密级:[单击键入：涉密论文填写密级，公开论文不填写]

中国科学院大学
University of Chinese Academy of Sciences

硕士学位论文

基于迁移学习的微博分类研究

作者姓名：张帅

指导教师：王斌 副研究员

中国科学院计算技术研究所

学位类别：工学硕士

学科专业：计算机应用技术

研究所：中国科学院计算技术研究所

2013 年 5 月

Microblog Classification Based on Transfer Learning

By

Zhang Shuai

A Thesis Submitted to

University of Chinese Academy of Sciences

In partial fulfillment of the requirement

For the degree of

Master of Computer Application Technology

Institute of Computing Technology

May, 2013

声明

我声明本论文是我本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的地方外，本论文中不包含其他人已经发表或撰写过的研究成果。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作了明确的说明并表示了谢意。

作者签名：　　日期：

论文版权使用授权书

本人授权中国科学院计算技术研究所可以保留并向国家有关部门或机构送交本论文的复印件和电子文档，允许本论文被查阅和借阅，可以将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编本论文。

（保密论文在解密后适用本授权书。）

作者签名：　　导师签名：　　日期：

摘要

近年来，随着微博的快速发展，微博数据的规模呈现出爆炸式的增长，通过分类对海量微博数据进行有效组织变得越发重要。微博分类面临的一个问题是每天产生的大量新数据使得数据的统计性质不断变化，从而导致在旧数据上训练的分类器难以适用于新数据的分类。为了保证对新数据分类的效果，需要持续地对新数据进行标注，而这需要付出极大的人力物力成本。迁移学习技术旨在在不同但是相似的领域、任务或分布之间迁移知识。本文利用迁移学习的方法，使用新闻数据作为外部资源，减少人工标注的工作量，提升微博分类的效果。本文的工作从以下两个方面展开：

1，充分利用旧数据，同源迁移知识。本文从两个思路出发，探索从旧数据中迁移出知识用于新数据分类的方法。从时间因素出发，假设时间越近的数据越重要，我们采用了指数衰减的方法；考虑到微博中突发词和稳定词分布变化性质的不同，采用了选择性指数衰减方法。从数据分布因素出发，我们采用了迁移学习算法 TransferBoost，根据和新数据分布的相似性和分类效果调整旧数据的权重。实验结果显示，在标注数据较多的情况下，选择性时间衰减方法对分类效果有所提升；在标注数据较少的情况下，几种方法都难以取得理想的效果。

2，利用新闻数据，跨源迁移知识。根据新闻网站详细的类别体系，我们可以获得具有类别标签的新闻作为辅助数据。但是，新闻和微博在文字表达方面存在着很大差异，不能直接使用新闻数据对微博进行分类。本文基于参数先验的思想提出了迁移学习方法 WpriNaiveBayes，把在新闻数据上训练得到的参数作为微博模型参数的先验知识，依靠微博标注数据对这些先验知识进行修正。但是，基于参数先验的方法没有考虑词项间的差异，有些词项在新闻和微博中表现出相似的统计性质，迁移有关这些词项的知识可以对微博的分类提供帮助；如果迁移不相似的词项，则不能对微博分类产生帮助，甚至有负作用。据此，本文提出了基于可迁移度的迁移学习方法 WtrNaiveBayes，有选择的从新闻中迁移知识辅助微博分类。同时，同源迁移学习方法可以很自然地融入到 WtrNaiveBayes 框架中。实验结果表明，在微博标注数据较多的情况下，WpriNaiveBayes 取得了较好的效果；在微博标注数据很少的情况下，WtrNaiveBayes 方法的分类效果要优于基于先验的迁移学习方法和原始的朴素贝叶斯分类算法，在微博标注数据 5%的情况下，分类的 F1 值高于 90%。

关键词：微博，新闻，微博分类，文本分类，迁移学习，朴素贝叶斯

Microblog Classification Based on Transfer Learning

Zhang Shuai (Computer Application Technology)

Directed By Wang Bin

In recent years, with the rapid development of Microblog service, the scale of Microblog data grows explosively. The efficient organizing of the massive Microblog data by classification has become more and more important. One problem of Microblog classification is that millions of Microblogs are posted per day, making the statistical properties of data changing every day. So the classifier trained on the old training data can't be directly used in the new data. To get good classification result, we have to manually label the new data continually, which will cost much manpower and material resources. Transfer Learning technology aims to transfer knowledge between different but similar domain, task or distribution. This thesis utilizes transfer learning technology, exploiting the news as source data to reduce the work of manual labeling for Microblog classification. The works of this thesis are given as follows:

1, Exploit old labeled Microblog data and transfer knowledge intra data source. We explore two ways to take advantage of the old labeled data. In the light of time factor, we assume that newer instances are more essential for classifying the current instance and present an exponential decay method. Considering the difference of burst words and stationary words, we utilize a selective exponential decay method. In the light of data distribution, we utilize a transfer learning method called TransferBoost, which will increase the weight of old instances which obeys the similar distribution with the target data. We conduct experiments on SinaWeibo dataset. Experimental result shows that given enough training data, the selective exponential decay method performs better than the baseline. When lacking of training data, none of the methods achieves good result.

2, Exploit news data and transfer knowledge cross data sources. According to the detailed category hierarchy of news web sites, we can get new data with class label as auxiliary data. Because of the difference between news and Microblog in expression and choice of words, news data can't be used directly as training data for Microblog classification. First of all, we present a prior based method, which utilizes the parameters trained on the source data as the prior of the model for the target task. The model will adjust the prior knowledge according to the target data. The prior based method doesn't consider the difference of words. Some words shows similar statistical properties in Microblog and news, so transfer the knowledge of them from news to Microblog can be helpful. But the others may be useless or even harmful. To solve this problem, we propose a transfer learning method named Word Transfer Naïve Bayes (WtrNaiveBayes). Moreover, the intra source transfer method can be integrated into WtrNaiveBayes framework. The experimental results show that WtrNaiveBayes performs

significant better than the standard Naïve Bayes and the prior based method, in the circumstances that there are few labeled Microblog data. When five percent of the Microblog data are labeled, the WtrNaiveBayes method achieves 0.9 in macro F1 measure.

Keywords: Microblog, News, Microblog Classification, Transfer Learning, Naïve Bayes

目录

图目录

表目录

第一章 引言

近年来，微博在世界范围内的兴起，受到了社会各界的广泛关注，吸引了大量的互联网用户，已经成为了网民获取信息、传播信息的重要平台。微博内容的数量也随之呈现爆炸式增长，新浪微博日更新量已经超过 1 亿。面对海量的微博数据，通过自动分类的方式对微博进行主题划分变得尤为重要。

采用有监督的机器学习方法，在已标注的数据上训练得到分类器，可以实现自动对微博进行分类。但是由于每天都会有大量新微博内容产生，会导致训练数据和测试数据具有不同的统计性质，不能将在旧训练数据上得到的模型直接应用于新产生的数据上，这就需要持续地对新数据进行人工标注，重新训练模型，而这将耗费极大的人力物力资源。

为了减少微博标注的工作量，我们利用新闻数据作为辅助数据。一方面，微博和新闻具有话题相关性，微博上数据的变化在相关的新闻上会有所体现；另一方面，国内的新闻门户网站都有详细的类别体系对新闻进行归类。因此，我们可以源源不断的获得带有类别标注的新闻数据，用来辅助学习微博的分类模型。

但是，新闻和微博作为两种不同来源、不同形式的文本，有着重要的区别，表现在微博的长度更短，文字表达更口语化，话题变化更快等方面。因此，同一类别下的新闻和微博数据的分布是不同的，直接将新闻数据用于微博分类会带来错误的信息，难以取得满意的分类效果。

本文采用了基于迁移学习的方法，从新闻数据中学习到对微博分类有帮助的知识，用来提高微博分类的效果。在真实的微博数据和新闻数据上进行了实验，结果表明，在微博标注数据量较少的情况下，本文提出的迁移学习的方法效果要明显优于非迁移的方法，并且随着微博标注数据量的增加，分类效果可以获得稳定提升。

1.1 研究背景

微博（MicroBlog）是微型博客的意思，是一个基于用户关系的信息分享、传播以及获取平台，用户可以通过 WEB、WAP 等各种客户端组建个人社区，以 140 字以内的文字更新信息，并实现即时分享[57]。简单来讲，微博提供了这样一个平台，你既可以作为观众，在微博上浏览你感兴趣的信息；也可以作为发布者，在微博上发布内容供别人浏览[1]。

2006 年 7 月全球首家微博网站—Twitter[2]的创立标志着微博的出现，作为一个全新

[1] http://baike.baidu.com/view/1567099.htm
[2] http://www.twitter.com

的互联网交流平台，微博在全球快速发展。以新浪微博为代表[3]，国内的各大互联网公司相继推出了微博服务，包括腾讯微博[4]，搜狐微博[5]，网易微博[6]等。

图 1 显示了从 2010 年 12 月到 2012 年 12 月中国互联网微博用户规模的发展情况。根据中国互联网中心（CNNIC）2013 年 1 月发布的《中国互联网络发展状况统计报告》[15]，国内微博用户规模在 2012 年达到 3.086 亿，网民使用率达到 54.7%，并仍保持着 23.5%的年增长率。手机微博用户规模达到 2.02 亿，即高达 65.6%的微博用户使用手机终端访问微博，可见随时随地使用微博浏览和发布消息已经成为很多用户的习惯。美国总统奥巴马、创新工场董事长李开复，谷歌公司，微软公司等社会各界名人，组织机构都是微博用户。可见微博已在国内外获得了广泛的应用，已成为一种具有强大影响力的信息发布，获取和传播平台。

通过微博，随时随地任何人都可以成为信息传播者。微博打拐，微博反腐，微博募捐，微博招聘等一系列事件的发生让我们看到，微博的影响力已远远超过了一个社交网络或者是媒体平台，李开复[61]在《微博：改变一切》中指出，微博会对人们的生活方式、社交方式和商业模式发生的深刻改变。

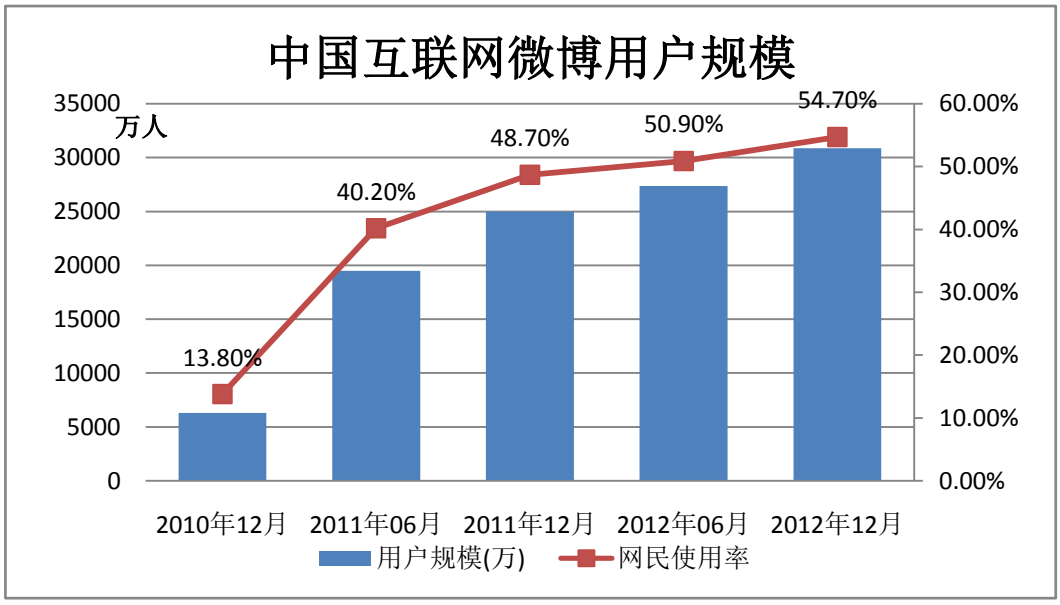


图 1 2010 年至 2012 年间中国互联网微博用户规模

微博的迅速发展引起了国内外学术界研究者的关注，出现了很多关于微博的研究方向。在图 2 中，简要地对目前微博研究的热点进行了分类。根据研究的侧重点不同，我们可以将对微博的研究可以分为两大类[57]：一种关注微博社交网络特性的研究，另一种则关注微博文本内容的研究。下面分别对这两类研究做简要的介绍。

[3] http://www.weibo.com
[4] http://t.qq.com/
[5] http://t.sohu.com/
[6] http://t.163.com/

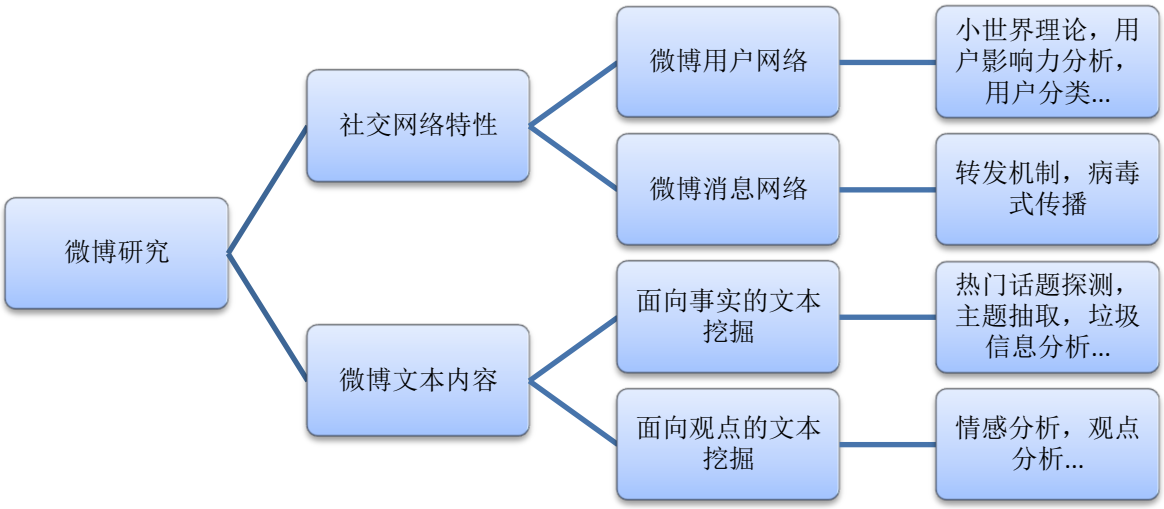


图 2 微博研究方向分类

1. 关注微博社交网络特性的研究

微博上的用户通过关注的方式建立起社交关系，形成了微博用户网络；用户发表消息，关注他的其他用户可以收到这些消息并通过转发来传播，形成了微博消息网络。对微博的社交网络特性研究分别从这两个方面展开。

微博用户网络方面，Java 等人[1]对 Twitter 的社会网络特性进行了初步的分析，结果表明 Twitter 表现出一定的小世界特性和幂律分布和小世界特性。Wu 等人[7]将 Twitter 的用户分为名人、媒体、博主和组织四个类别，发现相同类别的用户更容易存在关系，认为 Twitter 中用户间存在明显的互惠性。微博中用户影响力探测也是研究热点，研究者将在网页排序中使用的 PageRank，HITS 等算法应用于对 Twitter 社交网络图中用户影响力的分析[1][2]。Weng 等人[3]在 PageRank 的基础上，分析了用户的社交关系和所关注话题的相似度，找出话题相关的具有影响力的用户。Jie Tang 等人[36]提出了基于话题的个人影响力传播模型（Topical Affinity Propagation, TAP）来分析用户在话题级别的影响力。研究根据微博用户的社会网络特征对用户进行分类也是一个重要的研究内容，如 Krishnamurthy 等人[4]将用户分为广播者，普通人和垃圾虫三类。

微博消息网络方面，由于微博具有文本短，单向关注等机制，使得微博比一般的在线社交网络消息传播速度更快，范围更广。Kwak 等人[2]认为微博中消息传播最有效的方式是转发，通过构建转发树进行研究发现，用户获取到的大部分消息都是通过转发而非是直接接收的，并且微博中消息传播呈现出病毒式传播特点。Romero 等人[5]利用微博中的哈希标签，分析了不同类型话题的传播特性。Sadikov 等人[6]则研究了在消息传播过程中的信息丢失问题。

2. 关注微博文本内容的研究

微博不仅具有社交网络的特性，且微博每天产生的海量文本内容也具有很大价值，通过对微博内容进行挖掘，发现有价值的信息是面向微博文本内容研究的主要工作。根据文本挖掘任务的目标不同，可以分为面向事实的挖掘和面向观点的挖掘两类。

面向事实的文本挖掘包括用户标签推荐，热门话题挖掘等。Wu 等人[7]利用用户发表的微博内容，提出了一种 TextRank 算法自动给用户标注标签。Zhao 等人[8]利用上下文相关的 PageRank 算法从微博中提取出关键词，并根据这些关键词间的相关性和兴趣度，对微博中某时间段的特定话题进行自动摘要生成。Zhao 等人[9]还提出了对 LDA[55]模型进行了改进提出了 Twitter-LDA，并将 Twitter 和传统媒体纽约时报进行了对比。此外，根据主题对微博进行分类[23][31]、聚类，微博搜索等也属于面向事实的微博文本挖掘。

面向观点的文本挖掘主要是指根据用户发表的微博内容，挖掘出用户对特定主题的潜在情感或观点。网络用户每天发表的微博内容为情感分析提供了丰富的数据来源，微博内容短，结构自由等特点也给情感分析带来了挑战。文献[14]的研究表明，对微博进行情感分析要比博客更加有效。Go 等人[10]采用机器学习的方法，对微博的情感倾向进行分类（正面或负面）。Jansen 等人[11]的研究表明，19%的微博包含了对某产品或品牌的评论信息，并利用机器学习的方法对这些信息进行情感分类。还有研究人员利用微博来预测股票走势[12]，总统大选[13]等。

1.2 研究内容和思路

1.2.1 研究内容

基于上一节的研究背景，我们可以看出微博的重要性已经引起了学术界和工业界对微博研究的关注。本文研究的微博分类是微博研究的一个重要问题，属于微博文本内容研究中的面向事实的文本挖掘。

随着微博的流行，微博上产生的信息呈现爆炸式增长。根据 Twitter 公司发布的数据，Tweet 日更新量超过 3000 万。在国内，新浪微博注册用户也已突破 3 亿，微博的日更新量超过了 1 亿条。面对日益增长的微博内容，如何从海量的微博数据中高效的找出有价值的信息变得尤为重要。

根据主题对微博自动分类，是一种有效管理和利用海量微博信息的方法。微博分类对于微博系统的作用可以体现在以下两个方面：

首先，随着微博的发展，微博的用户活跃度也越来越高。用户每天都会受到大量的信息，信息种类是非常多样化的，包括了社会新闻，专业话题，兴趣爱好，生活心情，转发消息等等，而这些并不都是用户所感兴趣的。因此，对微博进行分类，可以给用户提供一种新的浏览方式，选择自己感兴趣的类别，免受无关信息的打扰，节省用户的操

作时间，提升用户体验。

另一方面，微博网站为了更好的满足用户的需求和自己的商业模式，需要向用户提供微博搜索功能，根据用户的兴趣给用户推荐好友和内容，精准的投放广告等，而这些应用都需要以用户发表的微博文本内容作为数据源。微博分类可以解决信息杂乱的现象，在微博主题挖掘、建立高效索引等方面都发挥着重要作用，是一个重要的基础技术。

由此可见，研究微博分类方法具有重要的应用价值。目前，通常使用有监督的机器学习的方法进行分类，其过程如图 3 所示。首先从数据源中选出一部分，由人工进行标注类别构成了训练数据集；然后采用机器学习的算法在训练集上学习到一个分类器；将未标注的数据或新产生的数据输入分类器，就可以得到这些数据的类别标签即分类结果。

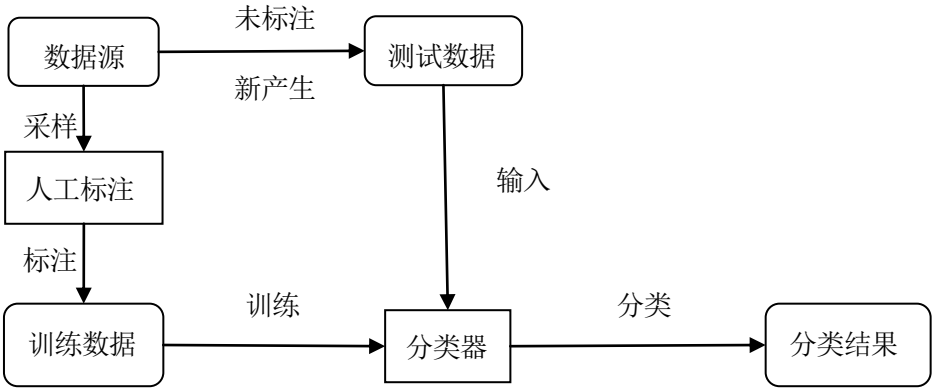


图 3 监督学习的分类过程

这种监督学习方法的假设是：训练数据和测试数据是服从相同分布的。这样的话，在训练数据上训练得到的分类器才能适用于对测试数据的分类。但是这种假设对于微博数据却很难成立。

微博平台每天产生的大量新数据导致微博数据的统计性质不断变化，这种变化主要体现在两个方面：类别分布的变化和词项在类别中分布的变化。

1. 类别分布的变化

类别分布是指因为人们在不同时间，对类别的关注和讨论程度是不同的。比如在 2013 年 3 月份两会期间，关于政府政策、经济民生，环境问题等在微博上引发了广泛地讨论，“政治”类的微博的数量的比例则会比以往上升很多，相应的其他类别的比例会下降。类似地，在奥运会、世界杯等热点体育赛事会导致“体育类”上升；苹果、三星等著名公司召开新款手机发布会时导致“手机科技类”上升；股市大涨、大跌或有重要政策出台时“股票类”会上升等。

2. 类别中词项分布的变化

随着时间的变化，同一个类别中词项出现的概率也在发生变化。因为同一个类别下，微博上讨论的具体话题或事件变化是很快的，这样也会导致相关的词项概率变化。

我们以类别“手机科技类”和词项“锤子”为例进行解释。“锤子”本意是一种带柄的锤击工具，在四川或陕西方言里会有对人或事的不屑或不认同意思。“锤子”本身跟“手机科技类”关系很小，所以它在该类别下出现的概率很小。但是在 2013 年 3 月 27 日，发生了很大改变。原因是微博名人罗永浩创立了一家科技公司，该公司开发了一款基于 android 的手机操作系统，命名为“锤子”。该系统在那天召开了发布会，该话题在微博上热门起来，这也导致在“手机科技类”下，“锤子”这个词出现的概率迅速上升。

表 1 显示了 2013 年 1 月 27 日，2 月 27 日和 3 月 27 日三天在新浪微博搜索“锤子”的结果，可看出，在 3 月 27 日，搜索结果的数量从一两个月前的 400 条左右，急剧上升到 32 万多条。而在搜索结果的前 20 条中，与“手机科技类”相关的从 1 条，6 条上升到了 18 条。由此可见，在“手机科技类”下“锤子”出现的概率发生了很大变化。

表 1 “锤子”在 2013 年 1 月、2 月、3 月 37 日的新浪微博搜索结果统计信息

	1 月 27 日	2 月 27 日	3 月 27 日
搜索结果总数	448 条	422 条	322896 条
Top20：手机 ROM	1 条	6 条	18 条
Top20：工具含义	4 条	2 条	0 条
Top20：四川方言	15 条	12 条	2 条

通过上面的分析可以看出，微博上的数据是在不停变化的，不能确保新数据和旧数据服从相同的分布，也就是说旧数据很容易过时。如果仍用原来的模型对新数据进行分类，则会导致分类效果的下降。目前解决数据变化的方法都需要对新数据进行人工标注，然后采用实例选择，实例加权等方法，利用较新的数据，重新训练分类器。

众所周知，人工标注工作需要耗费很大的人力物力，在微博分类的应用场景下，人工标注的瓶颈问题愈发凸显。大量新的微博不断产生，数据分布变化极快，这就需要持续有新的标注数据补充进来，保证对新数据的分类效果。一方面持续的人工标注要耗费很大的代价，另一方面受人工标注速度的限制，分类系统不能快速对新数据的特点做出响应。

因此，本文研究一种迁移学习的方法，利用新闻数据作为外部数据，面对微博数据的变化，减少人工标注的工作量，提高分类效果。

1.2.2 研究思路和目标

面对微博数据变化快易过时的特点，文本采用迁移学习的方法，利用有类别标注的新闻数据作为外部数据，用来辅助训练微博分类器，减少人工标注的工作量，提高分类效果。

迁移学习技术的目标在于在不同但是相似的领域之间迁移知识[16]，其基本思想来源于人类学习知识的过程，在一个领域学习到的知识，对于相似领域的学习是有帮助的。鉴于微博数据分布变化快，易过时，迁移学习的方法很适合微博分类的应用场景。如图

4 所示，本文研究从以下两个方面进行知识的迁移：

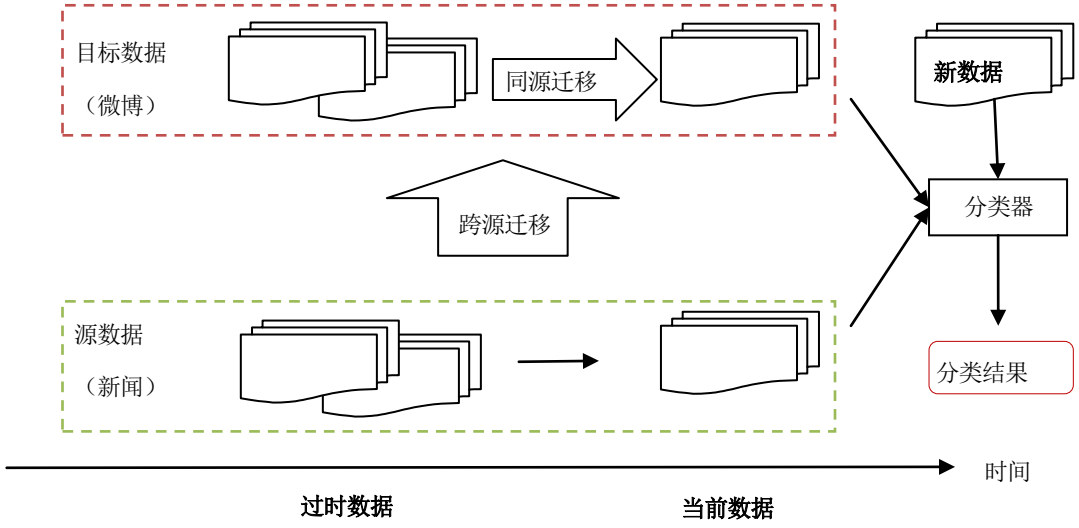


图 4 基于迁移学习的微博分类方法示意图

1. 同源迁移：利用旧的微博数据

新的微博数据和旧微博数据在分布上发生了变化，直接使用旧数据难以达到好的效果，但是他们之间仍然有很多共通的知识，有效利用旧数据可以提高在新数据的分类效果。因为新旧微博数据是相同来源的数据，因此，我们称为同源迁移学习方法。

2. 跨源迁移：利用外部的新闻数据

新闻网站如新浪[7]、网易[8]、腾讯[9]等都有着详细的类别体系，方便用户浏览新闻，如图 5 为新浪网类别导航页面截图。我们可以利用该类别体系，通过网络爬虫，就可以源源不断地获取有标注的新闻数据。由于同一个类别的新闻和微博之间存在着相关性，我们可以利用同一时间段的新闻数据作为源数据，微博数据作为目标数据，采用迁移学习的方法提高对微博数据的分类效果。因为新闻和微博是不同来源的数据，我们称从新闻到微博的迁移学习方法为跨源迁移学习方法。

[7] http://www.sina.com.cn
[8] http://www.163.com
[9] http://www.qq.com

图 5 新浪网导航页面截图

微博数据是不断产生动态变化的，新闻数据也是不断产生，动态变化的，二者之间有关联也有区别。微博和新闻具有话题相关性和分布差异性。

微博内容随着用户新内容的发表而发生变化，微博上的热门话题也随之发生改变。我们认为，新闻和微博具有相关性，一方面，因为很多微博上的热门话题是由某些新闻事件引发的评价和讨论，如“房产税”，“国五条”等。另一方面，从微博上产生的热门话题，也会及时的被网络新闻所报导和关注，如“郭美美”，“房叔”，“PM2.5”等。

微博和新闻相比，文本更短，在表达上更加简单自由，不会讲究句法句式标点和语言的完整性，有时仅仅是几个词或短语，也会经常出现网民创造的新词汇。而新闻在表达上则要正式规范很多，措辞上也会较多采用书面语。

总结上述内容，历史微博数据虽已过时，但仍具有很大的使用价值；外部新闻数据和微博有着话题相关性，但是也有很大的差异，特别在用词和文字表达方面。

本文研究的目标是，综合利用历史微博数据和外部新闻数据，提出一种基于迁移学习的方法，既可以这些数据中学习到知识，又不被他们和当前微博数据之间的差异所误导，减少对微博标注数据的依赖，提高分类的效果。

1.2.3 文中所用符号和概念

下面对本文中所用到的概念和符号进行解释：

表 2 文中所用的概念和符号

概念	符号\缩写	解释
文档集	D	包括微博数据集、新闻数据集等
文档	d	在本文中，微博和新闻都被看做是文档
词典	V	文档集中所有词的集合
词	w	词，词语，词项，单词
源领域	S	迁移学习中，知识来源的领域或任务

目标领域	T	迁移学习中，知识的目标领域或任务
同源迁移	Transfer intra source	在相同的来源、不同时间的数据间迁移知识，本文主要是从旧的微博、新闻向新的微博、新闻迁移知识。
跨源迁移	Transfer cross source	在不同来源的数据间迁移知识，本文中知识从新闻向微博迁移

1.3 本文的主要工作和贡献

为解决微博数据更新快带来的微博标注问题，减少标注工作量，提高分类效果，本文主要做了以下几个方面的工作：

1. 目前针对微博的研究主要在 Twitter 上，很少在国内的中文微博数据上进行研究，为了能得到最真实的研究结果，我们采集了新浪微博数据和新浪新闻网易新闻的真实数据，根据微博的标签构造了六个类别的微博数据集。并且分析了微博数据中类别分布的变化和词项分布的变化。分析了微博和新闻的相关性和用词的差异性。
2. 研究适合微博的同源迁移学习方法，以有效利用旧的微博数据。从时间的角度出发，采用了选择性指数衰减的分类算法，从分布相似性的角度出发，采用了基于 Boosting 的迁移学习算法 TransferBoost。实验结果表明，在训练数据较多时，选择性指数衰减方法对分类效果有所提升，但是在训练数据较少时，实验的几种算法的效果均不理想。
3. 根据微博和新闻的联系和区别，提出了 2 种跨源迁移学习方法，并将同源迁移学习方法融合到跨源迁移学习的框架中。实验证明，在标注微博数据量很少的情况下，基于可迁移度的迁移算法 WtrNaiveBayes 要明显优于非迁移学习的方法，在标注数据占 5%的情况下，分类 F1 值超过了 90%。并且随着标注数据的增加，分类效果可以获得稳定提高。

1.4 论文的组织

第一章介绍了本文的研究背景，研究内容和思路，总结了本文的主要工作和贡献。

第二章从三个方面介绍了本文的相关工作，首先介绍了文本分类技术，包括文本分类的过程，算法和评价方法，重点介绍了本文用到的朴素贝叶斯分类法。然后介绍了迁移学习的技术，包括基于实例的迁移学习，基于特征表示的迁移学习，基于参数的迁移学习和基于相关关系的迁移学习。最后介绍了目前微博分类的研究现状以及迁移学习和微博挖掘相关的工作。

第三章讲述微博的采集、预处理和数据分析。首先介绍了数据集构建的方法，包括类别标签集合的定义，新浪微博数据和新闻数据的采集策略。然后介绍了微博预处理过

程，在微博文本表示方面采用了词袋模型，介绍了分词的方法及效果。最后，从微博数据的变化和微博与新闻数据之间关系两个方面进行了数据分析，数据显示微博数据分布变化很快，微博和新闻存在着话题相关性和词项分布差异性。

第四章介绍同源迁移学习方法，利用旧的微博标注数据，提高对新微博的分类效果。从时间因素和数据分布因素两个方面出发，比较了时间衰减方法、选择性时间衰减方法和基于实例的迁移学习方法 TransferBoost。实验结果表明，在微博标注数据较充分的情况下，选择性时间衰减方法可以提升微博分类的效果。TransferBoost 适用于源领域和目标领域差别较大的情况下，不适合本文的同源迁移场景。

第五章介绍跨源迁移学习方法，利用新闻数据作为辅助数据，提高微博分类的效果。本文提出了两种跨源迁移学习算法，分别是基于参数先验的方法 WpriNaiveBayes 和基于可迁移度的方法 WtrNaiveBayes。实验结果显示，WtrNaiveBayes 在微博标注数据量很少的情况下可以显著提高微博分类的效果，达到减少人工标注工作的目的。

第六章对本文的工作做了总结，并且从类别先验概率的估计、词项间关系的建模和新闻类别标签的限制三个方面提出了下一步的研究方向。

第二章 相关工作

本文主要利用迁移学习的方法研究微博分类的问题，本章将从以下 3 个方面介绍相关的研究工作：文本分类技术，迁移学习技术和微博分类的研究现状。

2.1 文本分类技术

文本分类中，给定文档$d \in \mathbb{X}$和一个固定的类别集合$\mathbb{C} = \{c_1, c_2, \ldots, c_K\}$，其中$\mathbb{X}$表示文档空间，$\mathbb{C}$表示类别空间，类别（class）也通常称为 category 或 label。训练集合是指有类别标签的文档，如$< d, c > \in \mathbb{X} \times \mathbb{C}$表示文档 d 属于类别 c。某种分类方法或者学习方法是指我们希望从训练数据中学习到某个分类函数γ，它可以将文档映射到类别[61]。

$$\gamma: \mathbb{X} \to \mathbb{C}$$

文本分类系统的一般过程如下图 6 所示：

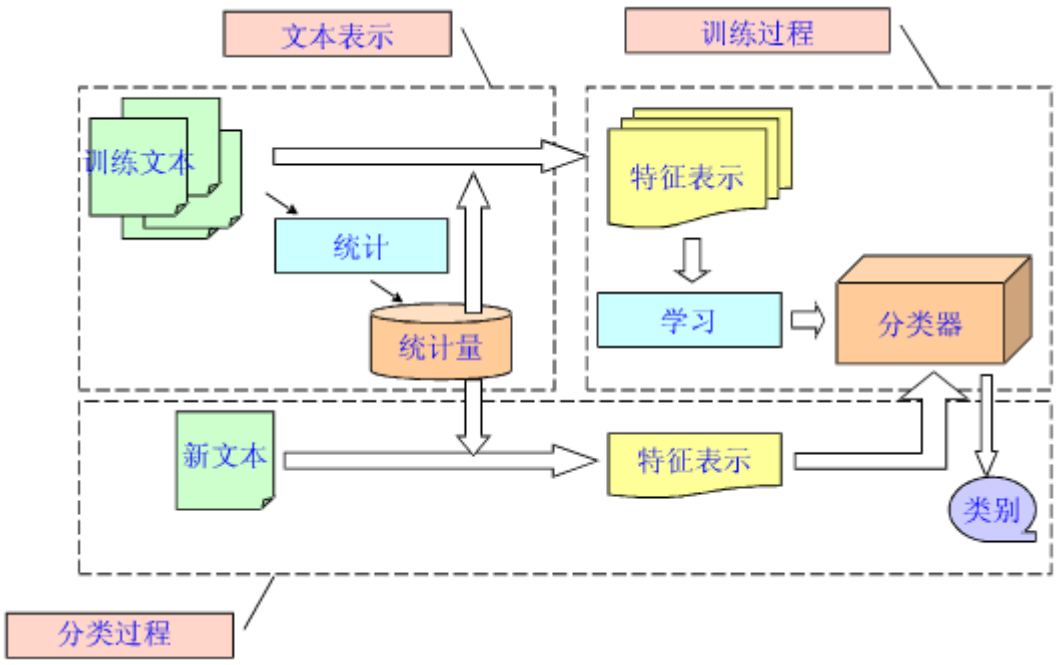


图 6 文本分类过程

从上图我们可以看到，文本自动分类一般包括 3 个阶段，首先需要把文本表示成计算机可以处理的格式，这个过程称为文本表示。文本表示需要选出某些重要的特征（Feature）来表示原始的文本，通常通过从文本集中构造一些统计量，利用他们筛选出特征，将文本表示成特征向量。然后分为训练和分类两个阶段。在训练阶段，使用有类别标注的训练数据，采用机器学习的方法学习到一个分类器。在分类阶段，输入新的未标注的文本，分类器会自动输出该文本的类别。一般的文本分类系统会包括文本预处理，

文本表示，特征降维，分类器训练和预测，评价分类效果等部分，下面对这些技术分别做简要的介绍。

2.1.1 文本预处理

文本预处理主要是从语料库中提取出主要的内容，过滤掉一些会对分类产生干扰的噪音。通常包括字母大小写转换、词干还原、中文分词、去停用词等。

对于英文文本，大小写的字母一般并没有区别，将他们都统一转换为小写或大写可以简化操作。另外在英文中，出于语法上的要求，文档中常常会使用词的不同形态，如organize，organizes 和 organizing。词干还原为了减少词屈折变化的形式，将派生词转化为其基本形式，这样也可以压缩词空间。

对于中文文本，词语之间并没有空格或其他的分隔标记，因此需要通过分词的算法将连续的文本转换为一个个的词语。分词的算法有基于词典的方法、基于统计语言模型的方法等。本文中采用了 Lucene 中的中文极易分词法。

另外，有些常见的词语在所有的文档中频繁出现，但是其本身并没有任何语义上的含义，对于区分文档类别也没有任何帮助，如中文中的“的”、“地”、“得”，英文中的“am”、“a”、“the”等。通常采用构造一个停用词表的方式，将这些词过滤掉，可以提高分类的速度和效果。

2.1.2 文本表示和特征降维

向量空间模型（Vector Space Model， VSM）[56]是最常用的文本表示模型，将每个文本表示成向量空间中的一个向量，空间的每一个维度代表一个特征，文本向量的每一维的取值代表该特征在文本中的权重。根据前人的研究结果，向量空间模型在文本分类中取得了很好的效果。进行文本表示的重要工作是选取特征以及计算特征权重的方法，下面分别做简要介绍。

特征不一定是词语，可以有不同的粒度。特征的最基本单位是字符，如中文中的字、英文中的字母。词是使用最广泛的特征，如中文文本分词后得到的词语，英文中直接用空格分开的单词。N 元组（N-gram）并不直接对汉语进行分词，而是用连续的 N 个字作为特征的基本单位，如二元组表示“中华人民共和国”为“中华，华人，人们，民共，共和，和国”，N 元组的方式避免了分词的难题，但是其不足在于计算量大，噪声多，容易过拟合。短语级别的特征比词语级别粒度更大，它考虑到了词语之间的关系，可以在一定程度上提高特征的语义含量，但是也会使得向量更加稀疏。

常用的特征权重计算方式有布尔权重，词频权重，TFIDF 权重等。布尔权重是用布尔表达式计算特征的权重，特征出现权重为 1，不出现则为 0。布尔权重实现简单，但是没有考虑特征出现的次数，无法体现特征的重要程度。词频权重（TF）以特征项在文档中出现的次数计算特征的权重，出现的频率越高，越重要。TFIDF 根据词频（TF）和反向文档频率（IDF）来计算特征的权重，DF 是文档集中包含该特征的文档数，IDF 是 DF

的倒数，IDF 越大说明 DF 越小，即特征出现在文档集中较少的文档中，具有较好的区分性。TFIDF 兼顾了特征在文档中的重要度和特征的区分能力两个方面，常用的 TFIDF 计算方法如公式所示。

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

采用向量空间模型，采用词或字做特征的话，特征的总数将是整个词典的大小，导致原始特征空间会达到几十万维甚至更高。高维的特征空间一方面会增加分类算法学习的时间和空间代价，另一方面，原始特征空间会包含很多噪声，影响分类的准确性。因此，在文本分类前常常需要对原始特征空间进行降维。特征降维可以分为特征选择和特征抽取。特征选择指根据某个准则从原始特征空间中直接选出最优的特征子集，经典的特征选择方法有文档频率，信息增益，互信息，卡方统计量等。特征抽取指通过某种方法构造一个从原始的特征空间到维度较低新的特征空间的映射，如从词项的特征空间映射到主题空间，特征抽取的主要方法有主成分分析（Principal Component Analysis，PCA），潜在狄利克雷分配（Latent Dirichlet Allocation，LDA），潜在语义索引（Latent Semantic Index，LSI）等。

2.1.3 文本分类算法

在文本挖掘领域，研究者已经对文本分类方法进行了广泛地研究。早期的文本分类方法采用知识工程的方法，由领域专家手工制定一系列的规则来进行分类。这种方法的缺点在于可推广性差，适用于一个领域的规则集很难推广到另一个领域，而且专家制定规则的代价极高。后来，研究者采用机器学习的方法进行文本分类，常用的算法有朴素贝叶斯分类器（Naïve Bayes Classifier）[40]，Recchio 分类法[38]，K 近邻[39]，支持向量机（Support Vector Machine）[41][42]，决策树算法[43]，以 AdaBoost 为代表的 Boosting 算法[44][45][46]等，这些算法在传统的文本分类问题上取得了不错的效果。

本文所提出的方法均是建立在朴素贝叶斯分类法的基础上，因此在本节先对朴素贝叶斯分类法的模型，参数估计和平滑方法进行简要的介绍。

朴素贝叶斯（Naïve Bayes）分类法是基于贝叶斯定理和特征的条件独立假设的分类方法[60]。对于给定的训练数据集，首先基于特征条件独立性假设学习输入/输出的联合概率分布；然后基于此模型，对给定的输入 x，利用贝叶斯定理求出后验概率最大的输出 y。朴素贝叶斯分类法实现简单，学习和预测都非常高效，并且在文本分类领域取得了不错的效果，是一种常用的分类算法。

1. 基本方法

下面将简单介绍文本分类中常用的多项式朴素贝叶斯模型，它是一种基于概率的学

习方法[60]。根据贝叶斯公式，文档 d 属于类别 c 的概率计算方法如下：

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)}$$

其中$P(c)$是文档出现在类别 c 的先验概率，$P(d|c)$是给文档 d 出现在类别 c 的条件概率。$P(d)$是文档 d 出现的概率，对于同一篇文档 d，$P(d)$是固定不变的，由此可得：

$$P(c|d) \propto P(c)P(d|c)$$

因为在朴素贝叶斯分类法中，假设文档中的词项w_k之间是相互独立的，故有：

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(w_k|c)$$

其中$P(w_k|c)$是词项w_k出现在类 c 的文档中的条件概率。$<w_1, w_2, w_3, \dots, w_{n_d}>$是文档 d 中的词项，$n_d$是 d 中词项的数量。上述公式表示，在词项条件独立性假设的前提下，联合条件概率等于条件概率的乘积。

在文本分类中，我们的目标是找出文档 d 最可能属于的类别 c，对于朴素贝叶斯法，最可能的类别是具有最大后验概率(Maximum a posteriori, MAP)估计的结果c_{map}：

$$c_{map} = \underset{c \in \mathbb{C}}{\operatorname{argmax}} \widehat{P}(c) \prod_{1 \le k \le n_d} \widehat{P}(w_k|c)$$

由于我们不知道参数的确切值，因此，在上述的公式中，采用了从训练数据中估计出的参数$\widehat{P}$来代替 P。在公式中，对所有的$1 \le k \le n_d$，计算其条件概率的乘积，这可能会导致浮点数下届溢出。因此，更好的方法是引入对数，进而转变为多个概率的对数和，因为对数函数是单调递增的，可得：

$$c_{map} = \underset{c \in \mathbb{C}}{\operatorname{argmax}} [\log \widehat{P}(c) + \sum_{1 \le k \le n_d} \log \widehat{P}(w_k|c)]$$

2. 参数的估计

对于朴素贝叶斯分类法，学习过程就是要估计参数$P(c)$和$P(w_k|c)$，可以采用极大似然估计法来估计参数，先验概率$P(c)$的估计值为：

$$\widehat{P}(c) = \frac{N_c}{N}$$

其中N_c是训练集合中类别 c 的文档数，N 为训练集合的文档总数。

$$\widehat{P}(w|c) = \frac{T_{cw}}{\sum_{w' \in V} T'_{cw'}}$$

其中T_{cw}是词项 w 在训练集合类别 c 中出现的词频，V 是词汇表。在对每篇文档计算时，使用的是其在文档中多次出现的总次数，w 出现的位置不被考虑。

用极大似然估计可能会出现概率值为 0 的情况，为了避免 0 概率造成的偏差，常采

用的方法是贝叶斯估计，贝叶斯估计的一个特例是加一平滑，或称为拉普拉斯平滑（Laplace smoothing），即在每个词项出现的次数上加 1，即每个词项都多出现了一次：

$$\widehat{P}(w|c) = \frac{T_{cw} + 1}{\sum_{w' \in V}(T'_{cw'} + 1)} = \frac{T_{cw} + 1}{\sum_{w' \in V} T'_{cw'} + |V|}$$

其中$|V|$为词典大小，类似的，对类别先验概率的拉普拉斯平滑估计为：

$$\widehat{P}(c) = \frac{N_c + 1}{N + |C|}$$

其中$|C|$为类别标签的个数。

2.1.4 分类效果评价指标

分类效果的评价是一个重要问题，针对不同的应用需求，人们提出了很多不同的评价方法。最常用的评价指标有正确率（Precision），召回率（Recall）和 F 值（F-Measure）。

对于某一类别而言，正确率是指分类器判断为某类的样本中判定正确的样本所占的比例，召回率是指分类器正确判断为某类别的样本占该类别所有样本的比例。用表 3 表示分类结果，正确率和召回率计算公式如下:

$$\text{Precision：} P = \frac{A}{A + B}$$

$$\text{Recall：} R = \frac{A}{A + C}$$

表 3 分类结果关联表

	属于此类	不属于此类
判定属于此类	A	B
判定不属于此类	C	D

实际应用中，一种分类方法很难同时提高正确率和召回率，两者常常是此消彼长的关系。一个融合了正确率和召回率的指标是 F 值，它是两者的调和平均值，F 值的计算公式如下：

$$F_\beta = \frac{(1 + \beta^2)P \times R}{\beta^2 P + R}$$

其中参数β来控制两者的重要度，$\beta < 1$强调正确率，$\beta > 1$强调召回率。$\beta = 1$时，它们权重相等，常记为F_1。在本文中，使用 F1 作为主要评价指标。

上述几个指标的定义是针对一个类别的，当具有多个类别时，需要给分类效果一个综合的评价。通常有宏平均和微平均两种做法，宏平均的做法是先对每个类别求出上述的指标，然后求这些指标的平均值。微平均是先将所有类别的分类结果融合到一起，得到一个总的分类结果邻接表，然后再计算上述指标。宏平均的做法是考虑每个类别的权

重是一样的，微平均则认为每个样本的权重是一样的，微平均的结果会偏重样本多的类别。

2.2 迁移学习技术

传统的有监督的机器学习或数据挖掘算法都是建立在一个假设上的，那就是训练数据和测试数据在相同的特征空间下服从相同的数据分布。然而，在实际应用中，这种假设并不总是成立的。

迁移学习的思想源于人类学习知识的过程，以前学到的知识，对于学习新的类似的问题将会有帮助，如图 7，学习过 C 语言的人，再学习 Java，因为有了编程基础，不需要从零学起。但是 C 语言中的语法知识对于写诗这个任务来讲，可能就没有多少帮助了。迁移学习的目标在于在不同但是相似的领域、任务、分布之间迁移知识[16]，利用从源问题中学到的知识，帮助解决目标问题。

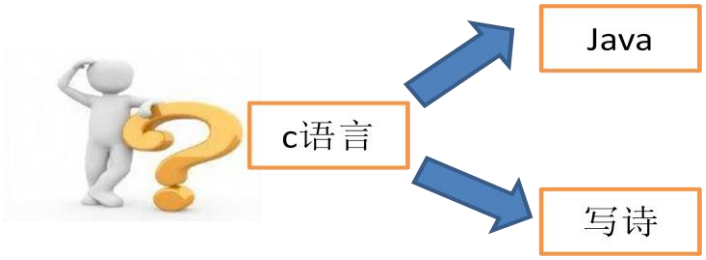


图 7 迁移学习的基本思想示例图

迁移学习的目标是从源领域向目标领域迁移知识，只关心目标领域任务的执行效果，源领域作为一种辅助。

根据迁移学习的应用情景不同，可以将迁移学习分为归纳式迁移学习，直推式迁移学习和无监督迁移学习，如表 4 所示。归纳式迁移学习源领域和目标领域的任务可能不同，需要有目标领域的标注数据，利用源领域标注或未标注数据，提高源领域预测效果；直推式迁移学习源领域和目标领域的学习任务相同，没有目标领域的标注数据，但是有源领域的标注数据，利用源领域的标注数据对目标领域数据进行预测；无监督迁移学习源领域和目标领域都没有标注数据，任务通常是聚类，降维等无监督学习任务。

表 4 归纳式、直推式和无监督迁移学习的设置

	相关领域	源数据是否有标注	目标数据是否有标注	任务
归纳式迁移学习	多任务学习	有	有	分类，回归
	自学习	无	有	分类，回归
直推式迁移学习	领域适配，样本选择偏置	有	无	分类，回归
无监督迁移学习		无	无	聚类，降维

根据迁移的知识不同（即 What to transfer？）可以将迁移学习分为四种，基于实例的迁移学习，基于特征的迁移学习，基于参数的迁移学习，基于相关关系的迁移学习，下面分别介绍这四类方法。

2.2.1 基于实例的迁移学习

基于实例的迁移学习在源领域和目标领域之间迁移的是数据样本的实例，这类方法是建立在这样的假设基础上的：虽然由于源领域和目标领域的差异，不能将源领域的数据直接应用在目标领域，但是，在源领域中存在部分数据是适合应用于目标领域学习问题的，可以通过实例加权的方法提高这些实例的权重或者使用重要性采样的方法将这些实例从源数据中甄选出来。

戴文渊等人[17]提出一种基于 Boosting 的迁移学习算法 TrAdaBoost，假设目标领域和源领域的特征空间和类别空间都相同，但是数据的分布不同。TrAdaBoost 对 AdaBoost 框架进行了修改，通过自适应地增加对目标领域分类效果有利的源领域实例权重，降低有损效果的实例的权重，提升对目标领域分类的效果。

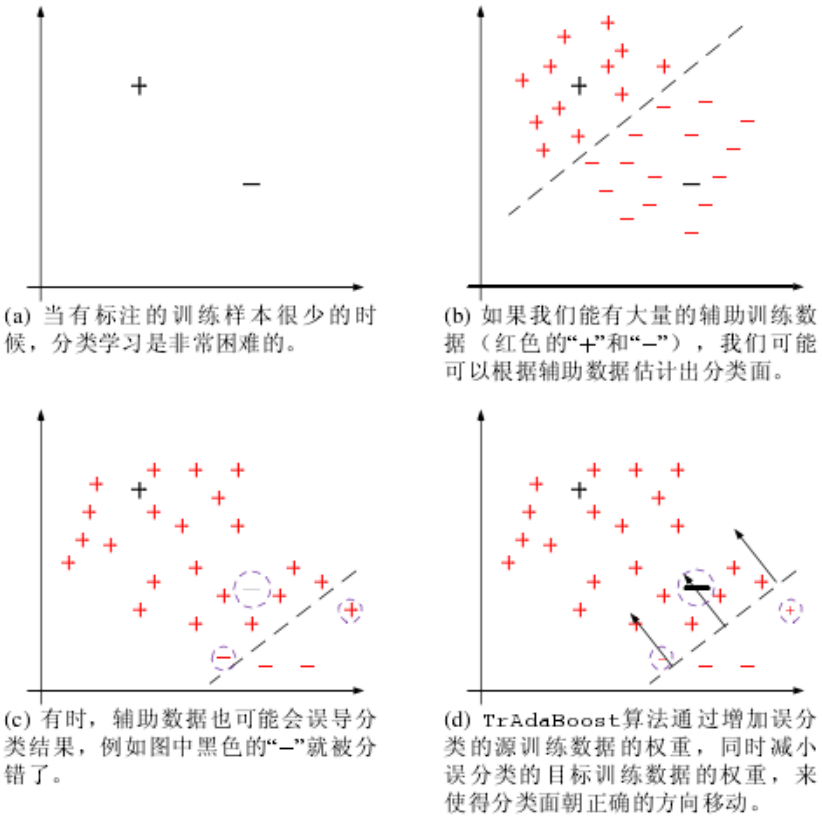


图 8 TrAdaBoost 方法工作基本思想的直观示例

如图 8 是一个直观的示例，展示了 TrAdaboost 算法工作的基本思想。随后 Eaton 等人[68]提出了基于任务选择性迁移的 TranferBoost 方法，对 TrAdaboost 进行了改进，在迭代过程中，针对不同的源领域任务和目标任务的可迁移力(transferbility)，改变该任务中所有样本的权重，同时根据样本在当前分类器上分类的正确性，调整样本的权重。从而使得和目标领域相似的源领域样本权重提升，不相似的样本权重降低。而在 TrAdaboost 中，只会降低源领域中和目标领域不相似样本的权重，并不会提升相似样本的权重。

Jiang 和 Zhai[47]提出了一种启发式的方法，根据实例在源领域出现的条件概率

$P(Y_S|X_S)$和在目标领域出现的条件概率$P(Y_T|X_T)$的差异，移除掉源领域中会对目标领域产生误导的实例。Liao 等人[48]提出了一种主动学习的方法，在源领域数据的帮助下，从目标领域中选择出部分实例进行标注。

迁移实例的方法的优点在于方法简单，效果显著，缺点在于只能适用于源领域和目标领域很相似的情况，要求源领域和目标领域的类别空间和特征空间都相同，且基于迭代的训练过程效率不高。当应用场景中源领域中的部分实例符合目标领域的分布时，适合使用基于实例的迁移学习方法。

2.2.2 基于特征表示的迁移学习

基于特征表示的迁移的方法为目标领域学习一个“好的”特征表示方法。这类方法的特点在于将不同领域的迁移的知识被编码为学习到的特征表示，采用这个“好的”特征表示方法，可以更好地解决目标领域的任务，性能上有所提升。在归纳式迁移学习问题中，“好的”特征表示方法，目标在于可以最小化源领域与目标领域数据的差异或者最小化回归误差。

当有大量标注的源领域数据时，常采用有监督学习的方法来寻找这样的特征表示方法，其基本思想和多任务学习（MultiTask Learning）有些相似，即学习到一个在多个相关的任务中共享的低维的特征表示。Argyrious 等人[70]提出了一个稀疏特征学习的方法，通过最小化在源数据和目标数据的训练误差，在损失函数中加上正则项来学习到稀疏的特征表示。

若源领域数据没有标注，则采用无监督学习的方法来构建特征表示，Raina[63]等人提出了一种稀疏编码的方法，使用无监督的方式来学习源领域和目标领域之间的高层次的特征。该方法分两步进行，第一步利用源数据学习出一组基元素(bases)，用这些基元素作为特征空间。优化的目标是学习一组稀疏的激活系数，使所有源数据尽量能被基元素来恢复。优化函数如下：

$$\min_{a,b} \sum_i \left\| x_{S_i} - \sum_j a^j{}_{S_i} b_j \right\|_2^2 + \beta \|a_{S_i}\|_1$$

$$\text{s.t.} \|b_j\|_2 \le 1 \; \forall j \in 1, \dots, s.$$

第二步使用将目标数据用基元素来表示，通过解下面的优化函数可以得到目标数目在基元素空间的系数。

$$a^*_{T_i} = \underset{a_{T_i}}{\operatorname{argmin}} \|x_{T_i} - \textstyle\sum_j a^j{}_{T_i} b_j\|_2^2 + \beta \|a_{T_i}\|_1$$

基于特征表示的迁移学习方法，基本思想是通过目标数据和源数据，学习到一系列隐含的“主题”，目标数据和源数据都在“主题”的空间下表示，把源数据中的知识通过学习到的“主题”来表示，迁移到目标数据中。适用于只有少量的目标领域标注数据，但是有大量的源领域未标注数据的情况。

2.2.3 基于参数的迁移学习

大多数基于参数的迁移学习方法任务源任务和目标任务的模型之间，应该共享一些参数或者超参数的先验分布。在领域之间迁移的知识就被编码为这些共享的参数或超参数的先验，通过发现它们，知识在不同的领域间得到了迁移。

很多参数迁移的方法原本是针对多任务学习任务提出的，包括基于正则化框架的和层次贝叶斯框架的方法。多任务学习和迁移学习的区别在于多任务学习同时关注多个任务上的学习效果，而迁移学习只关注目标任务，因此多任务学习的方法很容易应用到迁移学习上。

Lawrence 和 Platt[71]提出了基于高斯过程的 MT-IVM 算法来处理多任务学习情况，MT-IVM 算法通过共享高斯过程先验知识在多个任务之间学习高斯过程的参数。Bonilla 等[72]也在高斯过程的范畴下对多任务学习进行了研究，他们提出了在多个任务基础上使用自由形态的协方差矩阵来对任务之间相互依赖关系建模，高斯过程先验知识用来归纳任务之间的相关性。Schwaighofer 等[73]提出了层次贝叶斯框架和高斯过程相结合的方法来进行多任务学习。

在自然语言处理领域，迁移学习有时会被称为领域适配（domain adaptation），下面介绍几种领域适配的方法。

Chelba 和 Acero[69]提出了一种基于先验的方法，其基本思想是，现在源领域数据上进行训练，得到模型 S。然后在目标领域数据上进行训练模型 T 时，讲模型 S 的参数作为先验。他们将这种方法应用在最大熵分类器上，具体做法时，在训练模型时，为了避免过拟合需要加上一个正则项$\lambda\|w\|_2^2$，将模型 S 的参数w^S作为先验的方法是将正则项做如下修改$\lambda\left\|w - w^S\right\|_2^2$，其含义是，当目标数据不反对时，模型 T 的参数倾向于和模型 S 的参数接近。

Daumé III 和 Marcu [18]提出了一种基于最大熵分类器的领域适配方法。其核心思想是，学习 3 个独立的模型，一个是源领域专属模型（Source Specific Model），一个是目标领域专属模型（Target Specific Model），一个是通用模型（General Model），分别捕获源领域特有的，目标领域特有的，和两者通用的信息。通过这种方法，对目标领域和源领域的共性和区别进行了建模。处理目标领域数据时，使用目标领域专属模型和通用模型。该方法在效果上要优于基于先验的方法，但是该方法的缺点也很明显，实现复杂，训练效率低，不适用于大规模的数据。

Daumé III[19]后来提出了一种非常简单的基于特征扩展的方法用于领域适配。该方法将特征空间扩充为原来的 3 倍，分别表示源领域，目标领域和通用部分，扩充的方法很简单，将源领域的值复制到新的特征表示的通用部分和源领域部分，目标领域也是一样，如图 9 所示。该方法适用于带有正则项的机器学习方法，如最大熵，SVM 等，正则项会使模型参数尽量小，那么该参数如果在目标领域和源领域通用的话，就会在“通用”

的特征空间有较高的权重，否则，只会在源领域空间或目标领域空间有较高权重。此方法非常简单，并且在部分数据集上达到了和基于先验的方法接近的效果。

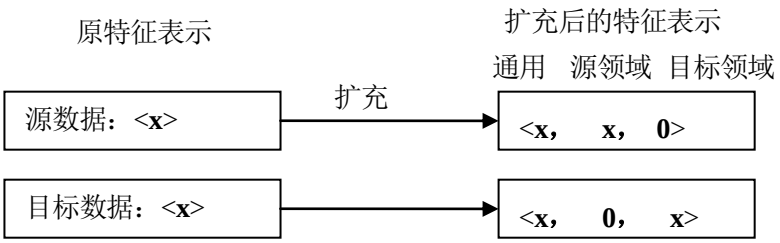


图 9 基于特征扩充的领域适配方法示例

2.2.4 基于相关关系的迁移学习

这种方法主要应用于关系领域数据，样本之间并不是相互独立的，而是存在关系的，如网络数据领域，社交网络领域等。基于相关关系迁移学习假设源领域和目标领域，数据间的关系是相似的，因此可以在他们之间迁移关系的知识。如导师、学生、论文之间的关系和经理、员工、项目之间的关系相似，可以在导师、学生、论文数据上学习到的关系特点，然后将这种关系直接或间接迁移。通常相关关系的迁移问题都采用基于统计关系学习的方法来解决。因为关系领域数据挖掘不在本文研究内容之内，在此不对基于相关关系的迁移学习做详细介绍。

2.3 微博分类的研究现状

微博分类是指将微博自动分到预先定义好的一个或多个类别中，包括按主题分类（体育，政治，财经等），按情感分类（喜欢，厌恶，普通等），按类型分类（私信类，评价类，事件类）等。微博内容既包含微博中的文本，还包括微博的作者、微博中的链接、哈希标签(#Hash Tag#)、点名(@username)、微博的转发、评论等信息。

2.3.1 微博分类中的主要研究问题

目前对于微博分类的研究可以分为下面三个方面：

1. 文本稀疏性问题

微博一般都会有长度限制，如 twitter 的 140 个单词，新浪微博的 140 个字，据统计 Twitter 的平均长度为 10 个单词。因此，微博的文本内容大多简单精炼，且常常依赖于上下文环境，有时只是一句话或者几个短语。微博文本过短将会产生严重的数据稀疏性问题。

为解决稀疏性问题，研究者主要采用了利用 Wikipedia、搜索引擎的搜索结果，新闻等外部资源的方法来丰富短文本的内容。Liu[29]等人提出了一种特征选择方法，先选取

词性丰富的词作为特征，再利用 Hownet 扩充和他们语义相关的特征，提高分类的效果。Sun 等人提出一种对 Wikipedia 进行语义分析的方法，然后利用得到的概念信息提高分类的效果[22]。Phan 等人提出一种框架[26]，将大规模的外部语料库作为全局数据集，使用话题模型在全局数据集上挖掘隐含话题，将短文本数据映射到隐含话题空间下训练分类器，解决数据稀疏问题。

2. 微博数据变化快，会产生概念漂移问题

对于微博流中存在的概念漂移问题，常见的解决方法有根据时间进行样本选择，样本加权，集成学习，特征选择等方法。

1. 样本选择方法指对以前的训练样本进行选择，使用一个滑动窗口选择时间最近的样本是常用的方法。Klikenberg 等人提出了一种自动确定窗口大小的方法，在新闻话题分类上展示了很好的性能。

2. 样本加权方法假设样本越新，对分类当前实例越重要，根据时间给样本不同的权重。Lebanon 等人提出了一种根据时间对样本加权的朴素贝叶斯分类算法。

3. 集成学习的方法构建一个分类器的集合，分类的结果由集合中的分类器共同决定。当观测到分类的准确率下降时，将旧的模型或者效果差的模型从集合中移除，增加新的模型来适应当前的数据，以保证分类的效果。

4. 特征选择的方法通过动态地选择最有价值的特征（词项）来应对概念漂移现象。特征选择的方法常常和其他方法结合使用。

Nishida 等人[64]提出了一种 P-switch 算法解决微博分类中的概念漂移问题，将词分为稳定词和突发词，对突发词使用最新的数据估计概率，对稳定词使用全局的数据估计概率。最后，他们在自己定义的 Twitter 数据集上的实验结果表明，该方法要优于其他时间敏感的分类方法。

3. 微博社交特性

微博本身是一个社交网络，除了微博文本内容外，微博还有其他一些元数据，如作者信息，转发，评论，微博中的点名，哈希标签等。利用这些数据可以对微博进行分类。

Bharath 提出了一种通过构建 8 个特征对微博进行分类的方法。该方法预先定义了“新闻”、“事件”、“观点”、“营销”和“私人信息”五个类别，根据 Tweet 发布者和 Tweet 本身提取出 8 个特征，分别为：作者，是否出现缩略词或俚语，是否有时间-事件词组，是否包含观点词汇，是否包含对词语的强调，是否包含货币或百分号，“@用户名”出现在 tweet 的开始，是否包含“@用户名”。采用朴素贝叶斯分类器在 Twitter 上实验表现出了不错的分类效果，优于传统的词袋模型。

2.3.2 迁移学习在微博挖掘的应用

迁移学习的技术已经成功应用到很多领域，如文本挖掘[51][50]，图像分类[52]，命

名实体识别[53]，跨语言分类[54]，Wifi 定位[49]等问题。将迁移学习应用于微博分类的研究工作还不多，将迁移学习方法用于微博挖掘或者可以用来解决微博分类的问题的工作如下：

Zhang 等人[20]提出了对部分观测数据的辅助学习方法(Assisted Learning for Partial Observation, ALPOS)，用来解决微博中短文本的问题，是一种基于特征表示的迁移学习方法。该方法认为微博文本是普通长文本的一部分，文本短是因为一些文字没有被观察到。利用长文本作为源数据（辅助数据），提高微博分类的效果。该方法对自学习方法的框架进行了扩展，要求源数据和目标数据具有相同的特征空间和标签空间，且需要有标注的源数据。其采用了一种有监督的特征构建方法，在学习基元素时，同时考虑了对源数据的恢复程度和分类的准确性。

在用基元素恢复目标领域数据时，只对考虑微博中出现的词，因为该方法假设没出现的词只是未观测到而已。在接下来的工作中，Zhang 等人在[21]中将该方法扩展为一种隐含空间学习的方法，并在监督学习的分类和无监督的标注两种应用场景下展示了该方法的有效性。

Long 等人[28]提出了一种迁移学习的方法解决稀疏短文本进行分类的问题(TCSST)，该方法属于归纳式迁移学习，基于实例的迁移学习方法，可以用来解决微博中的短文本和稀疏性问题。该方法对 TrAdaBoost 框架进行了扩展，为了解决源数据中标注数据少的问题，使用半监督学习的方法，对源数据进行采样。利用源标注数据，采样后的源未标注数据和目标标注数据，采用 TrAdaBoost 的框架训练分类器。TCSST 方法的框架如图 10 所示。该方法在 20-Newsgroup 数据集和一个真实的研讨会评论数据上实验展示了有效性。

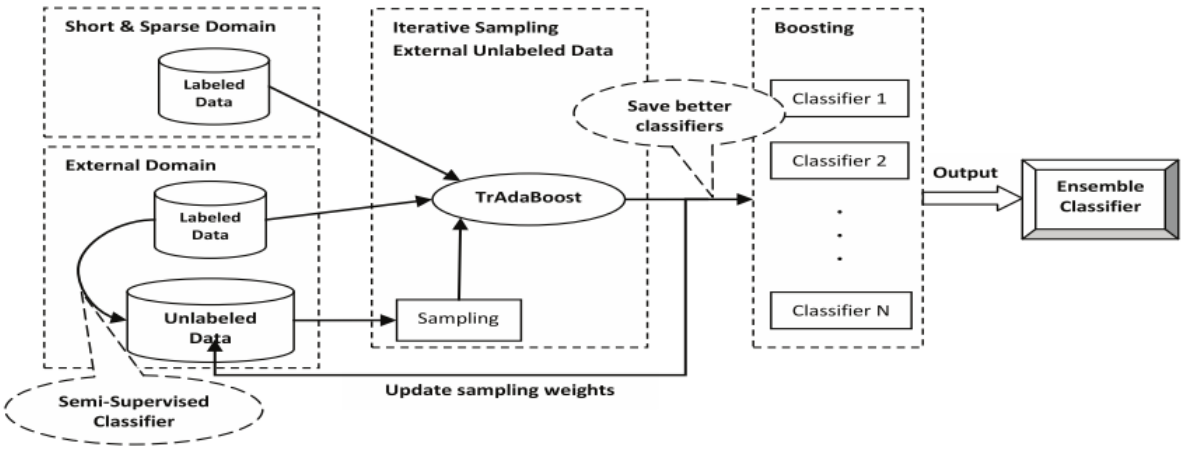


图 10 TCSST 方法框架

2.4 本章总结

一、本章简要介绍了文本分类技术，包括文本预处理，文本表示，文本分类算法和

分类效果评价方法，在文法分类算法中重点介绍了本文采用的朴素贝叶斯方法。本章还介绍了迁移学习技术，根据迁移的知识的形式不同，从基于实例，基于特征表示，基于参数和基于相关关系四类分别进行了介绍。

二、目前针对微博分类的研究主要围绕微博文本稀疏性，数据变化快和社交元数据三个方面展开。其中针对微博数据变化快的研究关注点在于怎样应对数据变化，仍然需要有持续的有新的标注数据。根据我们的调研，还没有对减少对人工标注微博数据的研究，而人工标注数据需要耗费极大代价，在保证分类效果的前提下减少人工标注工作具有重要意义。

三、目前将迁移学习的方法应用于微博分类的研究还较少，Zhang 等人采用了迁移特征表示的方法，Long 等人采用了迁移实例的方法，都是用来解决微博中短文本稀疏性问题。目前还没有研究利用迁移学习的方法解决微博数据变化快导致的人工标注问题。

第三章 数据集构建和数据分析

随着近年来微博的流行，对微博的传播、检索、挖掘等研究吸引了国内外科研工作者的注意，多数的研究工作都是在 Twitter 的数据上进行的，在国内的中文微博数据上进行的研究目前还很少。国内的微博和 Twitter 存在着很多的区别，国内微博用户的思想、行为和表达方式与国外用户也不尽相同。为了得到国内微博数据的结果，本文采集了新浪微博 14 天的数据，并利用微博的哈希标签构造了六个类别的微博数据集。同时，我们采集了同时间段的新浪新闻和网易新闻数据，对这些微博和新闻数据进行了初步的分析，并在这些数据集上对本文提出的算法进行了有效性验证。本章主要介绍了数据集的构建方法、数据预处理方法和数据分析。

3.1 数据集的构建

3.1.1 数据源和类别定义

新浪微博是一个由新浪网推出，提供微型博客服务的类 Twitter 网站，是目前国内最流行的微博产品之一[10]。我们使用新浪微博作为微博采集的数据源。新浪网和网易网是中国著名的门户网站，我们使用新浪网和网易网作为新闻采集的数据源。

我们定义了六个主题类别，其中包括 3 个大类，每个大类包含两个小类。分别是：体育类-NBA，体育类-国际足球，财经类-房地产，财经类-股票，科技类-互联网，科技类-智能手机。

表 5 主题类别定义和设置

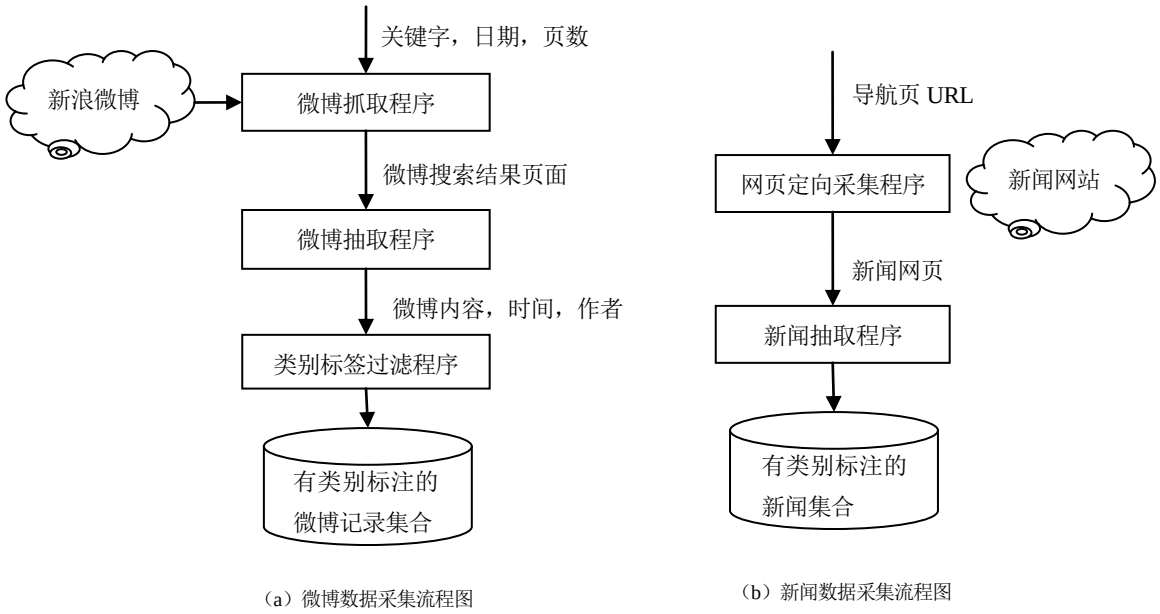
类别	描述	新闻网址	微博搜索关键字	微博标签
Sports.nba	NBA 篮球	新浪 NBA，网易 NBA	#NBA#,#美职篮#	#NBA#,#美职篮#
Sports.football	国际足球	新浪国际足球，网易国际足球	#英超#，#西甲#，#意甲#，#德甲#，#法甲#，#欧冠#	#英超#，#西甲#，#意甲#，#德甲#，#法甲#，#欧冠#
finance.house	房产类	新浪房产，网易房产	#房产#，#房价#，#房地产#	#房产#，#房价#，#房地产#
finance.stock	股票	新浪财经股票，网易财经股票	#股票#,#股市#,#港股#，#A 股#，#美股#，#概念股#	#股票#,#股市#,#港股#，#A 股#，#美股#，#概念股#
Tech.internet	互联网	新浪科技互联网，网易科技互联网	#互联网#，#互联网科技#	#互联网#，#互联网科技#
Tech.mobile	手机	新浪手机，网易手机	#智能手机#	#智能手机#

[10] http://baike.baidu.com/view/2762127.htm

3.1.2 数据采集方法

按照时间，分别从新浪微博和新浪新闻，网易新闻抓取了相关微博或新闻，构成了微博-新闻按时间平行数据集。对于微博采集，我们使用 Java 实现了微博采集、解析、抽取和存储程序。在对新闻数据的采集和抽取过程中，我们使用了公开的火车头数据采集程序，该程序提供了网页定向采集，根据规则或 Xpath 抽取数据，存储数据等功能[11]。

微博和新闻数据采集的流程图分别如图 11 中 a、b 所示。



（a）微博数据采集流程图　　（b）新闻数据采集流程图

图 11 微博和新闻数据采集流程图

1. 微博数据采集步骤:

1) 根据新浪微博的搜索接口，指定时间范围，抓取页面数，关键词等参数，定向采集了 2013 年 3 月 1 日到 2013 年 3 月 21 日 3 周的微博数据。根据类别特点自定义了若干搜索关键字，用于搜集和该类别相关的微博，类别对应的关键字信息见表 5 中微博哈希标签一列。
2) 对网页进行解析，抽取出微博 ID，用户名，文字内容，发表时间，是否为原创，评论数，转发数等字段。采用了自定义规则匹配的方式，通过正则表达式实现。
3) 在新浪微博中，哈希标签是被两个“#”括起来的文字，形如“#标签#”，顾名思义，标签是对微博内容的标注，标示着微博的主题或类别。通过观察微博内容，我们发现，很多微博会使用“【标签】”的形式来突出微博的主题，因此，我们把用“【”和“】”括起来的内容也认为是标签的一种。我们认为，标签中包含类别指定关键字的微博是属于该类别的，如微博内容中有标签，且标签中

[11]http://www.locoy.com/

包含“NBA”关键字，则被认为属于“NBA 篮球”这个类别。我们对采集到的搜索结果进行过滤，仅保留具有类别标签的微博记录，这样，就得到了具有类别标签的微博数据集。

4) 将微博数据按时间、类别存储在本地磁盘上，如图 12 所示为数据的存储结构。

2. 新闻数据采集步骤：

1) 新浪、网易、腾讯等互联网新闻门户网站，通常都具有详细的类别体系来对网站内容进行组织，以方便用户分门别类地进行浏览。因此，我们可以根据新闻网站的类别导航体系来获得某类别下的所有新闻。如从“体育 - NBA”版块得到所有属于“NBA”类别的新闻。

2) 从新闻网站的，类别版块下的“滚动新闻”栏目获得该类别的新闻 URL 列表，抓取这些 URL 对应的新闻网页。

3) 采用规则匹配的方式，通过正则表达式实现，对新闻网页进行解析，从中抽取出新闻的正文，标题和发表日期，按日期存储。这样，就得到了带有类别标注的新闻数据集。

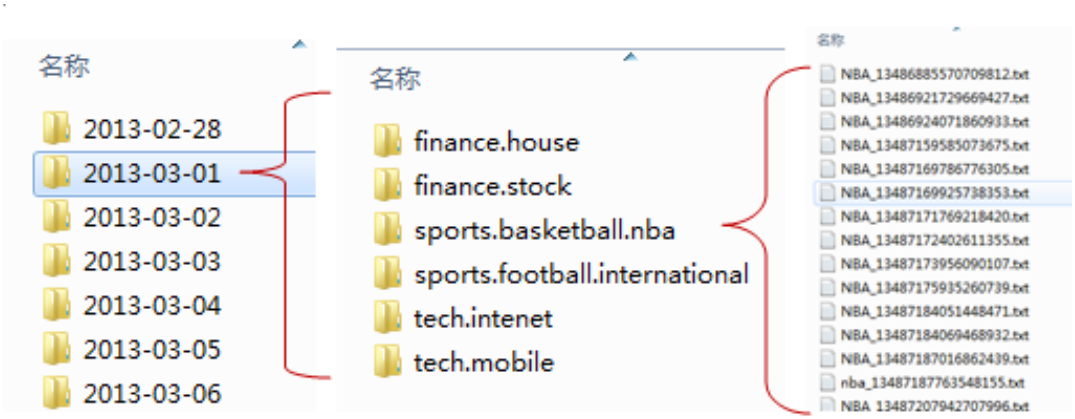


图 12 数据集存储结构示意图

3.2 数据预处理

采集到原始的微博数据后，为了去除噪音的干扰以及为分类做准备，我们对原始的数据进行了预处理，图 13 为预处理的流程，包含下面的步骤：

1. 因为我们是通过关键词集合搜索的方式获取的微博数据，这样就会导致所有的微博文本中都会出现搜索的关键词，这些关键词将会是非常明显的特征（比如仅仅通过判断是否包含某类别的搜索词就可以判断是否属于该类别），为了消除这些关键词的影响，我们将搜索词从微博文本中过滤掉。

2. 因为使用了哈希标签作为判断类别的依据，所以需要将包含类别关键字的哈希标签过滤掉，对于不包含类别关键字的哈希标签，则不过滤。

3. 微博文本中的外部链接对分类没有帮助，将其去除掉。点名信息@username 在本文的研究中暂时不用，为了排除其带来的影响，将点名信息去除掉。

4. 中文分词，采用了极易分词库（JE-MMAnalyzer），是一个开源的分词软件。为了提高分词效果，我们对每个类别构建了用户词典，将该类别的术语、人名、机构名等加入用户词典。

5. 我们采用了采用了哈工大自然语言处理小组的停用词表[12]，包含 767 个中文停用词，对停用词进行过滤。

6. 预处理后，如果一条微博记录的字数少于 5 个，那么将其过滤掉。

7. 采用基于文档频率的方法进行特征选择，因为它实现简单、高效，并且前人的研究结果表明 DF 特征选择方法具有很好的效果。

8. 采用词袋模型（Bag of Words，BOW）进行文本表示，将一条微博记录表示为一个向量，并将所有的数据转换为 Weka 可以处理的 arff 格式[13]，作为分类程序的输入。

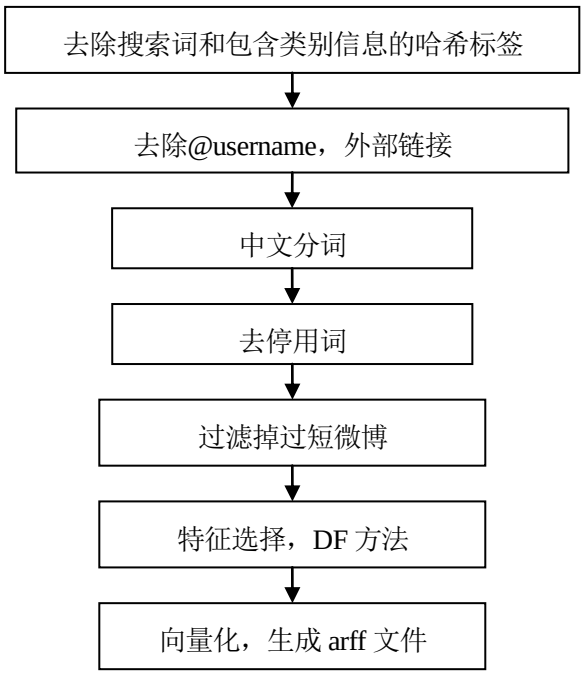


图 13 数据预处理流程图

3.3 数据分析

3.3.1 微博数据分析

微博作为一种新兴的社交网络，出现的时间并不长，但是发展速度非常快。微博消息内容短，信息传播速度快，热门话题变化快是微博给人们的普遍印象。“微博打拐”、“郭美美事件”，“北京雾霾天气”等一个个热点话题在微博上兴起、传播、又被新的话

[12] http://www.datatang.com/data/13281/
[13] http://weka.wikispaces.com/ARFF

题淹没。

从直观感觉来看，微博数据的统计性质是在不停变化的，前人的研究结果也有所说明[64][65][66]。下面，我们从两个方面来分析这种变化：1，类别分布的变化，即各类别所占比例的变化。2，类别中词项出现概率的变化，可以反映类别和词项的关系。

图 14 显示了从 3 月 1 日到 14 日，6 个类别中微博数量占微博总量的比例的变化，不同颜色的形状面积代表了该类别下微博的数量。总体来看，类别分布每天都表现出比较大的变化，而且变化是比较突然的而不是平缓的，如 3 月 2 日这天，足球类数量突然上升，股票类下降等。

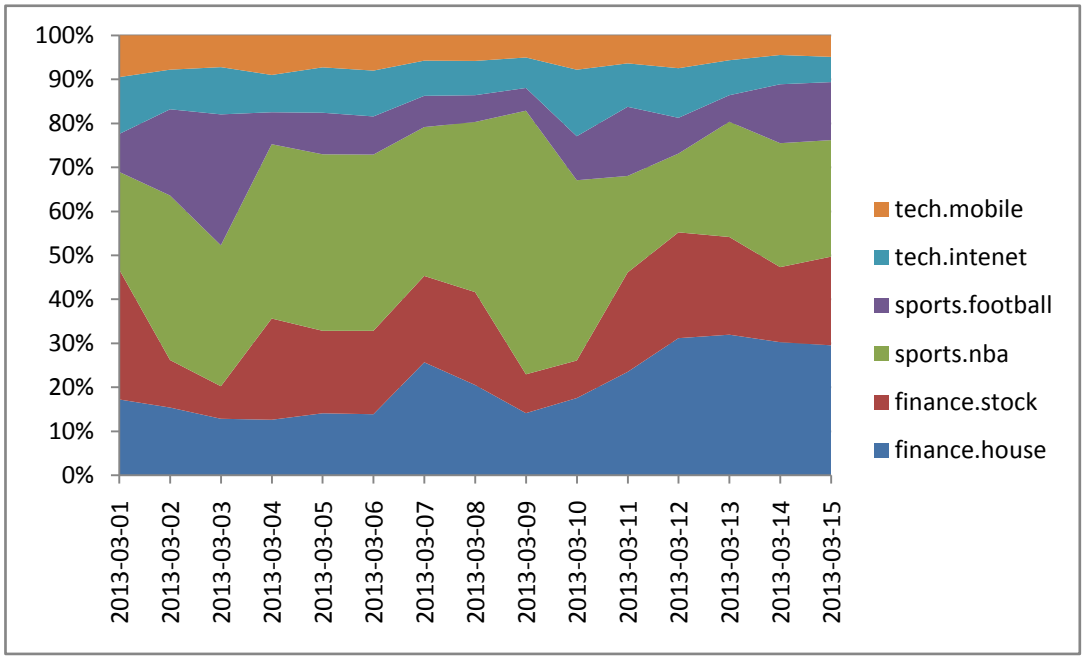


图 14 微博类别分布变化图

为了分析类别下词项分布的变化，我们统计词项在类别中出现的频率变化，即词项出现的次数占类别中所有词项出现的总数的比例。以“体育-篮球”类别为例，从中选取了典型的 12 个词，分析了它们出现频率的变化。其中“科比、热火、火箭、雷霆”是在该类别中出现频率很高的 4 个词，代表了高频词。“三分球、不敌、篮板、助攻、得分”这 5 个词是在篮球类别中稳定出现，但是频率并不是很高的词。“安慰、琼斯、左手”这 3 个词整体出现的频率不高，但是在某一天频率突然升高，属于突发词。

图 15 显示了上述 14 个词的频率从 3 月 1 日至 14 日的变化曲线。对于高频词，频率变化更加突出，升得快跌的也快。对于突发词，表现出频率一致很低，在某个时间点突然升高，之后又突然回落到很低的特点。对于稳定词来说，在稳定的时间段内，频率在较低水平范围内波动，当然，稳定只是一段时间内的状态，稳定词也可能在某个时间突发。综合来看，这些词的频率随着时间在不停的改变，而且变化速度很快。

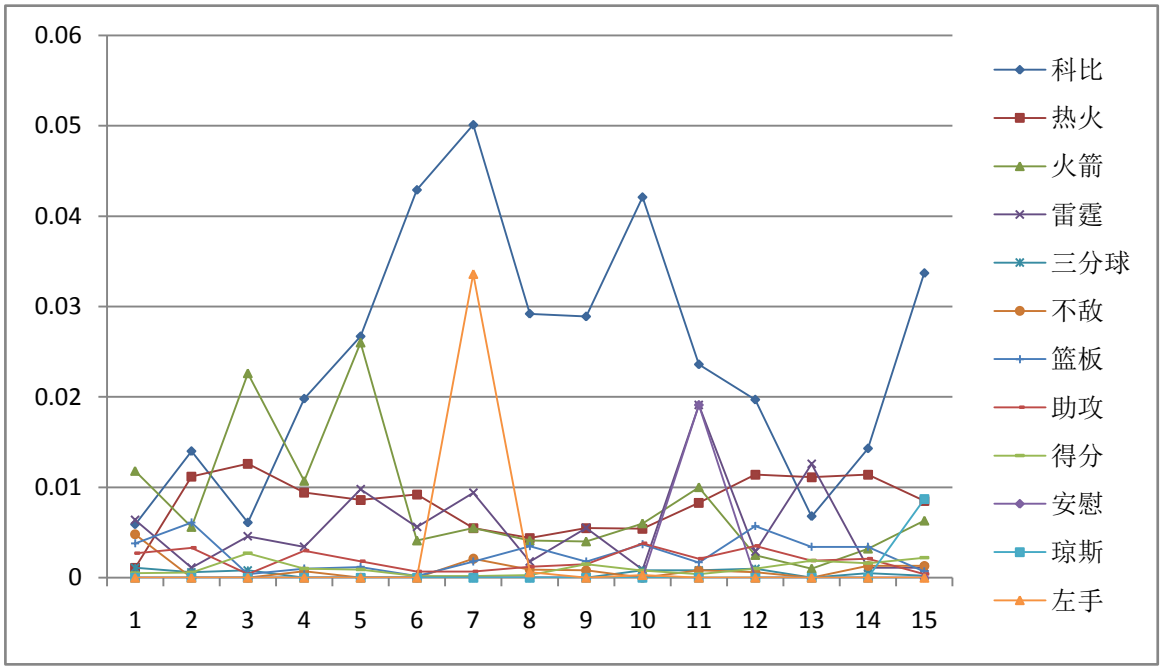


图 15 体育-NBA 篮球类下典型词项的频率变化

3.3.2 微博和新闻的关系

对于新闻上的热点事件报道，在微博上往往会引起广泛的讨论，如“国五条出台”，“两会政策”，“NBA 热火队连胜”等，微博用户会对这些话题发表见解或表达情感。而对于微博上产生的热门话题，也往往会受到社会各界的广泛关注，网络新闻也会及时的进行报导，如“流浪儿冻死垃圾箱”，“北京雾霾天气”等。因此，同类别的微博和新闻中的热门话题是相似或相关的。

因为微博 140 字的限制、网民自媒体的特质和其社交网络的特点，微博文本和传统的新闻文本有很大的区别，主要表现在：1，用户在发表微博时，语言表达上更加自由，不会拘泥于句法文法，很多微博甚至只是几个短语。而新闻作为正式的媒体，则在表达上则更加规范，不符合语法的病句极少出现。2，微博上的用词更加口语化，网络新词层出不穷，而新闻在措辞上更加书面化。

为了更加了解微博内容词分布的特点，发现微博和新闻之间的联系和差别，我们采集了六个主题类别下 2013 年 3 月份的新浪微博数据和新闻数据，并对其进行统计。

1. 微博和新闻的相关性

我们首先分析了每个类别中每天出现频率最高的前 100 个词中，微博和新闻重复的词的比例，如图 16 所示，总体来看，新闻和微博重复词的比例保持在 10%-50%，由此可见，在微博中出现频率高的词中一部分在新闻中出现频率也很高。由于数据的动态变化，每个类别每天重复词的比例也在改变，这也是符合我们之前关于微博数据变化快的假设的。由此可得，同类别下的新闻和微博具有相关性，随着时间变化，这种相关性依

然存在。

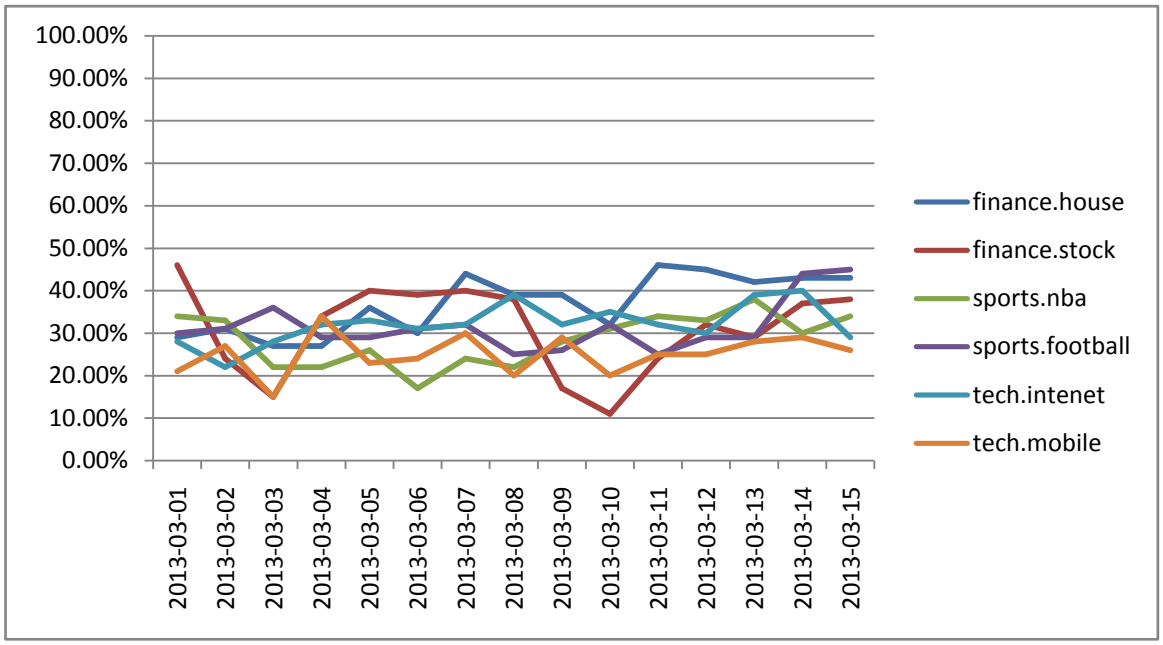


图 16 微博和新闻中频率最高的 100 个词重复的比例

为了统计微博上的突发词在新闻上是否也表现出了突发的特点，我们分析了“左手、琼斯”等突发词在新闻和微博上的变化情况，篇幅所限，仅绘出了“左手、琼斯”这两个词在微博和新闻中频率的变化曲线。如图 16、图 17 所示：

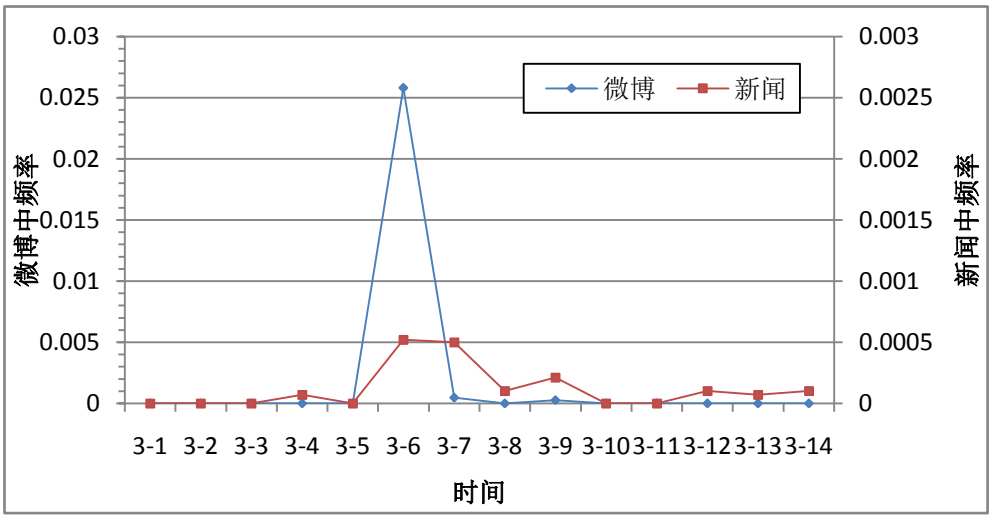


图 17 “左手”一词在微博和新闻出现频率变化图

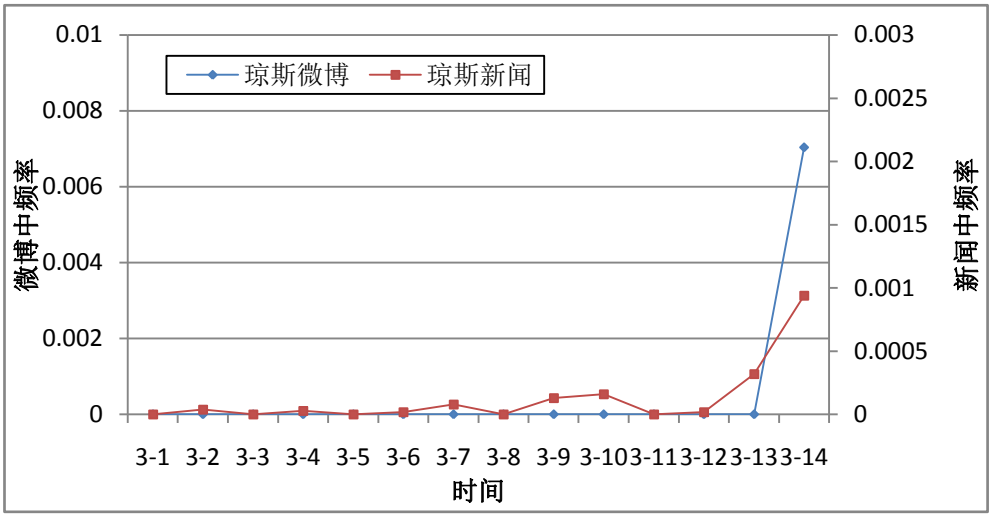


图 18 “琼斯”一词在微博和新闻出现频率变化图

如上图所示，当微博中的某个词突发时，该词在新闻数据中的频率也会大幅上升，由此可知，微博和新闻在突发词这点具有相关性，微博中的突发词在新闻也会有所体现。另外，微博中的突发词突发频率更突出，消失更快，如“左手”在 3 月 6 日前每天在微博的出现频率几乎为 0，在 3 月 6 日为 0.256，3 月 7 日又迅速下降。而在新闻中变化则要缓和很多，在 3 月 6 日后的几天，仍会出现比较高的频率，而非立刻下降为原来的状态。“琼斯”等其他突发词也表现出类似的特点。

注意到图 16、17 中，左坐标轴是词在微博中的频率，右坐标轴是词在新闻的频率，虽然变化趋势一致，但是词在新闻中的频率要小于微博。这是因为在特征选择前，新闻中的词数要远多于微博，因此词频率的绝对值会小于微博。表 5 是微博和新闻的总次数，所有词出现的总数，文档总数和平均文档长度的统计值。

表 6 微博和新闻文档数、词项数和文档长度对比

数据源	总词数	所有词出现总数	文档数	平均文档长度
微博	41253	282912	9809.0	28.8
新闻	178008	5758068	16517.0	348.6

2. 微博和新闻的区别性

根据上述的分析，我们可以判定同类别的微博和新闻具有相关性，但是两者之间也有差别，我们认为同类别的微博和新闻的差别主要在于文字表达方面。如表 5 所示，微博的平均长度 28.8 个字要远小于新闻的 348.6 个字。微博文字表达口语化、用词随意，不讲究语法规则；新闻文字表达书面化，用词讲究，严格遵守语法规则等是人们对两者之间区别的直接印象。

我们统计了微博和新闻中每个类别频率最高的 100 个词中，相同的词和微博特有的词和新闻特有的词，取频率最高的 10 个，统计结果如下表。

表 7 微博和新闻在用词上的区别示例

类别	重叠词	微博词	新闻词
Finance.house	城市，二手，调控，政策，房地产，房价，新，投资，市场，中国	房地产开发，信息，买房，泡沫，平米,行业，十年，quot，涨，货币	记者，建设，出现，执行，需求，一些，认为，目前，影响，增加
Finance.stock	指数，数据，股，股票，上涨，投资者，中，3月，市场，中国	收盘，股指，深，今天，今日，美股，震荡，成交，新高，道	业务，股东，产品，进行，认为，集团，上市公司，资产，相关，行业
Sports.nba	连胜，科比，詹姆斯，nba，湖人，火箭，胜，比赛，热火，季后赛	直播，来自，熊，人，篮球，视频，战，詹皇，vs，11	得到，表现，领先，命中，说，他的，体育，连续，之后，网易
sports.football	皇马，进，欧冠，米兰，曼联，中，赛季，比赛，球，主场	直播，这场，巴塞罗那，我看到，加油，视频，vs，英镑,西甲，凌晨	表现，本文，时，他的，前，作者，成为，之后，机会，网易
tech.internet	网络，移动，产品，广告，市场，服务，用户，互联网，手机，中国	app，派，微信，搜索，报告，美女，系，微博，百度，lbs	记者，时，进行，提供，表示，认为，品牌，称，电，去年
tech.mobile	使用，三星，android，功能，发布，市场，用户，手机，苹果，智能	3g，业务，老年人，2012年，针对，公司，增长，市场份额，讯，世界	像素，配置，设计，分辨率，支持，htc，机身，采用，galaxy，这款

统计结果验证了我们之前的假设，微博和新闻在用词方面存在较大区别。主要在于新闻为了保证文字表达的完整和规范，有很多书面语和助词。如“认为”、“目前”、“一些”、“之后”等。而这样的词在微博中很少出现，微博中的词多数具有明显的含义。原因很简单，微博有 140 的字数限制，因此表达更简练。

我们没有过滤掉分词的错误，因为目前的分类算法难以避免这些错误，而且它们确实会对分类效果产生影响，因此我们把这些错误保留了下来。

3.4 本章总结

目前国内外的研究大多在 Twitter 上进行，采用中文微博做实验的研究还很少，为了了解中文微博上的真实情况，我们采集了新浪微博的数据。因为现在没有公开的中文微博标注语料，我们利用微博的哈希标签，得到了有类别标注的微博，作为实验数据。

通过对 3 月 1 日-14 日微博数据的统计分析，验证了微博上数据变化快的特点，主要表现在类别分布的变化和类别中词出现频率的变化。微博数据变化快这一特点也使得减少微博标注工作的更加重要。

我们分析了微博和新闻的关系，统计数据表明，同类别下的微博和新闻存在着相关性，随着热门话题的改变，微博和新闻数据的变化具有一致性，如在微博上的突发词在新闻上也表现出突发性。因此，根据他们之间的相关性，我们可以利用新闻作为辅助数

据，减少微博标注的工作量，提高分类效果。

微博和新闻的区别主要表现在文字表达上，这会导致词的统计性质在微博和新闻上的差异，如新闻中的高频词在微博中不一定为高频词。因此，直接使用新闻数据而不考虑新闻和微博的区别，将会产生负面的影响，降低分类的效果。因此，研究迁移学习的方法，更有效的利用新闻数据具有重要意义。

第四章 同源迁移学习方法

因为微博数据是动态变化的，这种变化将导致数据的统计性质发生变化，因此不同时间的数据将会有不同的统计性质，旧数据不能反映新数据的特点，因此直接使用在旧数据上得到的分类器用于新数据的分类难以达到理想的效果。本章研究如何充分利用旧标注数据，降低数据分布变化带来的影响，提高对新数据的分类效果。因为研究不同时间的微博数据，因此我们称为同源数据。本章分别从两个思路出发，探索从旧数据中迁移出知识用于新数据分类的方法。

从时间因素出发，假设时间越近的数据越重要，我们采用了指数衰减的方法，随着时间的发展旧数据的权重不断减少；考虑到微博中突发词和稳定词分布变化性质的不同，采用选择性指数衰减方法，即只对突发词指数衰减。

从数据分布因素出发，不进行时间相关性的假设，我们采用了基于 Boosting 的迁移学习算法 TransferBoost，增加旧数据中和新数据分布相似的数据的权重，降低分布不相似的权重，以提高分类器在新数据的分类效果。

我们在新浪微博数据上进行了实验，结果显示，在标注数据较多的情况下，选择性时间衰减方法对分类效果有所提升，而 TransferBoost 方法效果较差，不适合在同领域同源数据间迁移。另在，在标注数据较少的情况下，几种方法都难以取得理想的效果。

4.1 任务描述

微博上每天产生大量信息，公众讨论的话题变化很快，这也就导致了微博数据变化快，容易过时的特点，第三章对微博的数据分析也证明了这一点。那么如何更有效的利用过时的数据，提高对新数据的分类效果，是一个很重要的问题。我们将本章要解决的任务描述如下：

如图 19 所示，根据时间顺序，以天为基本单位，将数据分块，D_T^1，D_T^2，D_T^3 ，…，D_T^t表示从时间 1 到 t 的微博数据，分类算法利用这些数据训练分类器，在新的测试数据上取得好的分类效果。

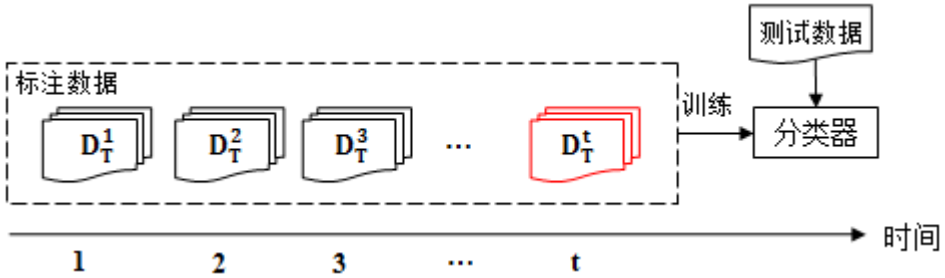


图 19 历史微博数据的利用方法

4.2 基于时间衰减的方法

如前文所述，微博上数据分布是变化的，意味着D_T^1，D_T^2，D_T^3 ,…，D_T^t的分布是不同的，直接使用时刻 t 之前的数据（旧数据）预测 t 时刻的类别是不合理的，需要研究更有效的方法利用旧数据。

基于时间衰减的方法假设时间越近的数据和新数据的分布越相似，因此其权重就越大，反之其权重就越小。本文采用了朴素贝叶斯分类法，指数衰减模型，这种思想的实现方式为采用指数衰减的方式对类别先验概率 p(c)和词项条件概率 p(w|c)进行估计。

指数衰减的方法对 p(w|c)进行估计时对所有词项都衰减，没有考虑不同类型词项概率变化性质的不同。选择性指数衰减把词分为突发词和稳定词，认为对突发词进行指数衰减的概率估计可以反应数据最新的特点，对稳定词不进行衰减，可以得到更平滑的概率估计，具有较好的推广性。

下面对指数衰减方法和选择性指数衰减方法分别进行介绍。

4.2.1 指数衰减方法

本文采用指数加权移动平均(Exponentially Weighted Moving Average，EWMA)的方法进行时间衰减。EWMA 作为是一种常用的序列数据处理方式，随着时间的流逝，旧数据的权重呈指数衰减，每过一个时间单位，旧数据的权重乘以一个衰减因子$(1-\lambda)$，其中$0<\lambda<1$，λ越大，表示衰减的速度越快，当$\lambda=1$时，旧数据等价于直接被丢弃掉，当$\lambda=0$时，新数据不起作用。$0<\lambda<1$时，旧数据的权重会随着时间越来越低，但是不会等于 0。

本文采用朴素贝叶斯分类法，取后验概率最大的类别，类别预测方法如下公式：

$$c_{map} = \underset{c\in\mathbb{C}}{\operatorname{argmax}} P(c)_{EWMA} \prod_{1\le k\le n_d} P(w_k|c)_{EWMA}$$

在 t 时刻，采用指数加权移动平均的方法对类别 c 的先验概率$P(c|t)_{EWMA}$估计如下：

$$P(c|t)_{EWMA} = \begin{cases} (1-\lambda)P(c|t-1)_{EWMA} + \lambda P(c|t)_{ML} & t>1 \\ P(c|t)_{ML} & t=1 \end{cases}$$

其中$P(c|t)_{ML}$是类别 c 在 t 时刻极大似然估计的先验概率，即采用 t 时刻的训练数据估计的概率值。$P(c|t-1)_{EWMA}$是在 t-1 时刻，采用指数加权移动平均估计得到的概率。当 t>1 时，类别 c 在 t 时刻的先验概率的估计值$P(c|t)_{EWMA}$为其在 t-1 时刻的概率估计值与 t 时刻的最大似然估计值的线性插值；当 t=1 时，$P(c|t)_{EWMA}$等于在 t 时刻的极大似然估计值。

类似的，在 t 时刻，对词项 w 在类别 c 的条件概率的估计方法如下：

$$P(w|t,c)_{EWMA} = \begin{cases} (1-\lambda)P(w|t-1,c)_{EWMA} + \lambda P(w|t,c)_{ML} & t>1 \\ P(w|t,c)_{ML} & t=1 \end{cases}$$

4.2.2 选择性指数衰减方法

根据我们对微博数据的分析，可以将类别相关的词项分为两类：稳定词和突发词。突发词指词项 w 在 t 时刻，在类别 c 的条件概率急剧上升，然后很快下降。稳定词则相反，在一定时间范围内相对稳定。

为了解突发词和稳定词概率变化情况，我们从类别“体育-NBA”中选出突发词：“左手”、“安慰”、“琼斯”；稳定词：“三分球”、“篮板”、“助攻”、“得分”，统计其每天频率变化情况，结果如图 20 所示：

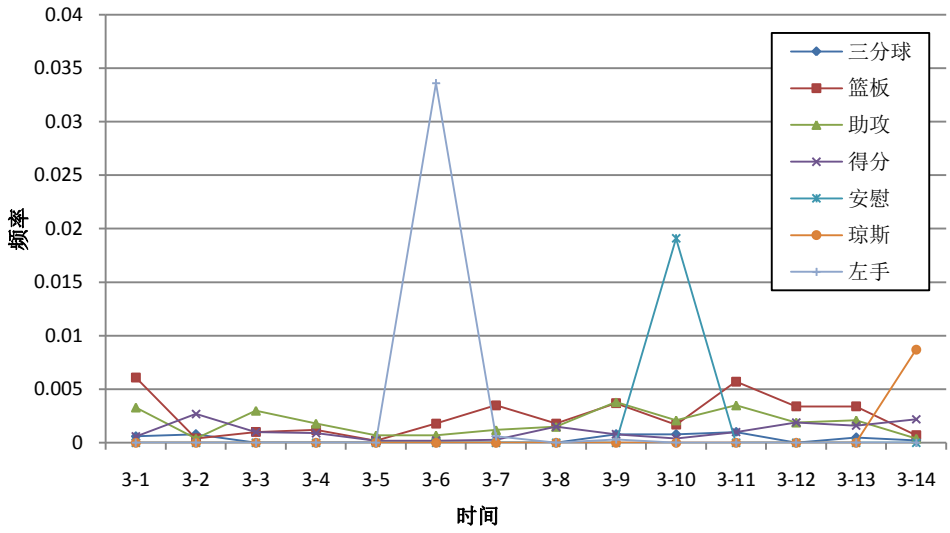


图 20 “体育-NBA”类别中部分突发词与稳定词的条件概率变化

如图所示，突发词的频率均是在某一天突然爆发。如“左手”这个原本跟“体育-NBA”并不是密切相关的词，在 3 月 7 日那天突然爆发，其在类别“体育-NBA”的条件概率从接近 0，突然上升到 0.035，第二天则跌回接近 0。“安慰”、“琼斯”也表现出类似的性质。

“三分球”、“篮板”、“助攻”、“得分”在一段时间内，频率保持在一定范围内，没有大的波动。这些词是和类别相关的常规词，在非突发的情况下，其条件概率应该保持稳定。

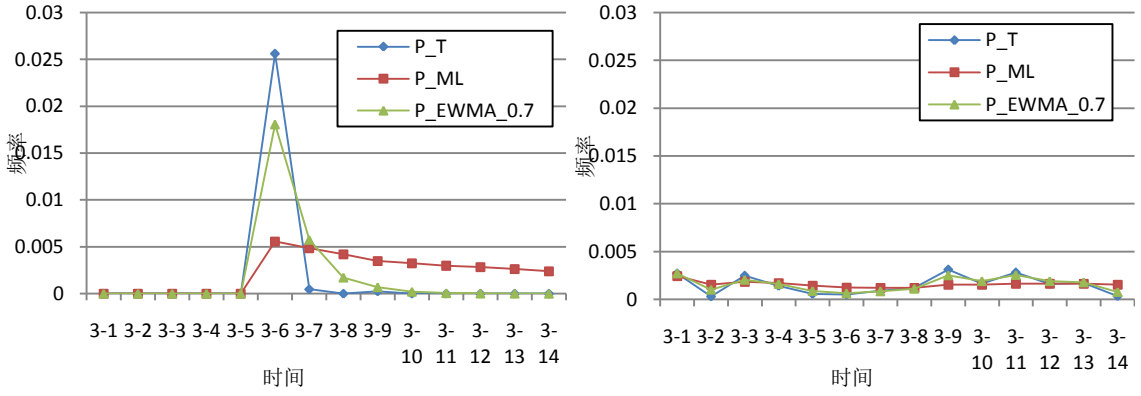


图 21 采用全局极大似然估计和指数衰减估计分布对突发词和稳定词的影响

图 21 显示了对突发词和稳定词分别采用全局数据的极大似然估计和指数衰减估计的效果。其中突发词以“左手”为例，稳定词以“三分球”为例。图中 P_T 是在 t 时刻数据上最大似然估计的概率，P_ML 是在 t 时刻之前所有数据上采用极大似然估计的条件概率，P_EWMA_0.7 是$\lambda = 0.7$时采用指数加权移动平均估计的效果。

如图 21 中左图所示，对于突发词，指数衰减估计能对反映出其概率的变化，而且影响消减很快。而极大似然估计因受到历史数据的影响，对突发词的概率估计反应要慢，且突发词产生的影响需要较长时间才能消失。如在 3 月 14 日“左手”在“体育 NBA”的真实条件概率已接近 0，而极大似然估计仍保持了较高的估计值。

如图 21 中右图所示，稳定词三分球的真实频率在 0.0003 和 0.003 之间波动，变化范围很小。这时，采用全局数据极大似然估计的结果更平滑更稳定，而指数衰减估计则波动较大，如在 3 月 2 日、5 日等，出现了概率接近 0 的情况，而这对于稳定词来讲是不合适的。

综上，我们认为对突发词，即时性更重要，对于稳定词，平稳性更重要。因此，对突发词采用指数衰减方法，对数据的剧烈变化做出快速反应；对稳定词采用全局数据的极大似然估计，对数据正常范围内的小波动，保持模型的稳定性。

判断词是否为突发词的方法如下：比较词项 w 的指数加权移动平均估计的概率值$\mathrm{P(w|t,c)_{EWMA}}$和最大似然概率估计的概率值$\mathrm{P(w|c)_{ML}}$，当前者比后者大超过一定阈值时，则认为 w 为突发词，否则，认为 w 为非突发词。

采用选择性指数衰减方法，在给定时刻 t 和类别 c，词项 w 的条件概率估计值$\mathrm{P(w|t,c)_{SEWMA}}$的计算方法如下公式：

$$\mathrm{P(w|t,c)_{SEWMA}} = \begin{cases} \mathrm{P(w|t,c)_{EWMA}} & \text{if } \mathrm{P(w|t,c)_{EWMA}} > \mathrm{P(w|c)_{ML}} + \mathrm{L}\sigma_{c,t} \\ \mathrm{P(w|c)_{ML}} & \text{otherwise} \end{cases}$$

$$\text{s.t.}\sigma_{c,t} = \sqrt{\mathrm{P(w|t)_{ML}}(1-\mathrm{P(w|c)_{ML}})}\sqrt{\lambda/(2-\lambda)}$$

其中$\mathrm{P(w|c)_{ML}}$表示词项 w 在所有数据上的极大似然估计概率。L 表示控制突发探测的程度，或者说是敏感程度，L 越小，对词项概率的变化越敏感，即判断为突发词的阈值越低。$\sigma_{c,t}$表示概率标准差的估计值。即认为，当概率的波动大于标准差范围，认为是发生了突变，否则认为是正常范围的波动。关于 EWMA 控制和$\sigma_{c,t}$推导在文献中有详细的描述。

4.2.3 算法实现

本章所采用的选择性指数衰减方法受到了 Nishida 等人[27]提出的 P-switch 算法的启发，但是与 P-switch 有所不同。首先在应用场景上不同，P-switch 算法属于流数据处理

中的训练-测试模型，每次输入一个新的标注样本，更新分类模型。本章提出的算法是分批-训练-测试模型，每次输入一块训练数据，批量更新模型。

我们给予 Weka 实现了增量的选择性指数衰减朴素贝叶斯分类法，对模型进行更新的更新算法流程图如下：

算法：选择性指数衰减朴素贝叶斯模型更新算法

输入：新数据D_T^t；

t-1 时刻，类别的先验概率：$P(c|t-1)_{ML}$，其中$c \in C$；

t-1 时刻，类别下词项的条件概率：$P(w|t-1,c)_{ML}P(w|t-1,c)_{EWMA}$，其中$c \in Cw \in V$；

1：对每个类别 c

for c=1，...,|C| do

2：计算类别先验$P(c|t)_{ML}$

3：对每个词 w

for w=1,…,|V| do

4：更新极大似然估计的概率 $P(w|t,c)_{ML}$

5：更新指数衰减估计的概率$P(w|t,c)_{EWMA}$

6：计算标准差估计值$\sigma_{c,t}$

7：计算$P(w|t,c)_{SEWMA}$

$$P(w|t,c)_{SEWMA} = \begin{cases} P(w|t,c)_{EWMA} & \text{if } P(w|t,c)_{EWMA} > P(w|c)_{ML} + L\sigma_{c,t} \\ P(w|c)_{ML} & \text{otherwise} \end{cases}$$

8： end for

9： end for

输出: 更新后的朴素贝叶斯分类器

4.3 基于数据分布的方法

章节 4.2 提出的选择性指数加权移动平均方法建立在一个假设上的，那就是数据在时间上越接近，他们的分布越接近。因此，样本越新，权重越高。

但是这个假设并不总是成立的，在微博中，数据变化不仅速度快，而且幅度大，往往是突变的，而非缓和的连续的。特别是对于爆发性的话题和突发词项，图 14 也证明了这一点。此外，对于具有周期性或者可能重复出现的话题，分布相似的数据不一定是时间最接近的数据。例如，在“体育-NBA”类别中，在 3 月 11 日发生了产生了绝杀球，这一天在微博上关于绝杀的话题和讨论立刻上升，这个时候，3 月 10 日的数据并不是和 3 月 11 日分布更接近的应该是 3 月 7 日的数据，因为这天也有类似的绝杀球事件发生。

因此，我们尝试另外一种同源数据间迁移知识的方法，旧数据的权重不是简单根据时间先后决定，而是根据它和新数据分布的相似度来决定，提高分布相似的数据的权重，降低分布差异大的数据的权重。

4.3.1 基本方法

我们采用一种基于实例的迁移学习算法 TransferBoost[68]，在 t 时刻，以$\mathbf{D_T^t}$为目标数据，历史数据$\mathbf{D_T^1}$，$\mathbf{D_T^2}$，$\mathbf{D_T^3}$,…,$\mathbf{D_T^{t-1}}$作为源数据。该算法是对 TrAdaBoost 算法的改进，支持从多个源数据向目标数据迁移。基于 AdaBoost 框架，在模型迭代过程中，根据分类器在目标数据和源数据的分类结果，调整数据实例的权重，以使得跟新数据分布相似的历史数据有较高的权重，跟新数据分布不相似的数据具有较低的权重，以提高目标数据的分类效果。

TransferBoost 把不同的源数据分为不同的任务（数据），认为不同任务（数据）对目标任务的可迁移力是不同的，需要给可迁移力强的任务较高的权重。根据 Eaton 在 2008 年给出的定义，源任务的可迁移力由迁移和不迁移在目标任务上的误差的差别决定，可迁移力的计算公式如下：

$$\mathrm{Transferbilitys(S,T)} = \epsilon_{\mathrm{T}} - \epsilon_{\mathrm{T\cup S}}$$

其中 Transferbilitys(S,T)表示源数据 S 向目标数据 T 的可迁移力，ϵ_{T}是只使用目标数据进行训练，在目标测试数据上的误差，$\epsilon_{\mathrm{T\cup S}}$是同时使用目标数据和源数据进行训练，在目标测试数据上的误差。

4.3.2 算法实现

下面描述了 TransferBoost 的算法过程，其中 K 为迭代次数，$\alpha_{\mathrm{iter}}^{\mathrm{i}}$ 和β_{iter} 的计算公式在文献[68]中有详细描述。

算法：TransferBoost

输入：源数据$\mathbf{D_T^1}, \mathbf{D_T^2}, \mathbf{D_T^3}$,…,$\mathbf{D_T^{t-1}}$，目标数据$\mathbf{D_T^t}$，其中$\mathbf{D_T^i} = \{(\mathbf{x_j}, \mathbf{y_j})\}_{\mathbf{j=1}}^{|\mathbf{D_T^i}|}$

1：合并源数据和目标数据 D=$\mathbf{D_T^1} \cup \mathbf{D_T^2} \cup \mathbf{D_T^3}, \ldots, \cup \mathbf{D_T^t}$

2：初始化样本权重$w_1(x_i) = 1/|D|$，for $(x_i, y_i) \in D$

3：迭代 K 次

for iter=1，...,K do

4: 在当前样本权重下训练分类函数$h_{\mathrm{iter}}: X \to Y$

5: 为每个历史数据集确定参数$\alpha_{\mathrm{iter}}^{\mathrm{i}}$

6: 确定参数β_{iter}

7: 更新所有样本的权重

$$w_{\mathrm{iter}+1}(x_j) = \begin{cases} \dfrac{w_{\mathrm{iter}}(x_j)\exp(-\beta_{\mathrm{iter}} y_j h_{\mathrm{iter}}(x_j) + \alpha_{\mathrm{iter}}^{\mathrm{i}})}{Z_{\mathrm{iter}}} & (x_i, y_i) \in D_T^1, D_T^2, D_T^3, \ldots, D_T^{t-1} \\ \dfrac{w_{\mathrm{iter}}(x_j)\exp(-\beta_{\mathrm{iter}} y_j h_{\mathrm{iter}}(x_j))}{Z_{\mathrm{iter}}} & (x_i, y_i) \in \mathbf{D_T^t} \end{cases}$$

8: end for

输出: 组合的分类器 hypothesis $H(x) = \mathrm{sign}(\sum_{\mathrm{iter}=1}^{K} \beta_{\mathrm{iter}}\, h_{\mathrm{iter}}(x))$

4.4 实验

4.4.1 实验设计

实验应用场景是，每天都对一定数量的微博进行人工标注，利用这些标注的数据，对新的微博进行分类。

实验采用在第三章中描述的数据集，对六个类别分别进行实验，每次实验都是二分类问题。从 2013 年 3 月 1 日至 3 月 14 日，以天为基本单位，将数据分为 14 块，3 月 14 日的数据为当前数据，3 月 1 日至 13 日的数据为历史数据。从当前数据和历史数据中分别随机抽样出一定比例的数据作为训练数据，剩余的当前数据作为测试数据，以测试数据上的效果来评价分类算法的好坏。为保证结果的鲁棒性，实验独立进行 10 次，每次的训练数据均是由随机抽样产生，实验结果取 10 次的平均值。实验采用 F1 值作为主要评价指标，六个类别的综合指标去 F1 的宏平均值。

本次实验主要为了验证 3 个问题：

1. 采用时间衰减的方法和选择性时间衰减的方法是否会提高分类效果？
2. 基于数据分布的迁移学习方法 TransferBoost 是否有效？
3. 比较上述不同方法，在不同微博标注数据量下的分类效果。

下表列出了本次实验比较的几种同源数据迁移方法。

表 8 同源迁移方法实验的对比方法

方法名	方法描述
NBC	标准朴素贝叶斯分类法，不考虑数据的时间关系，将历史数据直接和当前数据合并作为训练数据，是实验的基准方法。
EWMA	采用了时间衰减方法的朴素贝叶斯分类法，其中衰减系数$\lambda = 0.2$，取经验值。
S-EWMA	采用选择性的时间衰减的朴素贝叶斯分类法，其中衰减系数$\lambda = 0.2$，突发敏感度 L=0.01，取经验值。
TransferBoost	TransferBoost 算法，采用多项式模型的朴素贝叶斯作为基分类器，所有样本的初始权重相等，采用算法默认参数，迭代次数 100 次。

4.4.2 实验结果与分析

图 22 显示在每天微博标注数据为 5%，50%和 80%下的实验结果，评价指标采用了在六个类别的 F1 值的宏平均。从实验结果可以看出，在有较充分的标注数据（50%）的情况下，选择性时间衰减的效果最好，但是跟基准方法相比，提高的幅度有限。提高有限的原因一方面在于选择性指数衰减的改进主要在于突发词，突发词的个数是有限的，提高对突发词条件概率估计的准确率影响了那些根据稳定词无法正确分类的样本。另外在标注数据较充分的情况下，基准方法的分类效果也达到较高的水平，很难获得大幅提

高。

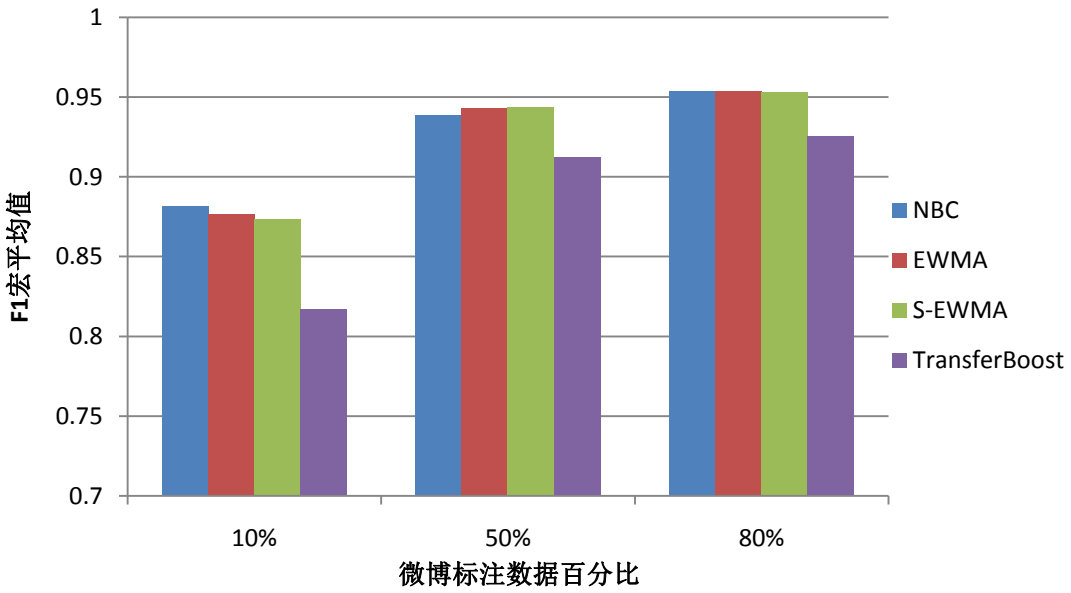


图 22 不同比例标注数据的情况下同源迁移分类实验结果

指数衰减的方法也要略优于基准方法，推测原因在于在标注数据较充分的情况下，侧重利用最新的信息，已经可以对稳定词给出较平滑的估计，又可以反映最新数据的特点故优于基准方法。

对于 TransferBoost 方法，在数据量少，充分和多的情况下都难以取得理想效果，推测原因在于 TransferBoost 最初提出用于在不用领域间进行迁移学习，如“足球”到“篮球”，该方法认为源领域中错分的样本是噪音，会降低其权重，对目标领域中错分的样本则提高权重。而同领域（类别）不同时间的微博数据，差异并不是那么明显，这种做法太偏重当前数据，会降低分类器的推广性，因此该方法不适合用于此应用场景。

当标注数据很多时，4 种方法都取得了较好的效果。而在标注数据较少时（10%），4 种方法的效果都不理想，利用全局信息的朴素贝叶斯方法取得了最好的效果，因为他充分的利用了所有信息，而其他三种方法会对部分数据进行减弱，在本来数据量就少的情况下，模型训练不充分，导致分类效果差。

上述给出的结果中，所有方法的参数均是取自经验值，没有对参数进行训练。下表的结果显示了在训练集上对 EWMA 的参数λ，S-EWMA 的参数 L 交叉验证的结果。

表 9 标注数据占 50%时同源迁移实验各类别的实验结果

类别	NBC	EWMA	S-EWMA	TransferBoost
tech.internet	0.9374	0.9432	**0.9442**	0.9145
tech.mobile	0.9315	0.9305	**0.9334**	0.8951
Sports.nba	0.9614	0.9685	**0.9687**	0.9433
finance.house	0.9568	0.9563	**0.9587**	0.9188
finance.stock	0.9478	0.9490	**0.9513**	0.9296
Sports.football	0.8978	**0.9192**	0.9186	0.8720

结果显示，选择性指数衰减方法在 6 个类别中的 5 个都取得了最好的效果，在“体育-足球”类略差于指数衰减方法。在每个类别上，选择性指数衰减方法都比基准方法有所提高。原因在于不同类别数据的分布和变化情况是不同的，因此，在实际应用中，根据分类任务和数据，使用选择性指数衰减方法，选择合适的参数可以提高分类的效果。

4.4.3 实验结论

通过本章的实验和分析，我们可以得出下面几点结论：

1. 在有充分标注数据的情况下，选择性时间衰减方法对分类效果有所提升。根据类别数据分布和变化特点，选择合适的参数可以提升选择性时间衰减方法在该类别的分类效果。
2. 相同类别下不同时间的微博数据分布差异并没有不同领域间数据差异那么明显，TransferBoost 方法不适用于在同源不同时间的数据间进行迁移学习。
3. 在标注数据很少的情况下，难以对词的概率进行准确的估计，分类模型训练不充分，因此实验中的几种方法都不能取得理想的效果。

4.5 本章总结

本章主要讨论了在同源不同时间的数据间进行知识迁移的方法，即如何有效利用历史微博标注数据，提高对当前微博数据的分类效果。分别从时间和数据分布两个因素出发，尝试了 3 种方法，分别为基于时间的指数衰减方法，基于时间的选择性指数衰减方法和基于数据分布的迁移学习方法 TransferBoost。实验结果显示，在具有较充分的数据时，选择性指数衰减方法可以提高分类效果。当微博标注数据量较少时，因为缺乏充分训练，上述几种方法都难以取得理想的效果。

因此，只利用微博数据本身的话，需要有较充分的标注数据才能达到理想的分类效果，而持续的标注工作将耗费很大的人力物力。因此，研究一种方法可以降低人工标注工作量具有重要意义。

第五章 跨源迁移学习方法

微博数据是动态变化的，每天都会产生大量新数据。新数据的统计性质和旧数据存在着很大差异，现有的解决方法都需要对新数据进行标注，重新训练分类器或更新原来的分类器。第四章实验的结果表明，当标注数据较少时，实验的集中同源数据迁移方法都很难达到理想的分类效果。众所周知，标注工作将耗费大量的人力物力，而微博数据更新快、数据量大的特点更加重了标注工作的代价。

本章提出的方法利用新闻数据作为辅助数据，因为同类别的新闻数据和微博数据属于同一领域、不同来源的数据，有相关的主题却有不同的文字表达形式。因为知识在不同来源的数据间迁移，我们称之为跨源迁移学习。

本章提出了两种跨源迁移学习的方法，分别为基于参数先验的方法和基于可迁移度的方法。实验证明，在标注微博数据较少的情况下，基于可迁移度的词概率迁移算法分类效果要明显优于其他方法，在只有 5%微博标注数据情况下，6 个类别上的 F1 宏平均值超过 90%。而且随着标注数据的增加，效果稳定提升，接近其他算法的最优水平。

5.1 任务描述

互联网新闻门户有着详细的类别体系，根据该类别体系，使用网络爬虫，我们就可以及时获取到最新的有类别标注的新闻数据。因此，本章探索基于迁移学习的方法，利用新闻作为源数据，在微博标注数据量很少的情况下提高分类效果，减少人工标注工作。

任务描述如图 23：D_T^1，D_T^2，D_T^3 ,…，D_T^t表示时刻 1-t 目标数据上的数据（微博），D_S^1，D_S^2，D_S^3 ,…，D_S^t表示在时刻 1-t 源数据上的数据（新闻），分类目标是在目标数据标注量较少的情况下，提高分类器在D_T^t数据上的效果。

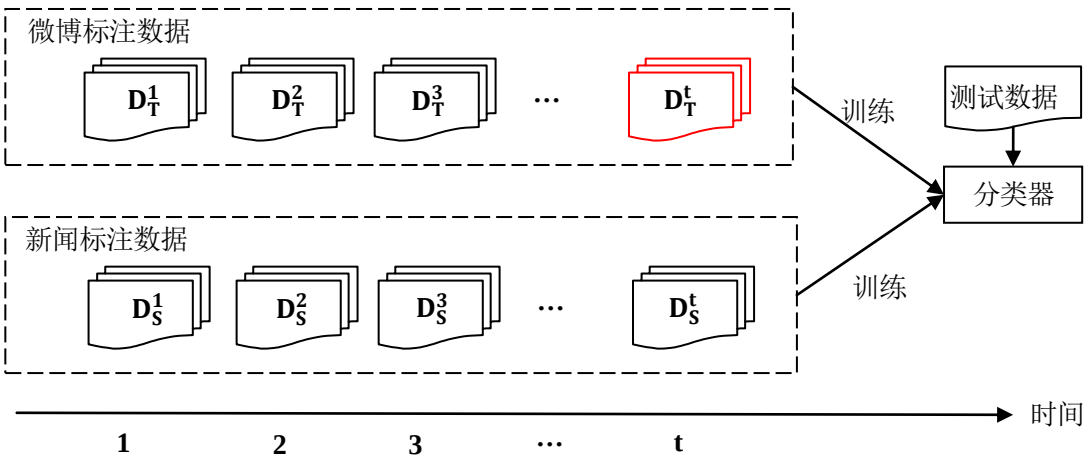


图 23 跨源数据迁移任务描述

5.2 跨源迁移学习方法框架

本文采用基于参数的迁移学习方法，即知识以参数的形式进行迁移。基于参数的迁移学习方法和基于实例、基于特征表示的迁移学习方法相比，形式上更简单，不需要解复杂的优化问题或者是多次迭代问题，因此学习和分类的效率都很高，适合于微博数据量大的特点。

本文采用多项式模型的朴素贝叶斯方法，是一种生成式模型，需要估计的参数是每个类别的先验概率P(c)和每个词在类别 c 的条件概率P(w|c)。类别的先验概率 P(c)表示在没有任何其他信息的时候，文档 d 属于类别 c 的概率，由类别 c 的重要程度决定。类别 c 下词 w 的条件概率P(w|c)表示类别 c 中的文档产生词 w 的概率，代表了词 w 和类别 c 的关系。因此，对概率 P(c)和 P(w|c)的估计越准确，朴素贝叶斯分类法的效果越好。

从应用场景设置来讲，属于归纳式迁移学习，要求源领域和目标领域有相同的类别空间和特征空间，源领域和目标领域都需要有标注的样本。

5.3 基于参数先验的方法

5.3.1 基本思想

基于参数先验的迁移学习方法的基本思想是：把在源领域训练数据上学到的知识作为目标领域的先验知识，即在没有任何目标领域训练数据进行指导的情况下，以源数据上得到的先验知识进行分类；当拥有目标领域训练数据时，根据目标领域数据的特点，对先验知识进行修正，以对目标领域的数据做出更好的分类判断。

图 24 显示了基于参数先验的迁移学习框架，在源数据和目标数据采用相同的模型，先在源数据上估计模型参数P_S，在目标数据上学习时，把P_S作为模型参数的先验，最终得到目标领域的分类模型的参数P_T。Chelba 和 Acero[69]最早提出基于先验的迁移学习方法[14]，在最大熵分类器的基础上采用了正则化的方法实现。

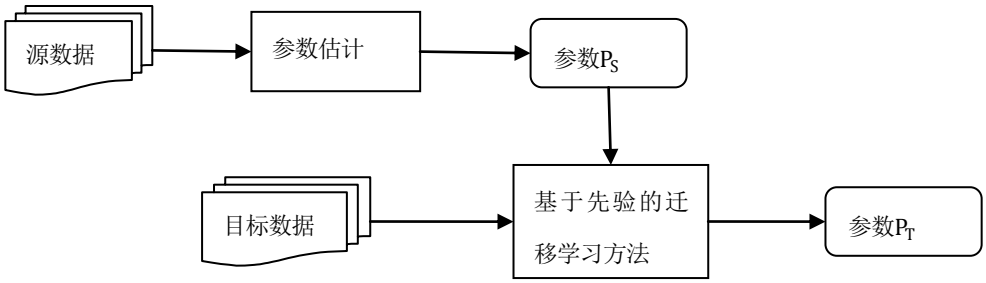


图 24 基于参数先验的迁移学习框架

[14]在自然语言处理领域，迁移学习方法也称为领域适配，文献[69]使用了这种说法。

5.3.2 算法实现

本文基于朴素贝叶斯分类法，提出了基于参数先验的迁移学习方法的一种实现WpriNaiveBayes（Word Priori Naïve Bayes）。WpriNaiveBayes采用贝叶斯估计的方法使用先验知识，是一种生成式的方法，和基于正则化的方法相比，更加简单高效。

在朴素贝叶斯分类法中，需要估计的参数为 P(c)和 P(w|c)，有极大似然估计和贝叶斯估计两种方法。在从一个已知的抽样中估计概率 P 时，极大似然估计认为概率就是频率，只考虑从样本中得到的信息，使似然度最大。贝叶斯估计认为，除了从样本中观测到的信息之外，我们对概率 P 还有一些先验知识，加入先验知识后，根据贝叶斯公式，通过最大化后验概率来估计概率 P。当先验知识可靠的时候，贝叶斯估计要比极大似然估计准确。朴素贝叶斯分类法中，常使用拉普拉斯平滑方法来避免零概率，假设词项的先验概率为均匀分布。

根据贝叶斯估计，先在源数据上估计的条件概率$\widehat{P}(w|c)_S$作为先验概率，目标领域数据条件概率$\widehat{P}(w|c)_T$的估计方法如下公式：

$$\widehat{P}(w|c)_T = \frac{T_{wc} + m\widehat{P}(w|c)_S}{T_c + m}$$

其中T_c表示目标数据中，在类别 c 中所有词出现的次数之和，T_{wc}表示在目标数据中，类别 c 中词 w 出现的次数，$\widehat{P}(w|c)_S$是先验概率，是从源数据上估计的概率值。参数 m 表示等效的样本数，相当于我们从先验分布中引入了 m 个虚拟的样本，$\widehat{P}(w|c)_T$由在目标数据中观测到的样本和这先验分布中 m 个虚拟的样本共同决定。m 越大，表示先验分布的权重越高，反之，似然度的权重越高。

类似的，类别 c 的先验概率$\widehat{P}(c)_T$的贝叶斯估计如下：

$$\widehat{P}(w|c)_T = \frac{N_c + n\widehat{P}(c)_S}{N + n}$$

基于贝叶斯估计的方法，也可以理解为用源数据上估计的概率作为先验，对目标数据概率估计做平滑。

5.3.3 融合同源迁移学习方法

第四章中我们比较了几种同源迁移学习的方法，其中选择性时间衰减的方法取得了较好的效果，选择性时间衰减的方法也是建立在朴素贝叶斯分类基础上的。因此，我们可以将它融合到上节提出的跨源迁移学习方法中。

回顾下我们所设定微博分类的场景，D_S^1，D_S^2，D_S^3 ，…，D_S^t为源数据，D_T^1，D_T^2，D_T^3 ，…，D_T^t为目标数据，分类算法的目标是在D_T^t上取得的最好的效果。我们可以使用“先跨源，后同源”的策略，将这两种方法融合起来，以取得整体上最优的效果。如图 25 所示，先使用每个虚线框内的数据进行跨源迁移，再沿着时间轴做同源迁移，具体做法分下面两步：

1. 在时刻 t，新闻数据D_S^t做源数据，微博数据D_T^t做目标数据，采用基于参数先验的迁移学习方法，得到条件概率$\widehat{P}(w|c)_T^t$。
2. 使用同源迁移学习中选择性指数衰减的方法，根据在 t-1 时候的概率值$P(w|t-1,c)_{ML}$和$P(w|t-1,c)_{EWMA}$，最终得到 t 时刻的参数$P(w|t-1,c)_{SEWMA}$。

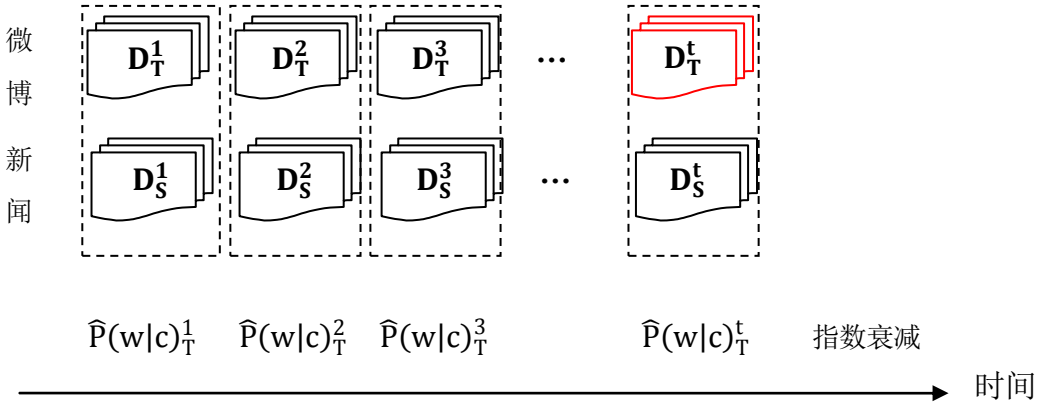


图 25 基于参数的跨源迁移学习方法融合选择性指数衰减方法示意图

5.4 基于词项可迁移度的方法

5.4.1 基本思想

基于参数先验的方法将从源领域数据中得到的所有知识都当做先验迁移给目标领域，依靠目标领域的训练数据对这些先验知识进行纠正。这种方法没有考虑到，当目标领域训练数据较少时，无法对那些具有“误导性”的知识进行有效的纠正，这样就会对分类效果造成损害，导致“负迁移”。上节提出的基于贝叶斯估计的方法将在源数据上估计的所有词的概率，都作为目标数据上词概率的先验。

根据第三章的数据分析，新闻和微博的区别主要表现在文字表达的不同，有些词在新闻和微博中表现出相似的统计性质，其概率可以从源数据迁移到目标数据，而新闻特有词则不行。

本节提出一种基于词项可迁移度的迁移学习方法 WtrNaiveBayes（Word Transfer Naïve Bayes），对词项进行有选择性的迁移。定义了指标词迁移度（Word Transfer Degree, WTD）来描述词从源领域向目标领域可迁移的程度。可迁移度高的词，说明该词在源领域和目标领域中通用性强，可以从源领域中迁移较多的关于该词知识用于目标领域；反之，若可迁移度低，则应少迁移或不迁移。

5.4.2 方法框架

如图 26 所示，基于可迁移度的方法分四步估计类别中词项的条件概率$P(w|c)_{tr}$：

1. 使用源数据，估计在时刻 t 的概率$\widehat{P}(w|t,c)_S$。
2. 使用目标数据，估计在时刻 t 的概率$\widehat{P}(w|t,c)_T$。

3. 计算词 w 从源数据向目标数据的可迁移度 wtd。
4. 使用上述 3 个值，计算出迁移后的条件概率$\widehat{\mathrm{P}}(\mathrm{w}|\mathrm{t},\mathrm{c})_{\mathrm{tr}}$

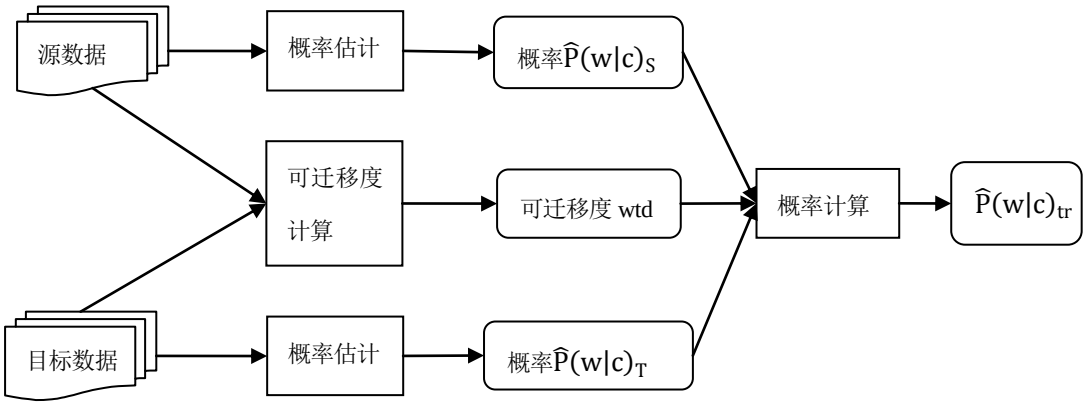


图 26 基于可迁移度的迁移学习方法框架

$\widehat{\mathrm{P}}(\mathrm{w}|\mathrm{t},\mathrm{c})_{\mathrm{tr}}$的计算方法为$\widehat{\mathrm{P}}(\mathrm{w}|\mathrm{t},\mathrm{c})_{\mathrm{S}}$和$\widehat{\mathrm{P}}(\mathrm{w}|\mathrm{t},\mathrm{c})_{\mathrm{T}}$的线性插值，二者的权重由可迁移度来决定，如下公式：

$$\widehat{\mathrm{P}}(\mathrm{w}|\mathrm{t},\mathrm{c})_{\mathrm{Tr}} = (1-\mathrm{wtd}_{\mathrm{wcST}})\widehat{\mathrm{P}}(\mathrm{w}|\mathrm{t},\mathrm{c})_{\mathrm{T}} + \mathrm{wtd}_{\mathrm{wcST}} * \widehat{\mathrm{P}}(\mathrm{w}|\mathrm{t},\mathrm{c})_{\mathrm{S}}$$

其中$\mathrm{wtg}_{\mathrm{wcST}}$表示在类别 c 下，词 w 从源领域 S 向目标领域 T 的可迁移度。

$\widehat{\mathrm{P}}(\mathrm{w}|\mathrm{t},\mathrm{c})_{\mathrm{S}}$和$\widehat{\mathrm{P}}(\mathrm{w}|\mathrm{t},\mathrm{c})_{\mathrm{T}}$的估计应利用局部较新的信息，以反应在时刻 t 最新的特点；计算可迁移度时，应利用 t 时刻之前的全局数据，以反应微博和新闻这两个不同源的数据整体的特点。

5.4.3 词的可迁移度

在类别 c 中，词 w 从源领域 S 向目标领域 T 的可迁移度定义如下：

$$\mathrm{wtr}(\mathrm{w},\mathrm{c},\mathrm{T},\mathrm{S}) = \frac{\mathrm{Sim}(\mathrm{w},\mathrm{c},\mathrm{T},\mathrm{S}) + \beta \times \mathrm{Fre}(\mathrm{w},\mathrm{c},\mathrm{T},\mathrm{S})}{\mathrm{Z}}$$

其中$\mathrm{Sim}(\mathrm{w},\mathrm{c},\mathrm{T},\mathrm{S})$表示类别 c 下词 w 在 S 和 T 上条件分布的相似度。分布越相似，说明词 w 在源领域和目标领域的通用性越高，可迁移能力越强。$\mathrm{Fre}(\mathrm{w},\mathrm{c},\mathrm{T},\mathrm{S})$是频繁度，表示词 w 在源数据和目标数据上出现频率的高低，因为低频词很有可能是噪音或者相关性不大的词，高频词往往比低频词更重要。而且在计算$\mathrm{Sim}(\mathrm{w},\mathrm{c},\mathrm{T},\mathrm{S})$时，低频词往往相似度更高，如计算 0.0000001 和 0.00000011 的相似度要高于 0.001 和 0.0011，但是在迁移的过程中后者更应该被迁移，加上频繁度指标，为了给高频词一个补偿。β控制相似度和频繁度的权重，Z 为归一化因子。

在计算相似度和频繁度时，均需要使用类别 c 下词 w 在源数据的概率$\mathrm{P}(\mathrm{w}|\mathrm{c})_{\mathrm{S}}$和在目标数据的概率分布$\mathrm{P}(\mathrm{w}|\mathrm{c})_{\mathrm{T}}$，因为可迁移度表示的是源数据和目标数据的整体性质，因此，使用全局的源数据和目标数据对其进行估计。下面分别给出相似度和频繁度的计

算方法。

1. **相似度**

类别 c 中，词 w 在目标数据 T 和源数据 S 的相似度的计算方法如下公式：

$$\mathrm{sim}(w,c,T,S)=1-\frac{1}{a\times e^{-b\times KL(P(w|c)_T,P(w|c)_S)}}$$

其中用 KL 距离来表示源数据 S 中词 w 在类别 c 的条件分布于目标数据 T 中词 w 在类别 c 的条件分布的差异。KL 距离，也叫相对熵，是衡量两个概率分布的指标，分布 P 和 Q 的 KL 距离计算方法如下：

$$D(P||Q)=\sum_{x\in X}P(x)\log\frac{P(x)}{Q(x)}$$

因为 KL 距离的值域是 0 到无穷大，我们采用了一个逻辑斯蒂方程对 KL 距离进行变换，其中 a>0, b>0，变换后的值域约束到$(\frac{1}{1+a},1)$，那么相似度的值域就为$(0,1-\frac{1}{a+1})$，表示词在源数据和目标数据的概率分布的相似度最高为$1-\frac{1}{a+1}$，通过调整参数 a，可以控制目标数据所占权重的最小值。

2. **频繁度**

频繁度$Fre(w,c,T,S)$由源数据中类别 c 词 w 的条件概率和目标数据中类别 c 词 w 的条件概率的调和平均值决定，调和平均值的性质是接近两者中较小的，因此，只有词在源数据和目标数据中的频率都较高时，才会有较高的频繁度。计算公式如下：

$$Fre(w,c,T,S)=\frac{2\times P(w|c)_T\times P(w|c)_S}{P(w|c)_T+P(w|c)_S}$$

5.4.4 融合同源迁移学习方法

在 WtrNaiveBayes 方法的框架中，在 5.4.2 中步骤 1、2 中，使用源数据估计在时刻 t 的概率$\hat{P}(w|t,c)_S$和使用目标数据估计在时刻 t 的概率$\hat{P}(w|t,c)_T$时，可以选择任意的概率估计方式：如只利用当前 t 时刻的数据进行概率估计，指数衰减概率估计，选择性指数衰减概率估计，利用全局数据进行概率估计等。

因此，在框架中融合同源迁移学习方法是非常自然的，对微博和新闻分别采用同源迁移学习方法进行概率估计即可。在本文中，我们采用了效果较好的选择性指数衰减方法。

5.5 实验

5.5.1 实验设计

本次实验采用第三章描述的微博数据集和新闻数据集，包括体育类-NBA，体育类-国际足球，财经类-房地产，财经类-股票，科技类-互联网，科技类-智能手机六个类别，时间从 2013 年 3 月 1 日到 14 日。以 3 月 14 日前的数据作为历史数据，3 月 14 日的数据作为当前数据，比较不同的分类算法，在 3 月 14 日微博数据上的分类效果。

为了比较在新闻标注数据充足，微博标注数据量不同时算法的分类效果，从每天的微博标注数据中选出一定比例用作训练。如采用 5%的标注数据，表示从每天的标注微博数据中随机抽样出 5%作为训练数据，以当天（3 月 14 日）剩余的 95%数据作为测试数据。所有的新闻数据均可用作训练数据，（因为新闻数据是辅助数据，可以得到大量有标注的新闻数据）。如图 27 所示，绿色部分为训练数据，红色部分为测试数据。

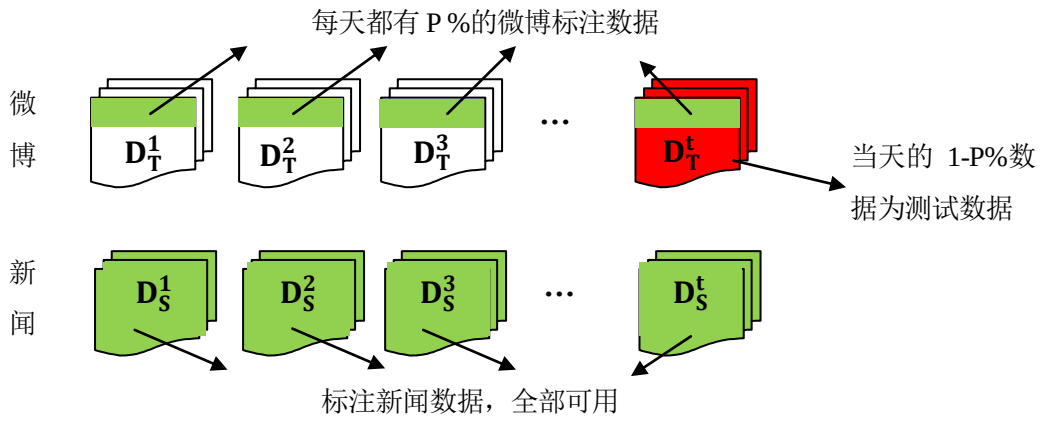


图 27 利用新闻数据的跨源迁移学习方法试验设计

每次实验重复进行 10 次，每次微博标注数据均是随机抽样产生。

本次实验为了验证以下 3 个问题：

1. 在微博标注数据量少的情况下，基于参数先验的迁移学习方法是否有效？
2. 在微博标注数据量少的情况下，基于词项可迁移度的迁移学习方法是否有效？
3. 在微博标注数据的比例变化时，上述两种方法的效果会发生怎样变化，和基准方法相比如何？

本次实验设计的基准方法为不使用迁移学习，使用原始的朴素贝叶斯分类法，将所有标注数据合并使用，根据使用标注数据不同，5 种基准方法如表 8 所示：

表 10 跨源数据迁移学习方法实验基准方法

类型	方法名	当前微博数据	历史微博数据	当前新闻数据	历史新闻数据
非迁移学习	NBM	是	否	否	否
	NBM-WEIBO	是	是	否	否

	NBM-SRC	是	否	是	是
	NBM-ALL	是	是	是	是
	NBM-CurSRC	是	否	是	否

本次实验需要验证的 2 种跨源迁移学习方法为：

1. Wpri-NBM：基于参数先验的迁移算法 WPriNaiveBayes，使用所有微博数据和新闻数据，参数选择 m=10000，融合了选择性指数衰减方法，衰减因子 lambda=0.35，突发度系数 BL=0.05，是经验值。
2. Wtr-NBM：基于可迁移度的词概率迁移算法 WtrNaiveBayes，a=2，b=1000，衰减因子 lambda=0.35，突发度系数 BL=0.05，和 Wpri-NBM 取一样的经验值。

5.5.2 实验结果及分析

1. 微博标注数据少时各方法效果比较

表 11 显示了在微博标注数据占 3%，5%，10%时的整体实验结果，评价指标取 6 个类别 F 值的宏平均。表 11 中斜体数字表示基准方法的最优值，粗体数字表示所有方法的最优值。

在所有的基准方法中，NBM-ALL 和 NBM-WEIBO 效果相对较好。在微博标注数据量极少时（3%，5%），NBM-ALL 要优于其他基准方法，原因在于此时模型处于训练不充分的状态，越多使用标注数据效果越好。当微博标注数据为 10%时，NBM-WEIBO 的效果超过了 NBM-ALL，说明随着标注数据量的增加，新闻和微博的差异性导致的“负迁移”现象逐渐明显，同时使用新闻和微博的方法比只使用微博数据的 NBM-WEIBO 方法要差。

WtrNaiveBayes 的方法在 3%，5%，10%时均取得了最优的效果，和所有的基准方法相比均有明显提高，因为 WtrNaiveBayes 在利用新闻数据的时候，有选择性的进行知识迁移，避免了不通用知识的误导。

基于先验的 WpriNaiveBayes 方法的整体效果要差于 WtrNaiveBayes，和其他基准方法相比，效果提高也不明显，当标注数据在 3%时，效果差于 NBM-ALL 和 NBM-SRC 方法。原因在于 WpriNaiveBayes 将所有的词概率都作为参数的先验进行了迁移，当标注数据较少时，无法给一些错误的先验进行充分的纠正，因此其效果要差于基于可迁移度的 WtrNaiveBayes。

表 11 微博标注数据占 3%、5%、10%时跨源迁移实验结果

标注微博比例	NBM	NBM-WEIBO	NBM -CurSRC	NBM -SRC	NBM -ALL	Wpri-NBM	Wtr -NBM
3	0.6276	0.8448	0.8675	0.8829	*0.8839*	0.8786(-)	**0.8949(+)**
5	0.7259	0.8787	0.8694	0.8830	0.*8849*	0.8921(+)	**0.9021(+)**
10	0.7787	*0.9056*	0.8742	0.8840	0.8871	0.9063(+)	**0.9130(+)**

表 12 显示了微博标注数据占 3%，5%，10%时，跨源迁移学习实验在各个类别上的

结果，其中粗体数字表示最优结果。结果显示，基于可迁移度的 WtrNaiveBayes 方法，在大多数类别上都取得了最优的效果，几乎所有的类别结果都要优于基准方法。另外，此处对于所有类别，WtrNaiveBayes 采用了相同的参数，根据我们的经验，针对不同类别的特点，采用不同的参数，可以提高 WtrNaiveBayes 方法在该类别的效果。

WpriNaiveBayes 在类别“财经-房产”上取得了较好的效果，但是在其他类别的效果普遍差于基准方法。推测其原因和“财经-房产”数据特点有关，该数据下迁移所有先验知识带来的提高要大于错误先验知识带来的损失。WtrNaiveBayes 因为没有迁移所有知识，效果略差于 WpriNaiveBayes，但是仍远好于基准方法。不过从总体上看，WTR-Bayes 在微博标注数据量较少的情况下的效果并不好。

表 12 微博标注数据占 3%、5%、10%时跨源迁移实验各类别的结果

类别	比例	NBM	NBM -WEIBO	NBM -CurSRC	NBM -SRC	NBM -ALL	Wpri -NBM	Wtr -NBM
finance.house	3	0.6861	0.8196	0.8236	0.8123	0.8139	**0.8651**	0.8591
	5	0.7345	0.8587	0.8323	0.8137	0.8166	**0.8807**	0.8692
	10	0.805	0.8931	0.8532	0.8175	0.8249	**0.9010**	0.8895
finance.stock	3	0.6512	0.854	0.8568	0.8762	0.8774	0.8696	**0.8884**
	5	0.7295	0.8704	0.8576	0.8771	0.8797	0.8803	**0.8917**
	10	0.8138	0.8957	0.8629	0.8792	0.8814	0.8936	**0.9042**
sports.nba	3	0.7949	0.901	0.9363	0.9432	**0.9449**	0.9173	0.9439
	5	0.8572	0.9159	0.9354	0.9431	0.945	0.9270	**0.9514**
	10	0.8894	0.9347	0.9365	0.9434	0.9458	0.9373	**0.9589**
sports.football	3	0.7398	0.8655	0.9155	0.9315	0.9311	0.9166	**0.9343**
	5	0.8008	0.9021	0.9178	0.9327	0.9327	0.9252	**0.9401**
	10	0.8425	0.9201	0.9211	0.9344	0.9344	0.9351	**0.9429**
tech.internet	3	0.476	0.8515	0.9109	0.9345	**0.9352**	0.9028	0.9236
	5	0.6734	0.8841	0.9103	0.9337	**0.9344**	0.9085	0.93
	10	0.7745	0.9158	0.9111	0.9331	0.9338	0.9117	**0.9353**
tech.mobile	3	0.4176	0.7771	0.7621	0.7993	0.801	0.8001	**0.82**
	5	0.5598	**0.8412**	0.7628	0.7974	0.8009	0.8307	0.8301
	10	0.5473	**0.8742**	0.7605	0.7966	0.8021	0.8593	0.8476

2. 微博标注数据比例变化时，各方法效果的变化

图 28 给出了在标注微博数据比例从 3%-90%之间变化时各种算法的效果，评价指标取所有类别上 F1 值的宏平均。

如图所示，当微博标注数据量小于 15%时，WtrNaiveBayes 方法效果明显优于其他方法。基准方法 NBM，NBM-WEIBO，NBM-SRC，NBM-CurSRC 都因为的标注数据较少，模型训练不充分，效果较差。NBM-ALL 因为利用了所有数据，效果要优于上面 4 种基本方法，但是由于是非迁移学习的方法，将新闻和微博同等对待，效果要差于迁移学习的方法。

随着微博标注数据量增加时，基准方法中 NBM 和 NBM-WEIBO 效果上升较快，因

为增加的微博标注数据解决了它们模型训练不充分的问题。基准方法 NBM-SRC，NBM-CurSRC 和 NBM-ALL 效果提升较缓慢，因为他们直接利用了新闻数据，此时，分类效果的提升受到了新闻数据带来负面影响的限制。

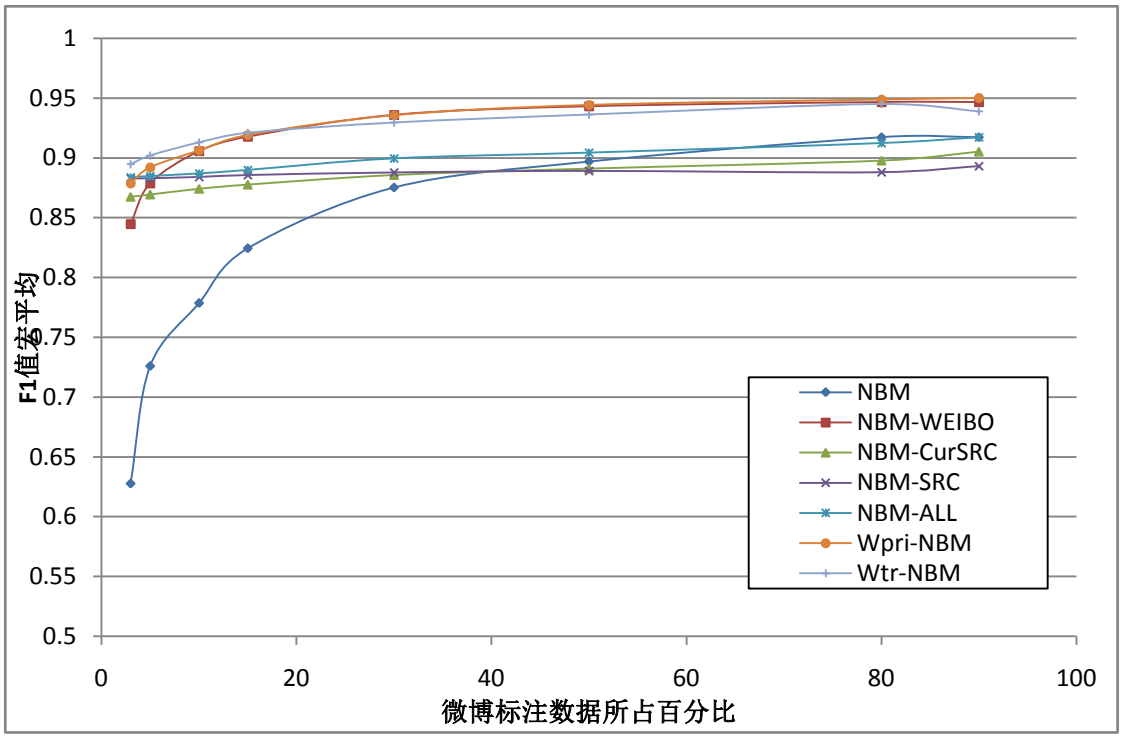


图 28 不同微博标注数据量下跨源迁移实验结果图

随着微博标注数据的增加，基于先验的迁移学习方法 WpriNaiveBayes 和基于可迁移度的迁移学习方法 WtrNaiveBayes 的效果均取得了提高，其中 WpriNaiveBayes 效果提升速度较快，这是因为增加的微博标注数据，对从新闻中迁移的“误导性”的先验知识进行了纠正。

当微博标注数据很较多时，基准方法中 NBM-WEIBO 的效果优于 NBM-SRC，NBM-CurSRC 和 NBM-ALL。因为此时微博训练数据充分，即使不利用新闻数据也可以取得较好的效果。NBM 方法由于只利用了当前微博数据，当前微博数据量还是比较有限，因此其效果差于 NBM-ALL。

当微博标注数据很较多时，基于先验的 WpriNaiveBayes 方法效果最好，略优于 NBM-WEIBO，因为其充分的利用了新闻数据，而充足的微博标注数据对先验知识进行了较好的修正，避免了其负面影响。说明，在有较充分的微博标注数据时，从新闻中迁移的知识仍对微博分类带来了帮助。

基于可迁移度的 WtrNaiveBayes 方法，也取得了与最优值非常接近的效果，远好于非迁移学习的 NBM-ALL 等基准方法，但是要稍微差于 WpriNaiveBayes 和 NBM-WEIBO。原因在于，当微博标注数据充足时，直接使用微博数据就可以对词的概率进行较准确的估计。即使是整体上可迁移度很高的词，在当前时刻，也存在着微博上估计概率和新闻估计概率有差异的情况，在这种情况下，从新闻中迁移概率会产生误导。

因此，当微博标注数据非常充分时，WtrNaiveBayes 方法会稍微差于 WpriNaiveBayes 和 NBM-WEIBO。

5.5.3 实验结论

根据上述的实验结果和分析，我们可以得出以下实验结论：

1. 当微博标注数据较少时，只利用微博数据难以取得很好的分类效果，如 NBM，NM-OLD 方法所示。
2. 基于迁移学习的 WpriNaiveBayes 和 WtrNaiveBayes 方法要优于非迁移的原始朴素贝叶斯方法 NBM-ALL 等，因为迁移学习的方法既利用了新闻中对微博分类有帮助的知识，同时可以抑制新闻中会对微博产生误导的负面影响。随着微博标注数据的增加，基于迁移学习的方法效果可以获得稳定提升，而非迁移的方法因为受新闻中负面知识的限制，分类效果提升幅度很小。
3. 在微博标注数据较少的情况下，基于可迁移度的 WtrNaiveBayes 方法的效果要明显优于基准方法和基于先验的 WpriNaiveBayes 方法。在只有 3%微博标注数据情况下，6 个类别上的 F1 宏平均值接近 90%。而且随着标注数据的增加，效果可以稳定提升，接近其他算法的最优水平，可以解决微博分类中标注问题，显著减少标注工作量，适合于本文提出的应用场景。

5.6 本章总结

本章主要讨论了微博数据更新快导致的微博分类中的人工标注问题：因为新微博的不断产生导致微博数据的统计性质在不停变化，在旧数据上训练的分类器无法直接用于新数据，因此需要对新数据进行标注，这将耗费极大的人力物力。第四章的实验也表明，只使用微博数据的话，无法在微博标注量很少的情况下取得理想的分类效果。为了降低标注工作量，本章使用新闻数据作为外部资源，根据新闻网站中详细的类别体系得到新闻的类别标签，提高微博分类的效果。

因为微博和新闻是不同来源不同形式的文本，同类别的微博和新闻数据间有相关性也存在差异性。为了有效利用新闻数据，抑制新闻和微博的差异带来的误导，本章提出了两种迁移学习的算法，分别是基于参数先验的 WpriNaiveBayes 方法和基于可迁移度的 WtrNaiveBayes 方法。

我们在第三章定义的数据集上进行了实验，结果表明：基于迁移学习的方法可以比非迁移学习的方法更有效的利用新闻数据，取得较好的分类效果。在标注微博数据较少的情况下，基于可迁移度的 WtrNaiveBayes 方法的分类效果要明显优于其他方法，在只有 3%微博标注数据情况下，所有类别上的 F1 宏平均值接近 90%。因此，采用 WtrNaiveBayes 方法可以显著减少微博标注量，降低微博标注工作带来的开销。

第六章 结束语

6.1 本文工作总结

近年来，微博在世界范围的兴起改变了人们的互联网生活，给人们提供了一个随时随地获取、传播信息的重要平台。随着微博的快速发展，微博上信息数量呈现爆发式增长，对微博信息进行有效的组织显得非常重要。微博分类是对微博信息进行组织的一种重要方式。

本文提出了微博分类中人工标注的问题：微博每天产生大量新的数据导致其统计性质发生改变，旧的标注数据上得到的模型无法用于新数据的分类，这就需要对新数据进行标注，而人工标注工作需要极大人力物力代价。为了解决这个问题，本文做了以下几个方面的工作：

1. **构建了微博和新闻数据集，并对数据进行了分析**

目前国内外对微博的研究大多在 Twitter 上进行，为了获得中文微博上真实的结果，本文采集了信息微博的数据；由于目前没有公开的有标注的微博数据，本文利用微博的哈希标签，构造了六个类别的微博数据集。然后，从类别分布和词项分布分析了微博数据变化的特点。利用新闻门户网站的类别体系，采集了六个类别的新闻数据，并分析了新闻和微博的联系和区别。

2. **探索同源的迁移学习方法，有效利用旧标注数据**

微博数据更新快导致微博数据统计性质发生变化，直接利用旧标注数据对新数据分类难以取得好的效果，为了更有效利用旧标注数据，本文探索了基于同源数据的迁移学习方法。从时间因素出发，本文采用了指数衰减的方法和选择性指数衰减方法；从数据分布特点出发，本文采用了迁移学习方法 TransferBoost。实验结果表明，在同源数据间进行知识迁移，选择性指数衰减的方法取得了更好的效果。当微博标注数据量很少时，模型训练不充分，上述几种方法均无法取得理想的分类效果。

3. **提出了跨源的迁移学习方法，用于从新闻向微博迁移知识**

为了在微博标注量很少的情况下提高分类效果，降低标注工作，本文使用新闻数据作为辅助数据，促进微博分类效果的提升。由于微博和新闻是不同来源、不同形式的数据，同类别的新闻和微博数据存在着文字表达方面的差异性。为解决这个问题，本文提出了两种跨源迁移学习算法，从新闻向微博迁移知识，并将同源迁移方法融合进跨源迁移的框架，以提高整体分类效果。

把新闻作为微博的先验知识，本文提出了基于参数先验的迁移学习方法；根据词在

新闻和微博中分布的差异，有选择的进行知识迁移，本文提出了基于可迁移度的迁移学习方法。

实验证明，本文提出的迁移学习的方法的分类效果要优于直接使用新闻数据的非迁移学习方法，可以更有效的利用新闻数据，避免其差异带来的误导。

在标注微博数据较少的情况下，基于可迁移度的迁移学习方法的分类效果要明显优于其他方法。在只有 3%微博标注数据情况下，所有类别上的 F1 宏平均值接近 90%。而且随着标注数据的增加，效果稳定提升，接近所有算法的最优水平。因此，采用 WtrNaiveBayes 方法可以显著减少微博标注量，达到降低微博标注工作带来的开销，解决本文提出的问题。

6.2 下一步研究方向

微博分类是一个综合的问题，在微博的表示如适合微博的分词算法，新词识别算法等，特征的选取如（解决稀疏性问题，词语失配问题，主题层次的特征表示等）等各方面都会影响到分类的效果，有很多的问题需要研究。限于时间和本人能力有限，在本阶段有些问题有待进一步研究，包括：

1. 类别先验概率估计

在微博中，识别单一的类别，将面对很大的类别非均衡问题。对类别先验概率的估计提高了难度，没有大量标注样本，很难对先验概率给予充分的估计。在本文工作关注于类别中词项的条件概率 P(w|c)，没有对 P(c)做改进。在微博分类中，内容繁杂，噪音多，类别非均衡问题很严重，下一步计划研究对 P(c)估计方法的改进。

2. 词项关系

本文提出的方法都是在朴素贝叶斯分类框架上的，朴素贝叶斯分类的假设是词项间条件独立。但这种假设在实际情况中常常是不成立的。特别是对于微博来说字数很少，同一个词，在不同的上下文环境中含义就不同。可以尝试两种方式来解决这类问题：1，采用 N-gram 方法，可以发现相邻词项间的关系。 2，发现强相关的词项，直接估计其联合概率。因为微博文本短，可以先挖掘频繁的词项模式，不受词项位置的限制。

3. 新闻类别标签的限制

本文提出的方法，需要利用新闻网站的类别体系来获得具有类别标注的新闻数据，如果微博中的类别在新闻网站的类别体系中没有，那么就不能直接使用本文提出的方法。可以从以下两个思路解决这个问题：1，仍利用新闻的类别体系，使用和微博中类别相近的新闻类别作为辅助数据，如使用新闻中“体育类”辅助微博中的“跳水类”，这样会造成新闻和微博更大的差异性，但是迁移学习的目的就在于在相似但不同的任务、领域或分布间迁移知识，我们可以将跨源迁移学习的方法和跨领域迁移学习的方

法进行融合，提高分类效果。2，对于非常特殊或具体的类别，如“某某事件”，可以利用其它外部资源，如根据该类别的特点，构造一个搜索词集合，利用该集合在搜索引擎的前 N 个结果构造该类别的数据集。

参考文献

[1] A．Java，x．song，T．Finin，eta1．Why we twitter：understanding microblogging usage and communities．Proceedings of the 9th web KDD and 1st SNA—KDD 2007 Workshop on Web Mining and Social Network Analysis，2007：56—65．

[2] Kwak, Haewoon, et al. "What is Twitter, a social network or a news media?". Proceedings of the 19th international conference on World Wide Web. ACM, 2010.

[3] Weng J, Lim E P, Jiang J, et al. Twitterrank: finding topic-sensitive influential twitterers. Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010: 261-270.

[4] Krishnamurthy B, Gill P, Arlitt M. A few chirps about twitter[C]//Proceedings of the first workshop on online social networks. ACM, 2008: 19-24.

[5] Romero D M, Meeder B, Kleinberg J. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter[C]//Proceedings of the 20th international conference on World wide web. ACM, 2011: 695-704.

[6] Sadikov E, Medina M, Leskovec J, et al. Correcting for missing data in information cascades[C]//Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011: 55-64.

[7] Wu W, Zhang B, Ostendorf M. Automatic generation of personalized annotation tags for twitter users[C]//Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 689-692.

[8] Zhao X, Jiang J, He J, et al. Topical key phrase extraction from Twitter [C]. Proceedings pf the 49th Annul Meeting of the Association for computational Linguistics (ACL), 2011: 379-388.

[9] Zhao W X, Jiang J, Weng J, et al. Comparing twitter and traditional media using topic models[M]//Advances in Information Retrieval. Springer Berlin Heidelberg, 2011: 338-349.

[10] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision[J]. CS224N Project Report, Stanford, 2009: 1-12.

[11] Jansen B J, Zhang M, Sobel K, et al. Micro-blogging as online word of mouth branding[C]//Proceedings of the 27th international conference extended abstracts on Human factors in computing systems. ACM, 2009: 3859-3864.

[12] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2011, 2(1): 1-8.

[13] Tumasjan A, Sprenger T O, Sandner P G, et al. Predicting elections with twitter: What 140 characters reveal about political sentiment[C]//Proceedings of the fourth international AAAI conference on weblogs and social media. 2010: 178-185.

[14] Bermingham A, Smeaton A F. Classifying sentiment in microblogs: is brevity an advantage?[C]//Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010: 1833-1836.

[15] 中国互联网络信息中心. 中国互联网络发展统计报告. 2013.

[16] Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359.

[17] 戴文渊. 基于实例和特征的迁移学习算法研究. 硕士学位论文. 上海交通大学, 2008.

[18] Daum, H., & Marcu, D. (2006). Domain Adaptation for Statistical Classifiers, 26, 101–126.

[19] Daum, H. (2007). Frustratingly Easy Domain Adaptation, (June), 256–263.

[20] Zhang, D., Liu, Y., Lawrence, R. D., & Chenthamarakshan, V. (2011). Transfer Latent Semantic Learning : Microblog Mining with Less Supervision, 561 – 566.

[21] Zhang, D., Liu, Y., Lawrence, R. D., & Chenthamarakshan, V. (2010). ALPOS: A Machine Learning Approach for Analyzing Microblogging Data. 2010 IEEE International Conference on Data Mining Workshops, 1265–1272.

[22] Sun, X., Wang, H., & Yu, Y. (2011). Towards effective short text deep classification. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11, 1143.

[23] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10,

[24] Sakaki, T. (2010). Earthquake Shakes Twitter Users : Real-time Event Detection by Social Sensors, 851 – 860.

[25] Raina, R., & Ng, A. Y. (2000). Self-taught Learning : Transfer Learning from Unlabeled Data, 759–766.

[26] Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. Proceeding of the 17th international conference on World Wide Web - WWW '08, 91.

[27] Nishida, K., Hoshide, T., & Fujimura, K. (2012). Improving tweet stream classification by detecting changes in word probability. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12, 971.

[28] Long, G., Chen, L., Zhu, X., & Zhang, C. (n.d.). TCSST : Transfer Classification of Short & Sparse Text Using External Data Categories and Subject Descriptors, 764 – 772.

[29] Liu, Z., Yu, W., Chen, W., Wang, S., & Wu, F. (2010). Short Text Feature Selection for Micro-Blog Mining. 2010 International Conference on Computational Intelligence and Software Engineering, 1–4.

[30] Forman, G. (2006). Approved for External Publication Tackling Concept Drift by Temporal Inductive Transfer, 20(August), 6–11.

[31] Nishida, K., Hoshide, T., & Fujimura, K. (2012). Improving tweet stream classification by detecting changes in word probability. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12 (p. 971).

[32] 吴薇. 大规模短文本分类技术研究, 硕士学位论文, 北京邮电大学, 2007.

[33] 张剑锋, 夏云庆, 姚建民. (2012). 微博文本处理研究综述[J].中文信息学报, 26(4), 21–27.

[34] S. Milgram. The Small World Problem, Psychology Today, 1967.

[35] Charu C. Aggarwal, Social Network Data Analytics,An Introduction to Social Network Data Analytics, 2011.

[36] Jie Tang, Jimeng Sun, Chi Wang and Zi Yang. Social Influence Analysis in Large-scale Networks, In ACM SIGKDD, 2009.

[37] Daum, H. (2007). Frustratingly Easy Domain Adaptation, (June), 256–263.

[38] T. Joachims. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for TextCategorization. In International Conference on Machine Learning (ICML), 1997:134-151

[39] Yang Y, Liu X. A re-examination of text categorization methods[C], Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999: 42-49.

[40] Langley P, Iba W, Thompson K. An analysis of Bayesian classifiers[C], Proceedings of the national conference on artificial intelligence. JOHN WILEY & SONS LTD, 1992: 223-223.

[41] Joachims T. Text categorization with support vector machines: Learning with many relevant features[M]. Springer Berlin Heidelberg, 1998.

[42] Joachims T. Learning to classify text using support vector machines: Methods, theory and algorithms[M]. Kluwer Academic Publishers, 2002.

[43] Quinlan J R. Induction of decision trees[J]. Machine learning, 1986, 1(1): 81-106.

[44] Freund Y, Schapire R, Abe N. A short introduction to boosting[J]. Journal-Japanese Society For Artificial Intelligence, 1999, 14(771-780): 1612.

[45] Freund Y, Schapire R E. Experiments with a new boosting algorithm[C]//MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-. MORGAN KAUFMANN PUBLISHERS, INC., 1996: 148-156.

[46] Schapire R E. A Brief Int roduction of Boosting. In : Proceedings of the 16 International joint Conference on Artificial Intelligence ,1999.

[47] J. J. a. C. Zhai, "Instance Weighting for Domain Adaptation in NLP," in Meeting of the Assoc. Computational Linguistics, June 2007, pp. 264-271.

[48] Y. X. X. Liao, and L. Carin,, "Logistic Regression with an Auxiliary Data Source," in Int'l Conf. Machine Learning, Aug. 2005, pp. 505-512.

[49] V. W. Z. S.J. Pan, Q. Yang, and D.H. Hu, "Transfer Learning for WiFi-Based Indoor Localization," in Workshop Transfer Learning for Complex Task of the 23rd Assoc. for the Advancement of Artificial Intelligence (AAAI) Conf. Artificial Intelligence, July 2008.

[50] W. Dai, G. Xue, Q. Yang, and Y. Yu, “Co-Clustering Based Classification for Out-of-Domain Documents,” Proc. 13th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, Aug. 2007.

[51] R. Raina, A.Y. Ng, and D. Koller, “Constructing Informative Priors Using Transfer Learning,” Proc. 23rd Int’l Conf. Machine Learning, pp. 713-720, June 2006

[52] P.Wuand T.G. Dietterich, “Improving SVM Accuracy by Training on Auxiliary Data Sources,” Proc. 21st Int’l Conf. Machine Learning, July 2004

[53] A. Arnold, R. Nallapati, and W.W. Cohen, “A Comparative Study of Methods for Transductive Transfer Learning,” Proc. Seventh IEEE Int’l Conf. Data Mining Workshops, pp. 77-82, 2007.

[54] X. Ling, G.-R. Xue, W. Dai, Y. Jiang, Q. Yang, and Y. Yu, “Can ChineseWeb Pages be Classified with English Data Source?” Proc. 17th Int’l Conf. World Wide Web, pp. 969-978, Apr. 2008.

[55] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. the Journal of machine Learning research, 2003, 3: 993-1022.

[56] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[57] 文坤梅, 徐帅, 李瑞轩等. 微博及中文微博信息处理研究综述[J]. 中文信息学报, 2012, 26(6):27-37.

[58] 靖红芳, 文本分类中特征选择的形式化研究, 硕士学位论文, 中国科学院计算技术研究所, 2009.

[59] 靳小波. 文本分类综述[J]. 自动化博览, 2006, 23(z1), 24.

[60] 李航. 统计学习方法.清华大学出版社. 2011.

[61] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze, 王斌（译）.信息检索导论, 人民邮电出版社, 2010.

[62] 李开复. 微博：改变一切. 上海财经大学出版社, 2011.

[63] Raina, R., & Ng, A. Y. (2000). Self-taught Learning : Transfer Learning from Unlabeled Data, 759 – 766.

[64] Nishida, K., Hoshide, T., & Fujimura, K. (2012). Improving tweet stream classification by detecting changes in word probability. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’12 (p. 971).

[65] Lee C H, Wu C H, Chien T F. BursT: a dynamic term weighting scheme for mining microblogging messages[M]//Advances in Neural Networks–ISNN 2011. Springer Berlin Heidelberg, 2011: 548-557.

[66] Lee C H, Yang H C, Chien T F, et al. A novel approach for event detection by mining spatio-temporal information on microblogs[C]//Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on. IEEE, 2011: 254-259.

[67] Eaton E, Lane T. Modeling transfer relationships between learning tasks for improved inductive transfer[M]//Machine Learning and Knowledge Discovery in Databases.

Springer Berlin Heidelberg, 2008: 317-332.

[68] Eaton E, desJardins M. Selective transfer between learning tasks using task-based boosting[C]//Twenty-Fifth AAAI Conference on Artificial Intelligence. 2011.

[69] Chelba C, Acero A. Adaptation of maximum entropy classifier: Little data can help a lot[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Barcelona, Spain. 2004.

[70] Argyriou A, Evgeniou T, Pontil M. Multi-task feature learning[C]//Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference. The MIT Press, 2007.

[71] Lawrence N D, Platt J C. Learning to learn with the informative vector machine[C]// Proceedings of the twenty-first international conference on Machine learning. ACM, 2004.

[72] Bonilla E, Chai K M, Williams C. Multi-task Gaussian process prediction[J]. 2008.

[73] Schwaighofer A, Tresp V, Yu K. Learning gaussian process kernels via hierarchical bayes[J]. Advances in Neural Information Processing Systems, 2005, 17: 1209-1216.

致谢

时光荏苒，三年的时间如白驹过隙，非常有幸能够来到中科院计算所，在这里领略大师的风采，聆听专家的教诲，潜心地科研提升自我的能力，激烈地讨论碰撞出思想的火花。回顾三年的生活，忙碌又充实，在这即将毕业之际，向所有陪伴、关心、帮助过我的人表示衷心的感谢！

首先，我要感谢我的导师王斌老师。王老师具有严谨的治学态度、渊博的学术知识和敏锐的洞察力，在学术研究上给予了我悉心的指导。王老师给我们空间去探索感兴趣的问题，又能在我无所适从时一针见血地指出问题的关键，把握正确的方向。感谢王老师分配的每一个项目，让我有机会真刀实枪地解决实际问题，提高了我的实践能力，也加深了对理论的理解。在生活上，王老师对我们的关心是无微不至的，关心我们的健康，关心我们的成长，他像大朋友一样分享他的人生经验，也会和我们一起谈古论今。当我们遇到难题时，老师总会设身处地的为我们着想，毫无保留地给我们真诚的建议和无私的帮助。王老师正直无私的品格、积极的人生态度、为人处世的方方面面都是我学习的榜样，能够成为王老师的学生是我的幸运。

感谢信息检索课题组的师兄弟姐妹们，和你们一起度过了太多难忘的日子。当我在研究上遇到难题时，你们会放下自己的工作和我一起研究解决方案；当我面对紧张的项目进度时，你们和我一起通宵达旦地加班；当我处于迷茫和低谷的时候，你们给了我无条件的支持、关心和鼓励。感谢李鹏、李锐、史亮、卫冰洁、张冠元、邱泳钦在学术研究、工程项目和论文写作方面给我的指导和建议。感谢鲁骁、王书鑫，不分彼此地赶项目、并肩战斗的日子是最难忘的。感谢吉宗诚、袁平广、鲁凯、徐飞，你们深厚的学术功底和活跃的思维给我的研究工作带来了很多启发。感谢李文娜、徐安滢，是你们的努力让实验室的生活更加丰富多彩。

感谢计算所 703 班的同学们，感谢你们对班委工作的支持和对班级的热爱。篮球赛、足球赛、羽毛球赛我们一起挥洒汗水；歌唱比赛、元旦晚会我们一起纵情歌唱；爬香山、游北海、我们一起放松身心、亲近自然。我们班虽然不是体育最强、歌声最动听的，但永远是那么的团结，在集体中的日子总是充满着快乐。

感谢前瞻研究实验室的刘玉东、任菲、刘卫玲老师，感谢研究生部的李琳、宋守礼、周世佳、李丹、张平等老师，感谢你们对我们平时工作和学习的支持，你们的辛苦工作使得我们的学习和科研顺利进行。

最后，我要感谢我的父母，你们给了我生命，你们抚育我成长，你们教会我做人。你们从来没有给我任何压力和要求，你们给我的是充分的信任和无条件的支持。感谢我女友，感谢你的一路陪伴，不论荆棘还是坦途，我们一起走过。你们无私的爱是我坚强的后盾和不竭的动力。

作者简介

姓名：张帅　　性别：男　　出生日期：1987.9.2　　籍贯：河南安阳

2010.9 – 2013.7　　中科院计算所计算机应用技术硕士研究生
2006.9 -- 2010.7　　武汉大学国际软件学院软件工程专业本科生

【攻读硕士学位期间发表的论文】

[1]. **Shuai Zhang**, Kai Lu, Bin Wang, ICTIR Subtopic Mining System at NTCIR-9 INTENT Task, Proceedings of NTCIR-9 Workshop Meeting, p.106-110, December 6-9, 2011, Tokyo, Japan

[2]. Bingjie Wei, **Shuai Zhang**, Rui Li, Bin Wang，A Time-Aware Language Model for Microblog Retrieval，trec 2012

[3]. Kai Lu, Guanyuan Zhang, Rui Li, **Shuai Zhang**, and Bin Wang, Exploiting and Exploring Hierarchical Structure in Music Recommendation, the 8th Asia Information Retrieval Societies Conference(AIRS 2012), pp.211- 225, 2012.

【攻读硕士学位期间申请的专利】

[1]. 李锐，**张帅**，张冠元，王斌，李鹏，鲁凯。专利名称：回归预测方法及装置。申请号：201110339244.1，申请时间：2011 年 11 月 1 日（已公开）

[2]. 史亮，王斌，卫冰洁，李锐，**张帅**，张冠元. 一种用于搜索引擎倒排索引压缩的基于排序的文档序号重排方法. 中国，发明专利，201210401317 ，2012 （已申请）

[3]. 史亮，王斌，李鹏，李锐，卫冰洁，**张帅**. 一种用于搜索引擎倒排索引压缩的文档序号表示方法. 中国，发明专利，专利号暂无，2013 （已申请）

【攻读硕士学位期间参加的科研项目】

[1] 第九届 NTCIR 信息检索评测子话题挖掘任务，2010 年 12 月 - 2011 年 12 月

[2] 国家电网智能云数据分析平台项目，2011 年 8 月 - 2012 年 6 月

[3] 基于微博的热门话题发现研究，2012 年 3 月 - 2012 年 8 月

【攻读硕士学位期间的获奖情况】

[1] 2010-2011 年，中国科学院计算技术研究所华为硕士生奖学金

[2] 2011 年被评为中国科学院研究生院“优秀学生干部”，“三好学生”，“优秀共产党员”

[3] 2012 年被评为中国科学院研究生院“三好学生”，获得研究生国家奖学金