****

**本科生毕业论文(设计)**

**毕业论文（设计）题目：**

**基于新浪微博热点话题的网络舆情分析**

**部 院 系： 政府管理学院**

**专 业： 信息管理与信息系统**

**学 号： 201111231031**

**学 生 姓 名： 黄勇**

**指 导 教 师： 刘晓娟**

**指导教师职称： 副教授**

**指导教师单位： 北京师范大学政府管理学院信息管理系**

**2015年5月10日**

**北京师范大学本科生毕业论文（设计）诚信承诺书**

本人郑重声明： 所呈交的毕业论文（设计），是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

本人签名： 年 月 日

**北京师范大学本科生毕业论文（设计）使用授权书**

本人完全了解北京师范大学有关收集、保留和使用毕业论文（设计）的规定，即：本科生毕业论文（设计）工作的知识产权单位属北京师范大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许毕业论文（设计）被查阅和借阅；学校可以公布毕业论文（设计）的全部或部分内容，可以采用影印、缩印或扫描等复制手段保存、汇编毕业论文（设计）。保密的毕业论文（设计）在解密后遵守此规定。

本论文（是、否）保密论文。

保密论文在 年 月解密后适用本授权书。

本人签名： 年 月 日

导师签字： 年 月 日

基于微博热点话题的网络舆情分析

摘　要

随着社交网络的发展成熟，人们在社交网站上更加关注公共事务，评论热点话题的意愿也很高。本研究通过获取“亚投行”这一热点事件相关的微博，对相关的微博进行了影响力统计分析；对于这些微博的内容作了短文本分析，不仅采用词向量模型搜寻相关话题中的上下文相关词语，还通过潜在语义分析获取部分主题的重要关键词；同时采用朴素贝叶斯的方法对于微博数据进行了情感态度分析。结果发现，关注“亚投行”的人群有着较好的经济基础，“亚投行”的相关主题主要涉及中日美澳等国家，关注点在于创建亚投行这一组织；同时，通过微博数据的情感分析，结果发现大部分微博数据对于亚投行这一事件都有着支持的态度，部分表达负面情感态度的微博主要集中在对于日本和美国的不满上。

**关键词：**新浪微博，短文本分析，词向量，情感分析，潜在语义分析

Text Mining of Public-Concerned Topic in Weibo

Abstract

With the development of social network, People get more and more concerned on public topics, they also get engaged in hot topics and express their personal emotions. In this research, we study the hot topic about AIIB, which means Asian Infrastructure Investment Bank. Not only do we focus on the descriptive statistics describing the importance of this topic, but also do we use text mining methods learning the hidden semantics in this topic. Finally, we can get clear that many people approve the idea Beijing will establish AIIB which will help China a lot. What’s more, China gets the highest attention above all countries.

**KEY WORDS：**Weibo, Text mining, term vector, Latent Semantic Analysis, Sentiment Analysis

**目　录**

[摘　要 5](#_Toc419809110)

[ABSTRACT 6](#_Toc419809111)

[1 引言 8](#_Toc419809112)

[2 相关研究概述 8](#_Toc419809113)

[2.1 网络热点话题的信息传播机制研究 9](#_Toc419809114)

[2.2 网络热点话题的时间和空间影响力研究 9](#_Toc419809115)

[2.3 网络分析和观点挖掘 10](#_Toc419809116)

[2.4 基于微博的推荐系统研究 10](#_Toc419809117)

[3 微博数据的统计描述分析 11](#_Toc419809118)

[3.1 微博数据的特征选择 11](#_Toc419809119)

[3.2 微博影响力的统计描述 11](#_Toc419809120)

[3.3 微博内容的数据预处理 13](#_Toc419809121)

[4 微博文本的舆情分析 14](#_Toc419809122)

[4.1 基于word2vec的词向量模型 14](#_Toc419809123)

[4.2 基于朴素贝叶斯的情感分析 15](#_Toc419809124)

[4.3 基于LSA的潜在语义分析 17](#_Toc419809125)

[5 结语 18](#_Toc419809126)

# 引言

随着社交媒体的日益成熟完善，信息的传播也在发生变化，面对网络热点事件，社交网络成为信息共享和交流的平台，也成为负面消息、虚假消息的传播平台，一些社交网络平台如微博、Twitter在信息传播过程中担任着重要的角色。如何更成熟应对复杂的网络环境，分析挖掘有意义的微博用户行为，了解用户的偏好和所属社区，并了解这些用户在热点事件传播过程中承担角色，成为一个重要的任务。同时，对网络热点话题的分析，有利于更好监控了解舆情，并针对舆情制定相应的方针政策。

微博也被称为微型博客，是一种通过关注机制分享简短实时信息的广播式的社交网络平台。微博是一个基于用户关系信息分享、传播以及获取的平台。用户可以通过Web、WAP等各种客户端组建个人社区，以不超过140字的文字发表信息，并实现即时分享。微博的用户关注机制分为可单向、可双向两种。微博是中国流行的分享和交流平台，其更注重时效性和随意性。微博更快更方便表达自己的思想和动态，逐步改变着信息传播和交流的方式。和微博相似，Twitter在全世界有广泛的用户，数据显示，截至到2012年3月，Twitter户每天会发表约3.4亿条推文。同时，Twitter每天还会处理约16亿的网络搜索请求。Twitter被形容为“互联网的短信服务” [[[1]](#endnote-2)]。如今Twitter全球注用户数量已经超过10亿，有70%的流量来自于美国以外的地区。

热点话题涉及一系列热点事件，包括事件的发生、传播、消亡和后续影响等。而微博和Twitter蕴含着海量的数据，其中很多数据涉及普通民众对于新闻热点事件报道、转发以及情感价值取向。对于这些海量数据的统计分析、情感分析，有助于我们更好了解热点话题事件发生的过程、影响范围、以及民众对于热点事件的价值判断和情感取向。如果能够有效发掘话题事件背后隐藏的规律，预测热点事件发生的时间、地点、波及范围，提前做好准备预防工作，并及时做出反应，可以有效减小某些突发事件带来的损失。而对于用户行为特征的分析，对于热点事件发生过程的中的传播规律研究，可以有助于我们做好舆情监测、舆情分析、甚至是舆情预测。其次，关注热点事件、转发热点事件的人物相互之间的关系组成一个复杂的社交网络，对于这个社交网络中每个节点人物的统计分析，可以更好帮助我们相关用户的行为特点。

# 相关研究概述

对于微博和twitter的数据分析最近一直是统计学、数据挖掘和机器学习、信息计量学、信息社会学等多个学科深入研究的热点，国内外也有很多从不同视角进行研究的论文。总体来说可以分为如下几类：（1）概括分析提出微博信息扩散的模型，指出热点事件传播的特点；（2）统计分析微博在时间和空间上的影响力分布，利用时间序列等模型分析预测突发事件对公众的影响方式；研究微博中人与人之间行为关系之间的相互联系和作用，发现有影响力的人物。（3）对微博等社交数据内容进行分析，采用自然语言处理和词向量模型，通过聚类分析挖掘公众对于突发热点事件的关注点；(4)利用微博数据进行文本处理，实现微博数据的短文本分析，在此基础上实现微博等社交网站的推荐系统。

* 1. 网络热点话题的信息传播机制研究

这类研究通过构建信息传播的模型和传统的传播学结合，并通过一定的数据发现热点话题事件的传播规律。蒋盛益[[[2]](#endnote-3)]等人指出微博研究主要包括话题趋势检测、情感倾向性分析等方面，用户关系挖掘主要包括用户群体特性、用户社区发现、意见领袖挖掘及微博传播模式等方面，并指出这些方法的局限性，并对微博信息挖掘的发展进行展望，期望能发现微博传播方式。李晓明，邱泽奇[[[3]](#endnote-4)]等人研究了网络热点的传播机制并指出三种信息传递方式：由影响力较大的信息源向外辐射信息；局部突发并在极短时间内产生级联效应，但会迅速消失；社会文化决定的周期性流行，如关于节日等，这类传播比较固定。这些信息传播有着不同的传播方式： 点源流行 、重复暴露同源流行、蔓延流行、 周期性爆发等特征。这些对于我们深入量化信息传播模型有着重要的意义。而在《危害公共安全的关联事件挖掘》[[[4]](#endnote-5)]一文中，作者利用海量数据的优势，对新闻事件分类标注、扩充语料库、语境过滤并实现对于不同事件的分类。并采用无监督自主学习获取词语的语境关联。通过对于新闻要素的多角度要素提取，对事件空间、时间、和媒体报道特征的关联分析和事件预测，掌握了公共事件的发生、传播特点，分析出危害公共安全的大事件后发生常常伴随着同类的小事件等结论。

* 1. 网络热点话题的时间和空间影响力研究

通过采集发布或者转发微博的时间、空间信息，利用时间序列分析或者基于LBS研究不同时间段、不同地域人们发布信息数量变化，发现在热点话题事件中人们情感变化等。Predicting Flu Trends using Twitter Data[[[5]](#endnote-6)]文中，Harshavardhan Achrekar[[[6]](#endnote-7)]等人分析了Google Trend和twitter关于流感数据的相关性，爬取twitter上与流感相关的推文，并分析了这些信息发出的地理位置；同时对每天、每周随着时间变化推文数量的变化，构建出相应的时间序列模型，作者还通过聚类分析，预测了流感在全美地区的爆发趋势，将美国的流感区域聚类划分。An Analysis of Twitter Messages in the 2011 Tohoku Earthquake[[[7]](#endnote-8)]文中，作者关注了日本地震中twitter和Facebook中用户对于突发公共安全事件的意见，分别分析英文和日文的tweets，采用时间序列和自然语言处理的方法，分析出不同时间阶段民众对于事件的了解和恐慌程度。最后指出，不同语言的推文数量有着强相关关系，关注的内容也很相近，但是国内语言在时间传播中起着重要的作用。同时，随着时间的变化，人们的情感态度也随之发生微妙的变化。

* 1. 网络分析和观点挖掘

利用网络分析，可以分析处网络中人物之间的亲密程度，一个人被关注和关注人数的多少，更重要的是利用图的相似度计算相同社区中是否出现相似的用户，这些相似的用户在同一热点话题事件中是否扮演者相同的角色。而自然语言处理技术，则可以挖掘用户在热点事件中的情感价值取向，发现用户的关注点。Sentiment Analysis and Time Series with Twitter[[[8]](#endnote-9)]一文中，作者分析了twitter上某个特定时间段内推文数量变化、以及推文中正面、负面情感变化趋势等。并期望能使用tweet和微博的数据同时进行分析。Sentiment Analysis for Tracking Breaking Events: A Case Study on Twitter[[[9]](#endnote-10)]中，作者收集11亿条推文，采用LIWC (Linguistic Inquiry Word Count)[[[10]](#endnote-11)]进行情感分析，发现人们在下午发布推文情感偏向负向而晚上发布偏向正向的推文，同时作者也通过大量推文预测出某些重要事件的发生规律。Network Analysis of Three Twitter Functions: Favorite, Follow and Mention [[[11]](#endnote-12)]一文中，作者分析了从网络分析的视角研究twitter上的三种功能（Favorite、Follow、Mention），并将这些功能对应于三种网络关系：简单图、二分图和多重图，通过现实世界中的twitter数据分析这些复杂网络中的无标度分布特征、主要网络的相互联系、Gini系数、Favorite和Mention功能在图的结构上的不同。并指出这些研究对于构建twitter用户偏好的模型非常有用。

* 1. 基于微博的推荐系统研究

除了简单的自然语言处理，更多的研究者关注于挖掘微博的文本信息，通过数据挖掘的方式找到用户对于热点事件的情感态度，并采用推荐系统满足用户在热点事件传播过程中的信息需求，为用户提供有实际意义的新闻、微博信息。Finding question microblog post and recommend answers for it[[[12]](#endnote-13)]文中，作者采用支持向量机分类器（SVM）来分析微博正文，提取情感特征、事件特征、词语情感特征、外部事件特诊。通过有监督的正向、负向情感学习，获得了较好的结果。谢丽星，周明[[[13]](#endnote-14)]等人采用情感分析研究。对三种情感分析的方法进行了深入研究，通过实验表明基于SVM的层次结构多 策略方法效果最好，并针对层次结构的多策略方法的特征选择进行了详细分析，包括话题无关、话题相关的特征。

# 微博数据的统计描述分析

考虑普通民众对于隐私的要求越来越高，新浪微博等社交媒体都采用了严格的数据保护手段，因而微博和Twitter数据获取变得越来越难。同时，考虑到很多热点话题涉嫌一些公司的炒作，我们最终选取了具有时事性的热点话题“亚洲投资银行”，为了考虑数据的完整性，我们将涉及“亚投行”、“AIIB”等相关的关键词也包含在自己的爬虫系统中，通过抓取从2015年3月12日到2015年4月14日期间的数据，获得亚投行相关的新闻的微博数据。

# 微博数据的特征选择

通过多次抓取新浪微博上的关于亚投行相关话题的微博数据，最终，我们获得了10M的数据集。考虑到多次的爬取新浪微博数据会产生数据的重复抓取，同时，为了将微博数据的JSON格式转化为二维表的格式，我们自己实现了数据的快速清洗存储。通过更新MySQL数据库的方式处理重复数据，最终获取数据5708条数据集，涉及21个相关属性，主要包括微博ID、微博发布者ID、微博发布工具、微博发布内容、微博被浏览量、微博被评论量、微博被赞次数、微博发布时间、微博发布图片地址、微博发布地理经纬度等数据。由于新浪微博对于用户隐私的限制等因素，微博被浏览量、微博发布地理位置等数据获得量较少，很难进行数据分析，此外还有用户的ID等属性也被筛选剔除。最终我们选出用户微博内容、发布时间、评论数、点赞数、转发数量、转发来源等属性。

# 微博影响力的统计描述

新浪微博是一个公共开放的平台，一个人在微博上发表博文被浏览的次数，微博被评论、点赞、转发的次数可以显示一个微博的影响力。我们对所有的抓取微博数据的相关属性进行统计分析，期望能发现在亚投行事件中，各个关注相关事件的微博影响力分布情况，从而了解亚投行事件总体的影响力情况。除了通过分析微博中内容（content）被浏览转发的次数，我们还可以收集用户了解这些微博内容的渠道（source），或者称之为发布工具，通过对这些发布工具的描述统计，我们可以看出相关的关注群体都有哪些特点。

通过对新浪微博数据的清洗和预处理，最终选出有效数据5708条，其中包括每条微博获得的评论数（commentCount）、被赞数（praiseCount）、转发数（reportCount）。这些数据超过75%都是0，我们通过统计软件对这些样本数据进行了统计描述。这些数据描述结果如下表1所示，通过对评论数、被赞数、转发数这三种属性的分析，我们发现近80%的微博获得0评论、0转发、0被赞。但是所有微博的平均转发数、被赞数都超过100，同时根据三种属性分布的偏度和峰度，可以看到不同的微博获得关注度差异非常大，而且这些数据和微博的用户行为习惯相吻合。此外，被赞数和被评论数量有着极高的相关度，甚至呈现出线性的相关关系，这可能和微博用户的行为习惯有关。

表 1 点赞、转发和评论统计表



除了对于三种相关属性的数据进行描述统计，我们还处理了用户发送微博采用的方式（source）。通过了解用户发送微博的方式、可以清楚知道用户更倾向于从那些渠道了解新闻、发布新闻。我们通过分析5708条微博发生的source，将这些来源分成：新闻媒体类（以人民日报、新华网、新浪网为首的全国各大官方新闻机构）、新浪微博网页类、新浪博客类、桌面浏览器类（如360浏览器、搜狗浏览器等）、iPhone类、Android类、其他社交媒体类（知乎、孔明社交管理器）、分享按钮类等，从而绘制了亚投行话题来源分布图，如图1所示。

图 1亚投行话题消息来源分布图

通过图1我们可以发现，超过1/15的亚投行消息来源于国内各大新闻媒体机构。而且发布新闻采用iPhone和Android数量相当，考虑到iPhone使用者消费水平较高，我们可以基本了解认为关注亚投行的人群生活水平比较高，社会地位也比较高，这说明关注亚投行的相关人群可能收入、消费等某些方面比普通民众要好。同时，亚投行新闻在多种来源中分布很均匀，这和一般的娱乐热点事件不同传播的渠道不同，也不同于新浪微博定位的移动社交平台。

# 微博内容的数据预处理

微博内容数据的预处理涉及问题很多。首先是数据的编码格式问题。由于微博是一个提供人们娱乐休闲方式的移动社交工具，微博因此设计了很多特有的编码符号，有些输入法厂商也为了吸引用户，特意设计很多不在utf-8编码格式下的符号。为了方便我们处理运行数据，本研究过程特意筛选剔除所有非utf-8的编码，并且为了研究方便，去除了表情和符号，如“//@”等。

微博内容数据的预处理另外一个问题在于中文分词。中文分词是一项非常有挑战性的工作内容，相关的算法设计“前向最大匹配”、“后项最大匹配”、“隐马尔科夫”等方法。同时，微博内容不能超过140个字，这就大大限制了中文表达的内容，更容易使用户采用精简的表达方式，表示自己的潜在语义。同时，这也给中文带来了非常艰难的挑战。为了能够最大限度的挖掘出微博内容的潜在语义，本研究中，对于微博数据进行了全切割（full segment）。全切割采用最大前向词匹配的方式扫描词典。生成所有基于句子的有向无环图。并采用动态规划的核心算法查找概论最大的路径，从而找出词语的最大次切分。除此之外，全模式切割词语可以提供所有备选词的分词结果，保证中文分词的准确率达到最大，但同时也会降低召回率，产生很多噪声。此外，我们分词过程中还使用了隐马尔科夫模型对某些未出现的新词进行识别，如“亚投行”等。

中文分词的存储是另外一个重要的任务。有很多重要的词语存储方式。本研究中采用最常见的词袋模型（bag of words），将所有的分词结果按照<词语,词语ID>的存储模式进行编码存储。从而期望能够基于词向量空间的模型获得词语在上下文语境中的隐含语义。

除了基于微博被转发、点赞次数等信息分析一个新闻事件的传播方式、传播影响力来源。我们更希望通过了解微博的文本关注的内容，通过对微博背后的文字进行分析处理，我们希望能了解一次新闻事件发生中民众对于微博事件的关注点、涉及相关的主题、高频词和与高频词相关的同类词、普通百姓对于新闻的支持/反对的态度等。我们采用自然语言处理的方法对微博的短文本进行了中文分词，并基于中文分词的结果，采用了词袋的方法对微博文本进行聚类分析。希望挖掘出潜在的语义，从而更加深入了解新闻关键点。图2是基于词频统计生成的词云。图中显示了词频超过100的词语分布情况，这些词的最大词频为7918，词频的中位数为193，平均数为423。图中可以明显看出中国、银行、基础、加入、澳大利亚等高频词语，除此之外，中日美澳等国家在词频中也占有非常重要的位置。

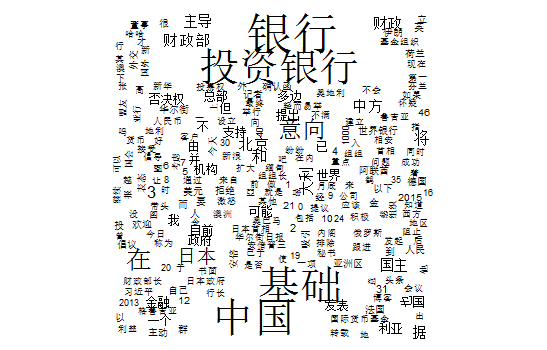


图 2 “亚投行”高频词词云

# 微博文本的舆情分析

基于微博文本的分析采用文本挖掘的方法，对文本进行数据中文分词处理，并采用词向量的方法对文本进行近义词、反义词查找，可以确定文本中的某些联系非常紧密的关键词、通过词向量(word2vec)[[14]](#endnote-15)的方法，不仅可以找到语义非常接近的关键词，而且通过获得每个词的向量，可以有效进行聚类分析，获得文本的相关主题；基于贝叶斯的情感分类[[15]](#endnote-16)，可以有效确定微博文本中的情感态度，通过计算微博中各个时间点每条微博postive的概率，可以研究分析整体的微博中情感态度的变化；基于LSA的文本主题聚类[[16]](#endnote-17)，通过研究每个主题中的最相关的词语，也可以有效发现文本中的不同的主题。

# 基于word2vec的词向量模型

词向量模型采用一种词袋模型（BOW）的方式对中文词语进行编码。词向量模型的输入通过对每个词语赋予一定的ID，使得每个词语在词向量空间中有自己独特的向量空间表示。我们采用word2vec这种基于深度学习的模型来训练自己的数据集，期望能通过词向量的深度学习，理解文本中隐含的语义关系，从而能获得对文本更加深入的分析。

词向量模型是一种三层神经网络模型，输入层是基于BOW的词模型；而中间隐层则是通过对词向量模型的训练生成模型，获得神经网络的输出；神经网络的输出层基于训练的结果进行分类，利用词频统计的方法生成最优的霍夫曼树；霍夫曼树的生成结果包含多个叶节点，每个叶节点都是一个具体的语料库中的单词，而输出中每个非叶节点都是表示一类词语。这样每个词语都能获得一个自己的独特的词向量，从而可以对这些词语进行聚类分析，了解一个语料库中的文档主题等信息；或者了解一个或者若干个词语中上下文最相关的若干个词语。本次研究中，中国和亚投行是两个极其重要的关键字，因而我们选取了关键词“中国”、“亚投行”，期望找到和这个词组最相近的一组关键词。同时考虑到“亚投行”这个事件背景在于北京在于中国期望借助亚投行增加自己的国际地位，需要和美国进行权力的斗争，我们设定“美国”作为最不相近的词语作为反义词，从而找出新的一组和“中国”、“亚投行”最相近的词语。

表 2“亚投行”中关键字的最近邻词表



分析可以发现，当以中国和亚投行作为目标词汇，搜索最邻近的词语时，可以发现“国主”、“否决”，“设立”等词汇和中国的关系最为相近。当设置“美国”作为最不相关的词汇进行搜索时，新词“国发”、“支持”、“一带”、“中国财政部”、“银行业”等关键词和“中国”、“亚投行”关系最接近。联系本研究的新闻事件的性质和主题，可以发现亚投行涉及相关“中国”、“美国”等词语中的相关词语。

# 基于朴素贝叶斯的情感分析

情感分类在公共话题中也有很重要的作用，通过对新浪微博中，关注“亚洲基础投资银行”话题的人的数据进行分析综合，可以了解这些人对于亚投行这种重要的热点事件的情感态度，从而分析民众对于热点事件基本的情感态度，获得人们对于热点事件的思考评价等信息。而了解掌握这些基本的信息可以帮助我们了解舆情舆论，掌握大众的情感态度，把握舆情、控制舆论有着非常深远的影响。

考虑到本次选取亚投行为新浪微博话题，这类新闻类事件并不存在人们通常表述的喜怒哀乐等情感人，人们表达态度更多的是采用一种赞同和反对的态度来看待事件，基于话题的特殊性。本研究决定将所有的微博进行二分类，期望发现在微博中人们对于亚投行事件的赞同和反对的态度。而选用训练集样本进行训练数据的时候，我们选择了常见的表达人们正面积极和负面消极的词汇组成了训练数据，这个数据集并不完全，希望能获得更多的训练标准。

我们对于情感分类建模的方法基于朴素贝叶斯理论。每条微博的情感态度决定于这条微博中所有词语表达的情感。我们通过收集常见的新浪微博中表达情感的词语，将词语分成积极的训练集和消极的训练集。利用本次研究数据集作为测试集。通过对于测试集进行预测发现正向和负向的微博情感。

我们通过朴素贝叶斯模型训练，对新浪微博这个话题正负情感分类做了预测。图3展示了新浪微博从3月12日到4月14日之间微博正面、负面消息的变化图。其中，每个散点纵坐标表示一个词语的可能为正向态度的概率，越接近1，表示这条微博表达的态度越积极正面，反之，越接近0，这条微博和积极正能量的相关性就越小，每个散点的横坐标表示这条微博创建的时间点，所有采集的微博数据范围介于2015年3月12日到2015年4月14日之间，拟合的曲线表示随着时间段的变化，所有采样微博总体情感态度的变化。从整个时间变化曲线来说，大部分微博都对于“亚投行”有着客观积极的评价，而少量负面情绪表达集中在对于美国、日本阻碍亚投行发展的情感表达的微博上。

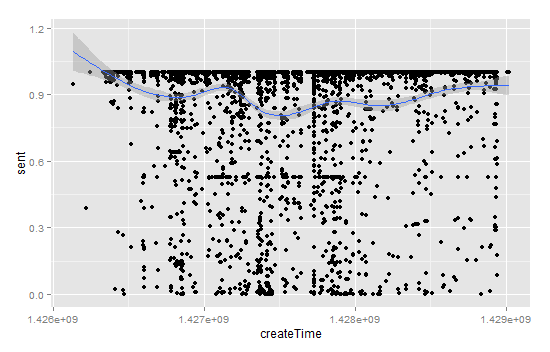


图 3 “亚投行”话题情感态随度时间变化图

具体来说，通过对于所有的微博数据记录每个微博可能为正面的概率，如对于微博“美国拆台亚投行的戏要演多久”这条微博通过计算其为积极的概率得到0.35，而相似的微博“韩国夹在中美之间左右为难”也只有0.23的积极的概率，类似的还有“唉，可怜的看门狗啊……”等微博。由此可见这类微博可能表达非正面的积极情感态度。同理对于“中国棒棒哒”“ 这是我们中国的魅力”这样的微博，计算情感概率为0.99，表示这条微博更倾向表达正向情感态度。从而获得了微博的情感分布图，结合获得的时间序列，从而可以了解这个热点话题表达的情感波动情况。由此可以看出，亚投行事件在中国国内的各家媒体中的宣传都偏向于正面、态度积极。部分少量消极的情感表达也都是表达了自己对于美国、日本行为态度的不满。

# 基于LSA的潜在语义分析

为了能深入研究微博的文本分析。不仅需要了解掌握微博中用户的情感态度，分析一些词与词之间的关联关系，更重要的是，通过对微博文档的分析处理，了解这些微博文档设计的文章主题思想，以及这些主题思想表达的关联关系等[[17]](#endnote-18)。

具体来说，LSA采用了奇异值分解的方法来处理文档-特征矩阵。而矩阵中的每一个元素使用这个文档中选择特征的TF-IDF值来表示。从而获得文档的特征矩阵。将文档的特征矩阵按照奇异值分解的方法处理，可以获得文档在各个特征值下的概率结果，从而我们可以深入了解文档的主题思想，包含的关键词等[[18]](#endnote-19)。

对于每个文档中每个词语的TF值，表示这个词语在这个文档中出现的频率，TF值越高，表明这个词语的出现频率越大，IDF表示总文档数和一个文档中某个词语在全部文档中出现频次的商，这个商取对数表示为逆向文档频率（IDF），IDF的值越大，表明这个词在所有文档中出现的频次越低，TFIDF的计算公式是TF和IDF的乘积，使用TFIDF代替原来的词频统计方法可以更容易找出一篇文档中的关键词，并且消除停用词语对于文档潜在语义发现的影响。

表 3“亚投行”中各个主题中关键词语



本研究选取了top-10个LSA，并列举每个潜在语义中最主要的前8个词汇，通过对比LSA1和LSA2关注的关键词，可以发现LSA1的关注主题在于亚投行、银行等，LSA2关注中国、美国等主题词语蕴含的潜在语义。通过对于top-10个LSA包含的关键词进行潜在语义分析，可以发现前十个LSA的话题关注主要集中于“亚投行”、“中国”、“否决权”、“北京”、“日本加入”、“日方申请”、“金融机构和国际货币基金组织”、“澳大利亚”这些相关事件。

总体来说，通过图三可以看出，亚投行的关键词集中在中国、亚投行、否决权、日本、新闻、澳大利亚这些主题上。而这些主题关键词正是2015年3月到4月期间，亚投行新闻事件的关键点，亚投行事件的这些关键词，基本可以代表亚投行中的事件关注的焦点。通过采用隐含语义分析，我们可以快速把握一个事件的主题关键词，从而发现普通群众在微博中关注、讨论比较热门的话题。

# 结语

新浪微博在国内重大新闻事件传播过程中的影响力越来越大，能快速传播突发事件。目前各个政府宣传机构都开设了自己的微博公共平台，这些平台不仅可以宣传积极正面的新闻事件，给社会带来正能量，宣传传播正确的思想价值观。而对于这些微博新闻数据的舆情监测，除了可以利用微博平台进行统计分析。我们还可以利用微博的短文本数据，对微博数据进行文本主题分析、文本情感分析。了解分析新闻事件在传播过程中，新闻事件的关注重点，普通的老百姓对于新闻事件的情感态度倾向等，从而可以更好为人民群众服务。

本研究首先抓取了新浪微博中涉及“亚投行”的数据。通过对于抓取的数据进行特征选择，筛选出能很好监测“亚投行”舆情的特征值。通过对于这些特征值的描述分析，可以初步了解“亚投行”事件的影响力，群众对于这一事件的关注讨论点。除此之外，通过对于新浪微博中的内容进行讨论研究，可以找出“亚投行”“中国”这类关键词在相关的新浪微博中的上下文近义词，从而确定和主题相关的关注点。除此之外，通过新浪微博的情感分析可以有效挖掘用户的情感态度和相关的情感态度变化趋势，并且通过对于情感态度的变化发现重要的舆论信息；最后采用LSA对于微博的文本进行了主题聚类，挖掘出重要的主题关键词，筛选出重要的文本主题信息。整个数据分析过程中，需要处理大量无关的文本噪音，采用合适的中文分词方法对于整个数据进行数据预处理和数据清洗。通过文本的描述统计可以初步了解事件的大概主题中心和人群的分布、受影响情况等问题。而采用自然语言处理的方式将文本数值化，向量化可以帮助我们深入了解微博主题变化、微博情感态度变化。这些都非常有利于我们进行舆情分析。

然而微博数据获取十分艰难，微博对于用户隐私的重视越来越强。对于中文分词的歧义消除、微博数据量较小、微博平台上营销账号众多、噪声筛选难度大等问题都严重影响了本次研究设计中的结果。因而本研究至少还有以下几个方面需要改进。（1）更好的中文分词和数据预处理算法。本研究采用全模式的方式对文中的词语进行抽取，这样在语料库较少的情况下，保证了语料库中必要特征的召回率，然而，语料库中的噪声也大大增加。（2）更大的数据集。本研究选用5708条微博数据进行文本挖掘，导致在主题分析中每个主题的关键词和主题的想关性不是十分明显。如果采用更大的数据集进行分析，则进行上下文判断时产生的结果就会更理想。（3）更好的训练集。在情感分析中采用朴素贝叶斯的算法对文本进行分类。然而分类的结果高度依赖于训练数据中积极和消极词汇的选择，而且具体到某类特定的主题还需要有更多专有的训练集合才能更好的提高文本情感态度分类的准确率。

综合来看，本研究中采用较合理的算法对微博文档进行了深入细致的研究，可以发现出很多重要的规律，这对于舆情的监测、分析，对于政府科学有效的管理治理社会都有很大的用处。后续研究可以继续寻找更加合适的语料库，选择更好的分类决策算法改进实验的结果，从而获得更好的分析预测能力，提高我们对于网络热点舆情的监测、分析、掌握、预测能力。

1. 参考文献

   [1]http://zh.wikipedia.org/zh-cn/Twitter[OL]. 2014. [↑](#endnote-ref-2)
2. [2]蒋盛益, 麦智凯等. 微博信息挖掘技术研究综述[J]. 图书情报工作. 2012(17)：136-142 [↑](#endnote-ref-3)
3. [3]李晓明，邱泽奇等.网络热点的社会传染现象[OL]. http://blog.sina.com.cn/s/blog\_8f48f4530100wxm3.html. 2014. [↑](#endnote-ref-4)
4. [4]陈夏明，强思维等.危害公共安全的关联事件挖掘[J]..中国大数据技术创新大赛2014. [↑](#endnote-ref-5)
5. [5]Harshavardhan Achrekar. Predicting Flu Trends using twitter Data [J]. Computer Communications Workshops. 2011:702-707 [↑](#endnote-ref-6)
6. [6]Son Doan, Bao-Khanh Ho Vo. An Analysis of Twitter Messages in the 2011 Tohoku Earthquake[J]. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. 2011(91): 58-66 [↑](#endnote-ref-7)
7. [7]Dongjin Choi and Pankoo Kim. Sentiment Analysis and Time Series with Twitter [J]. Twitter and Society 2014. [↑](#endnote-ref-8)
8. [8]Dongjin Choi and Pankoo Kim. Sentiment Analysis for Tracking Breaking Events: A Case Study on Twitter[J]. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 2011(62):406–418 [↑](#endnote-ref-9)
9. [9] Vapnik V. et al Support-vector networks[J]. Machine Learning 1995(20): 273-297 [↑](#endnote-ref-10)
10. [10] http://www.liwc.net/[OL].2014. [↑](#endnote-ref-11)
11. [11]Shoko Kato, Akihiro Koide. Network Analysis of Three Twitter Functions: Favorite, Follow and Mention [J]. 2012: 298-312 [↑](#endnote-ref-12)
12. [12]谢丽星, 周明. 基于层次结构的多策略中文微博情感分析和特征抽取[J]. 中文信息学报. 2012(26): 75-82. [↑](#endnote-ref-13)
13. [13]https://code.google.com/p/word2vec/[OL]. 2013 .

    [14] 唐晓波, 宋承伟.基于复杂网络的舆情分析[J].情报学报. 2012(11): 1153 -1163.

    [15]Bai Xue,Chen Fu, A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec[J], Big Data 2011(014): 358 – 363

    [16]Susan T. Dumais, Latent semantic analysis[J], Journal of the American Society for Information Science. 1991:391-407

    [17]Alec Go, Twitter Sentiment Classification using Distant Supervision[J]. CS224N Project Report. http://s3.eddieoz.com/docs/sentiment\_analysis, 2014.

    [18]Songbo Tan. Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis[J], 2009:337-349

    致　谢

    本研究以及论文都是在我的导师刘晓娟老师的亲切关怀和热情指导下完成的。她严谨认真、追求卓越的品质是我论文完成过程中的不懈动力；她对于细节的一丝不苟，对于困难的永不放弃精神值得我在以后的科研学习工作中模仿。刘老师不仅在论文进度上周周敦促，对于论文中建模的方法也要求严苛。除此之外，刘老师还为我提供了良好的学习试验环境，让我的论文完成有了足够的保证。刘老师的教导我会终身铭记。

    除了刘晓娟老师的大力指导外，师兄刘新哲和师姐宋婉姿也在实验室中为我提供了极大的帮助，从处理数据到理论建模，两位学长对我的论文思路改进也提供了很多修改意见。最后，马梁同学在论文完成过程中也为我提供了很多有用的思路和方法。在我的论文完成之际，谢谢这些同学的热心帮助，让我的大学最后一项科研生活能够大有不同。

    最后，感谢我的父母。他们让我从小支持我做自己喜欢的事情，让我能够开心的完成大学最重要的论文，无论我的论文碰到什么样的难题，他们都鼓励我坚持做下去。

    黄勇

    2015年5月　 [↑](#endnote-ref-14)
14. [↑](#endnote-ref-15)
15. [↑](#endnote-ref-16)
16. [↑](#endnote-ref-17)
17. [↑](#endnote-ref-18)
18. [↑](#endnote-ref-19)