会媒体管理平台上的品牌客户提供诸如舆情分析，媒体营销等服务。

1. 国内外研究现状

文本分类是文本挖掘的主要技术之一，其根据长度的不同可将文本分为长文本和短文本。自20世纪50年代美国IBM公司的H.P.Luhn提出词频统计思想开始，一直到今天，国内外学者对于长文本分类方法已经进行了较多的研究，其过程主要分为训练和测试两个阶段。大致如下：首先对得到的长文本语料库进行预处理得到训练集和测试集，然后按照所选的特征选择及特征权重计算方法将训练集和测试集表示成特征向量，最后通过分类器进行训练和测试。在训练阶段根据所选的分类器对训练集的特征向量数据进行文本建模生成模型文件，在测试阶段根据模型文件对测试集进行判别完成分类。迄今为止，基于机器学习的长文本分类方法已实现了许多经典的分类器[1]，例如支持向量机SVM、K-近邻、朴素贝叶斯、决策树、中心向量法等。

短文本由于字符较短，词的个数有限，数据集特征稀疏，且因为经常包含些网络用语，呈现一定的不规范性，区别于以前的长文本挖掘，长文本表示模型以及机器学习方法直接应用到短文本上效果不佳。短文本是伴随着Web2.0的兴起而逐步引起研究者的注意，研究过程前后不到10年，目前国内外只对短文本分类进行了少量的研究工作。现阶段对短文本的分类研究主要基于以下3种思路开展：

**2.1、对成熟的长文本分类器和特征选择方法进行改进**

鉴于以前的文本分类方法研究比较成熟，对成熟的文本分类方法加以改进应用到短文本分类是很多学者的研究热点。Healy[2]运用不同的特征类型与特征表达方法通过用实验比较朴素贝叶斯、支持向量机两种分类器对短文本的分类效果，证明了选择不同的分类器和特征方法组合对分类性能的影响。郭泗辉[3]在考虑特征值连接强度的情况下，对传统的贝叶斯进行改进，提出了适用于网络短文本的贝叶斯分类算法。张素智[4]等提出了一种基于语义的KNN短文本分类算法，采用基于字的分词策略提取短文本特征词，结合中国知网HowNet对关键词进行概念映射提高短文本语义表达，并使用LSA对特征词进行降维处理，实验表明该算法有效的提高了分类性能。曾佳妮[5]针对短文本的特征稀疏问题，提出在文本对预处理后，使用条件随机场CRF模型用类别作为标注进行标注预测的短文本分类方法，具有很好的分类效果。

在特征选择的改进方面，Monica[6]等人提出对于单一的特征选择方法，例如文本频数（Document Frequency，DF）、信息增益（Information Gain，IG）、卡方检验（CHI-square test，CHI）等，组合后综合考虑对提高分类效果有所帮助。Liu[7]等人对互信息和信息增益的这两种特征选择方法引入了正则化进行文本的特征选择。朱灏东[8]等人提出基于优化的模拟退火算法，对最优解增加记忆，应用于文本特征选择。路永和[9]等提出了一种基于粒子群优化的文本特征选择方法（POSTFS），引入了粒子群算法来挖掘文本特征选择规则。李太白[10]在路永和的基础上有所改进，其思路是首先采用模糊熵对原始文本集进行特征预选，然后对预选的特征集采用改进的粒子群优化算法进行第二次特征选择。

**2.2、利用其他资源对短文本的特征进行扩展**

一些学者基于短文本词个数有限，特征缺失，提出了引入其他知识资源对文本特征扩展的方法对短文本进行分类，主要集中在对文本间的概念相似度计算方面。Merhran Sahami[11] ，YIH W[12]等人提出使用基于Web语义核函数的方法，使用搜索引擎返回的信息对短文本进行扩展，来度量短文本的相似性。Somnath[13]，范运杰[14]，赵辉[15]等提出结合维基百科语义知识库对文本特征词进行扩展，提高语义描述能力。王细薇，宁亚辉，王盛[16-18]等均借助了知网HowNet帮助分类，王细薇[16]以训练数据作为背景语料，利用关联规则基于FP-Growth算法挖掘训练数据中的共现词对构造扩展词表，创建特征共现集，再通过知网HowNet引入语义信息实现分类。后续王细薇[17]又基于特征的共现关系，在改进特征词权重计算公式及特征扩展策略的前提下，提出了一种改进的短文本特征扩展算法。宁亚辉[18]等提出借助知网从语义方面将领域高频词扩展为概念和义元，计算不同概念包含相同义元的信息量衡量词的相似度。王盛[19]等人提出利用知网来确定短文本中概念对和词语对的上下位关系，来扩展测试文本的特征向量。张东娜[20]考虑WordNet中概念之间的层次关系以及数据集中的概率信息，提出了一种被称之为SS-CW（Semantic Similarity Based on Corpus Statistics and WordNet）的通用的短文本语义相似性计算方法。翟延冬等[21]在张东娜的基础上进而提出了一种新的计算短文本语义相似性的算法ST-CW（Short Text based on Corpus and WordNet）。

**2.3、挖掘短文本内部的隐含信息对分类提供帮助**

近年来，通过挖掘短文本内部语义信息的方法帮助进行分类引起学者的关注，Zelikovitz[22]使用潜在语义索引LSI（Latent Semantic Indexing），将词频矩阵转化为奇异矩阵，分解奇异矩阵来挖掘特征内部之间隐含的关联规则协助分类，但在对数据量的分解处理过程对计算机CPU处理能力要求过高。2003年在Blei提出LDA（Latent Dirichlet Allocation）后，为解决短文本的特征稀疏提供了新的思路，Phan[23]等使用LDA模型基于一个大语料库得到一簇隐含主题，将测试文本转化为文档—主题—词这样的三层的贝叶斯层次模型，该方法的缺陷是不能用于计算两篇文本之间的相似度。Quan[24]等将通过语料库生成的隐含主题作为第三方，考察词语间的关联性，提出了基于主题的相似性（Topic-Based Similarity，TBS）度量方法，度量了两篇文本之间的相似性，却未能解决上下文依赖性的问题。胡勇军等[25]针对短文本的特点，提出了一种基于LDA高频词扩展的方法，利用LDA得到每个文本的隐含主题并设定阈值，将概率大于阈值的主题所对应的高频词扩展到文本中来降低短文本的噪声和稀疏性。张志飞等[26]在Quan的基础上提出了一种基于隐含Dirichlet分布模型的短文本分类方法，解决了上下文依赖性的问题，并采用KNN近邻方法对网易页面标题数据进行分类，实验证明，分类性能在Quan的基础上有所提高。

1. 相关技术分析

3.1 LDA模型

LDA（Latent Dirichlet Allocation）模型是Blei提出的一种对离散数据集建模的概率增长模型，LDA是一种非监督机器学习技术，可以用来识别大规模文档集或语料库中潜藏的主题信息，其基本思想是每个文档都可以表示成若干潜在主题的混合分布，每个主题是词汇表中所有单词的概率分布，它属于完全的生成概率模型，具有清晰地层次结构，包含文档、主题和词三个层次。所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到的。

对于语料库中的每篇文档，LDA定义了如下生成过程：语料库中的每一篇文档与 *T*（事先给定）个主题的一个多项分布相对应，将该多项分布记为 *θ*。每个主题又与词汇表（vocabulary）中的  *V*个单词的一个多项分布相对应，将这个多项分布记为 *ϕ*。*θ* 和*ϕ*分别有一个带有超参数*α*和*β*的Dirichlet先验分布。对于一篇文档*d*中的每一个单词，我们从该文档所对应的多项分布*θ*中抽取一个主题*z*，然后我们再从主题*z*所对应的多项分布*ϕ*中抽取一个单词*w*。将这个过程重复*Nd（*文档*d*的单词总数*）*次，就产生了文档*d*。LDA模型生成过程可以用如下的图1表示：

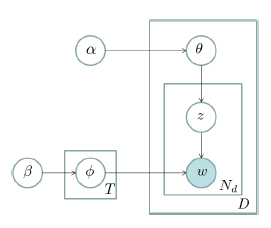


图1 LDA主题模型

该模型有两个参数需要推断：一个是“文档-主题”分布*θ*，随机变量*θ*服从Dirichlet分布，用1×T的列向量表示每篇文档中每个隐含主题的权重，另外是*T*个“主题-单词”分布*ϕ*，同样服从Dirichlet分布，用T×V的矩阵表示给定主题时词的条件分布。通过学习这两个参数，我们可以知道文档作者感兴趣的主题，以及每篇文档所涵盖的主题比例等。推断方法主要有LDA模型作者提出的变分-EM算法，还有现在常用的Gibbs抽样法。

3.2 支持向量机SVM

支持向量机方法是建立在统计学习理论的VC（Vapnik-Chervonenkis Dimension）维理论和结构风险最小化原理基础上的，它将降维和分类结合在一起，在模型的复杂性和学习能力之间寻求最佳折衷，目的是在样本集中找到一个最优的分类面使两类样本的分类间隔最大。

支持向量机在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势，由于SVM算法不受样本趋于无穷大理论的限制，也适合于大样本集的文本分类，并能够推广应用到函数拟和等其他机器学习问题中。

在SVM理论中，原始的输入空间通过一个核函数映射到一个高维的内积空间，这个高维的空间称为特征空间，在这个空间中，构造一个最优分类超平面将分类器的泛化能力最大化，并保证最小的分类错误率。它包含三个核心思想：求最优分类面以取得较好的推广能力；提出软间隔的概念以解决线性不可分问题；引入核函数使解平面从线性扩展到非线性。

支持向量机的基本思想可用图2说明，方形点和原型点分别代表两类样本，H是分类线，H1和H2分别通过各类中离分类线最近的样本且平行于分类线的直线，它们之间的距离被称为分类间隔。所谓最优分类面就是要求分类面不但能将两类正确分开（训练错误率为0，经验风险最小），而且使分类间隔最大。

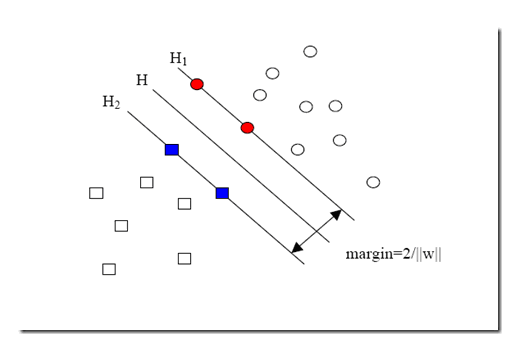


图2 SVM基本原理

假设给定一组线性可分的训练数据集(xi,yi)，xi就是文本向量（维数很高），yi就是分类标记。在二元的线性分类中，这个表示分类的标记只有两个值，1和-1（用来表示属于还是不属于这个类）。有了这种表示法，我们就可以定义一个样本点到某个超平面的间隔： δi=yi(wxi+b)，如果某个样本属于该类别的话，那么wxi+b>0，而yi也大于0；若不属于该类别的话，那么wxi+b<0，而yi也小于0，这意味着yi(wxi+b)总是大于0的，而且它的值就等于|wxi+b|。调整参数w和b的值将判别函数归一化，间隔可以写成：，引入约束条件yi[(w·xi)+b]≥1 (i=1,2,…,l) （l是总的样本数），求解最优分类面的两类分类问题也被我们转化成了求解一个带约束的最小值的问题。两类间的分类间隔为2/||w||，为使间隔最大化也就是使2/||w||最小。满足上式且使最小的分类面就叫做最优分类超平面。这就是SVM二类分类问题的基本思想。

1. 研究目标及内容

4.1论文研究目标

本文的研究工作是以实习公司的社会化媒体管理平台项目为背景，该平台是一个为品牌提供更加容易与消费者沟通并挖掘更多资讯和商机的互动平台。本文的研究目标是对微博短文本分类的技术及应用进行研究，设计并实现一个微博短文本的分类系统，分类识别出海量微博文本中的广告、活动、新闻类等类别文本。

4.2 论文研究内容

针对上述目标，本文研究内容包括：

1、系统需求分析：以实习公司社会化媒体管理平台项目为背景，对微博短文本分类系统涉及的业务实体、业务流程、功能和非功能性需求进行分析。

2、相关技术研究，本系统所涉及的相关技术主要有：

1. 以中文分词为核心的微博文本预处理技术；
2. 文本特征选择及权重计算技术，特征选择主要包括：文本频数（document frequency），信息增益(information gain)，卡方检验（chi-square)等；特征权重的计算方法主要有： (tf\*chi-square),tf\*ig(tf\*information gain),tf\*gr(tf\*gain ratio)等。
3. 特征文本建模技术：主要基于VSM（Vector Space Model）和LDA（Latent Dirichlet Allocation）。鉴于微博短文本的语言特点，使用传统的向量空间模型VSM建立‘文档——特征词’矩阵已不能很好的表现文本特征，而LDA把文档看作若干主题的混合分布，主题又是若干特征词的概率分布，是一种对文本数据的主题信息进行建模的方法，它实现了‘文档——主题——特征词’这样一个三层贝叶斯层次模型，在达到特征降维的同时又能很好的表现文档特征。
4. SVM（Support Vector Machine）分类器技术：SVM建立在VC维理论和结构风险最小化基础上，将文本分类转化为一系列的二类分类问题，它将降维和分类相结合，在样本集中寻找最优分类面使得两类样本的分类间隔达到最大，相对于其他的分类算法,如贝叶斯，KNN-近邻等分类算法，SVM在效果和稳定性上占有优势。

3、系统的设计与实现，包括：系统架构设计，文本预处理模块、特征选择及权重计算模块、特征建模模块、分类器模块等功能模块设计与实现，基于LDA主题模型的文本建模、基于SVM的微博短文本分类算法等关键技术研究。

4、系统测试：以实习公司从腾讯，新浪微博平台抓取的18万条微博文本数据人工筛选出2万余条数据作为训练集及1万余条数据作为测试集，对文本分类的相关指标如召回率（recall）、准确率（precision）、F1-measure等进行测试验证。系统性能要求1天实现对1000万条微博文本进行分类。

5、系统应用：本文将对系统上线运行3个月的应用效果进行分析和总结。

1. 技术路线及创新意义

本课题的关键问题主要有：

1） 利用正则表达式过滤掉微博文本中噪音信息：传统的文本预处理主要包括分词及去除停用词两个主要步骤，微博文本微博文本中包含有大量的超链接，用户名，表情符号等与分类无关的噪音信息，本文将在文本预处理阶段增加通过正则表达式技术过滤掉这些垃圾数据的过程，有利于减小模型的空间维度并提高分类的准确性。

2） LDA主题模型的建立：语料库中每篇文档与T（主题数，事先指定）个主题的一个多项分布相对应，记为θ，每个主题又与词汇表中的V个单词的一个多项分布相对应，记为φ，LDA的目标是根据已有的数据重建θ和φ，而多项式分布θ和φ分别有一个带有超参数α和β的Dirichlet先验分布，所以LDA事实上就是估计α和β参数。参数推理采用Gibbs抽样算法，迭代足够多的次数，每篇文档表示固定隐含主体集上的概率分布，最终得到整个语料库的“文档——主题”矩阵。传统的向量空间模型是“文档——特征词”矩阵，容易用余弦相似度等方式来度量文档间相似度，但由于词汇量庞大会导致计算缓慢，故本文将对整个文档集的表示以向量空间模型为基础，不过以选择“文档——主题”的分布概率矩阵来表示文本特征，可以实现对文本表示模型的有效降维，且易于计算文本间的相似度。

3） SVM多类分类器的实现：多类分类器是对二类分类问题的推广，目前常用的多类SVM分类方法主要有：“一类对其余”方法，即对于K类分类问题，构造K个将其中一个类别与其他所有类别分开的两类支持向量机分类器；“一类对一类”方法，即对于K类分类问题，一对一方法需要构造K（K-1）/2个分类器，分类时采用“投票法”，每个分类器都对待分类样本类别进行判断为相应的类别“投票”，最后将得票最多的类别作为待分类样本的类别；DAG方法，即对于K类分类问题，构造K（K-1）/2个分类器，与“一对一”不同之处在于它采用了有向无环图的组合策略，不必遍历所有的分类器，具有更高的分类效率，分类时从根节点开始，根据该节点的分类结果用下层节点继续分类，直到达到某个叶节点为止，该叶节点的类别即为待分类样本类别。相对于其他分类算法，SVM占有效果和稳定性上的优势，本文在基于类别信息的“文档——主题”矩阵上训练SVM分类器，可以将LDA模型良好的特征降维、文本表示能力和SVM强大的分类能力结合，有助于提高文本分类的性能。

1. 参考文献
2. 陈袆荻，秦玉平. 基于机器学习的文本分类方法综述[J]. 渤海大学学报：自然科学版，2010，6（2），201-205.
3. Healy M, Delany S, Zamolotskikh A. An Assessment of Case Based Reasoning for Short Text Message Classification[C]. In Proceedings of the 16th Irish Conference on Artificial Intelligence &cognitive Science (AICS’05), 2005: 257-266.
4. 郭泗辉，樊新华. 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.
5. 张素智，刘婧姣. 基于语义的KNN短文本分类算法研究[J]. 郑州轻工业学院学报：自然科学版，2012，27（6），01-04.
6. 曾佳妮. 基于条件随机场的中文短文本分类算法研究[D]. 上海：上海交通大学（2012）
7. Monica Rogati, Yiming Yang. High-Performing Feature Selection for Text Classification[C]. Proceedings of the eleventh international conference on Information and Knowledge management. (CIKM’02), 2002:659-661.
8. Huan Liu, Hiroshi Motoda. Computational Methods of Feature Selection [M]. Taylor & Francis Group, LLC, USA, 2008: 257-330.
9. 朱灏东，钟勇. 使用优化退火算法的文本特征选择[J]. 计算机工程与应用，2010，46（4）：8-11.
10. 路永和，曹利朝. 基于粒子群优化的文本特征选择方法. 现代图书情报技术，2011，208（7），76-81.
11. 李太白. 短文本分类中特征选择算法的研究[D]. 重庆：重庆师范大学（2013）
12. Mehran Sahami and Timothy D. Heiman. A Web-based Kernel Function for Measuring the Short Text Snippets[J]. www 2006, May 23-26: 377-386.
13. YIH W, MEEK C. Improving similarity measures for short segments of text[C]. // Proceedings of the 22nd Conference on Artificial Intelligence. Menlo Park : AAAI Press ,2007: 1489-1494.
14. Somnath Banerjee, Krishnan Ramanathan, Ajay Gupta. Clustering Short Texts using Wikipedia[C]. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, New York, USA, 2007: 787-788.
15. 范云杰，刘怀亮. 基于维基百科的中文短文本分类研究[J]. 现代图书情报工作，2012，217（3）：47-52.
16. 赵辉，刘怀亮. 一种基于维基百科的中文短文本分类算法[J]. 图书情报工作，2013，57（11）：120-124.
17. 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.
18. 王细薇，张凯. 一种改进的基于共现关系的短文本特征扩展算法研究[J]. 河南城建学院学报，2012，21（4）：48-50.
19. 宁亚辉，樊兴华，吴瀚. 基于领域词语本体的短文本分类[J]. 2009,36（3）：142-145.
20. 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.
21. 张东娜，周春光，刘彦斌，郭东伟. 一种基于WordNet和Corpus Statistics的语义相似性计算方法[J]. 吉林大学学报（理学版），2010,48（05）：811-816.
22. 翟延冬，王康平，张东娜. 一种基于Word Net的短文本语义相似性算法[J]. 电子学报，2012,40（3）：617-620.
23. Zelikovitz S. Transductive LSI for short text classification problems[C]. In Proceedings of the 17th International FLAIRS Conference, 2004: 556-561.
24. Xuan-Hieu Phan, Le-MinhNguyen. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collection[A]. WWW[C], Beijing, China 2008, April 21-25.
25. QYAN X J, LIU G, LU Z, et al. Short text similarity based on probabilistic topics[J] . Knowledge Information System. 2010 25(3) : 473-491.
26. 胡勇军，江嘉欣，常会友. 基于LDA高频词扩展的中文短文本分类[J]. 现代图书情报技术，2013,234（6）：42-48.
27. 张志飞，苗夺谦，高灿. 基于LDA主题模型的短文本分类方法[J]. 计算机应用，2013，33(6)：1587-1590.