基于异质话题模型的双视图财经数据分析

09300240004 计算机科学与技术 朱恬骅

# 概述

## 本文目的

伴随中国经济的发展，金融市场呈现出越发活跃的景象，这使得人工（专家）对数据的分析相对于数据的产生速度显得太过微弱，而市场又亟需可资参考的数据分析结果以利决策。这使得基于数据挖掘的财经分析具有丰富的应用前景。然而这也对计算机的机器学习方法、模型提供了一个不同的应用场景，需要有针对性地进行讨论。这一选题正是顺应了实践上和理论上的这些要求，以期建构一个财经分析（股票聚类）的适用模型，揭示金融市场的一些规律、线索，为财经分析提供参考。

## 本文的主要方法和研究进展

在信息处理和文本挖掘领域，对文本的表示方法主要采用向量空间模型或统计语言模型。文档常用表示其词频分布情况的向量来表示，称为向量空间模型[[[1]](#footnote-1)]。每一维都相当于是一个独立的词（组）。如果这个词（组）出现在了文档中，那它在向量中的值就非零。每一个维度被称为对应词（组）的权重。除了用词频作为权重之外，常用的权重取法还有TF-IDF权重等。通常而言，一个词组就是一个单一的词、关键词，或短语。向量运算能通过查询来比较各文档。常用的距离函数有欧几里得距离、Minkowski距离、余弦距离等。

统计语言模型[[[2]](#footnote-2)] 的发展引入了对文本生成过程的建模，从而允许引入更多的先验知识。由于运用了概率论的知识，统计语言模型有着更加坚实的数学基础，不像TF-IDF那样难以解释权重值的物理意义。然而，它与向量空间模型一样，认为文本是词和文档之间的映射关系，也就是说，它们都是在词典空间上表示的。

随着人们对于文本处理的需求不简单局限于个别字词的检索，而希望引入“相关信息”，将文档单纯看作词典空间中的映射关系，就显得不再足够了。正是在这样的背景下，话题模型得以产生，为词和文档之间增加了一个新的层面。这也就是话题（topic）的层面。其中，潜在语义分析（latent semantic analysis，LSA）[[[3]](#footnote-3)] 就是一个成功的早期工作。在LSA中，文档存在于在新引入的语义空间中，而语义空间中反映了词之间的相互关系。它的本质想法是考虑词在文档中的共现情况，以此为根据抽取出词和语义之间的映射关系，然后实现文档在相对低维的语义空间中的表示。

在LSA中，这样的抽取过程是通过线性代数的矩阵分解完成的。这同样使得它的结果不易于有良好的事实解释。pLSA（probabilistic latent semantic indexing/analysis）[[[4]](#footnote-4)] 通过引入概率模型解决了这一问题。在LSA中，每个语义对应一个特征向量；而在pLSA中，每个语义对应一个词典上的概率分布，文档对于每个语义的权重相应地就是在语义空间上的概率分布。

在数学上，pLSA并不是一个充分的贝叶斯模型，因为它的文档－话题分布和主题－词组分布被看作是参数而非随机变量。

针对这一点，Blei等人[[[5]](#footnote-5)] 提出了LDA（latent Dirichlet allocation）这种层次贝叶斯模型（hierarchical Bayesian model）。其思想如下：假设整个文档集合一共有个话题（topics），每个话题表示为词典上的一元语言模型（如图 1），即一个多项式分布。每个文档对应这个话题有一个关于文档的多项式分布。LDA假定，，对于文档中的每一个词，有话题，。采用Dirichlet分布的好处在于，Dirichlet分布中的每一个采样点对应一个多项式分布。同时，。 LDA的图模型表示如图 2。



图 1 一元语言模型



图 2 LDA的图模型表示

LSA、pLSA、LDA的特性使得它们易于就不同的适用场合进行扩展。已有的一些重要的扩展结果包括：McCallum等人的作者模型（Author Model）、作者－话题模型（Author-Topic Model）、作者－接收者－话题模型（Author-Recipient-Topic Model），以及考虑了各种要素的cFTM模型[[[6]](#footnote-6)] （Contextual Focused Topic Model，其图模型参见图）。围绕学术文章的发表过程展开了一系列作者、文档的聚类。这一些拓展展现出一个非常有趣的侧面，也就是作者是以其所发表过文章作为特征的，而文章又以作者为一个特征。这种实体间关系上进行的聚类，是多视图聚类的一个重要侧面。



图 3 LDA及其几种拓展[[[7]](#footnote-7)]

从左至右：LDA[5]、Author Model (Multi-label Mixture Model)[[[8]](#footnote-8)] 、Author-Topic Model[[[9]](#footnote-9)] 、Author-Recipient-Topic Model[7]

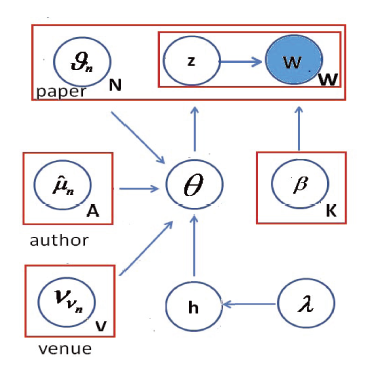


图 4 cFTM的图模型表示[6]

但本文所要研究的问题，与上述围绕实体间展开的聚类有本质的不同。其不同之处在于，我们所要处理的两类视图，是同一类实体的反映；亦即，如果以作者－文章的关系为例，我们研究的目光聚焦于对文章进行聚类，其中同时考虑到作者和文章中所包含的语词，而不考虑作者的聚类问题，其视图准确来说是作者和词语，而被聚类的实体则是文章。它较为容易进行的主要原因在于相对于文章而言作者比词语有更少的维度（只有少数的一些作者参加了文章的撰写），同时有着更高的重要性（学术论文的作者通常有着明确的学术兴趣，决定了其文章之间具有较高的关联性，而这些关联性可能牵涉到大量的同义词、同类词，从而在词语的维度上不能明确看出）。

在本文所试图解决的问题中，两个视图分别是文本和股价信息。它们首先具有不同的表征方式，文本是离散的，而股价则是连续量；文本在词袋（bag-of-words）模型中是不考虑前后次序的，而股价则有着明确的时间印记。此外，文本和股价之间不具有直接的相关性；二者都是取决于市场的反应，它们都是一种“后来的描述”，不像作者可以决定文章的内容，从而作为一个重要的标签。

在双视图的研究方面，Jordan等人[[[10]](#footnote-10)] 提出了用话题模型解决语言翻译的问题。该文讨论了利用英德文的非平行语料和一定的标签（例如维基百科中对于同一个主题的论述）发现两种语言中表达的对应关系的尝试。就某种意义上，我们的问题与之有一定的相似性：对于同一只股票，有文本（自然）语言对它进行的描述，也有股价（通过某种方式转化为词语）语言进行的描述，而我们的目的在于发现文本语言同股价语言之间的某种对应关系。然而不同于机器翻译领域中一对一的明确性，我们的映射关系是相当模糊的。

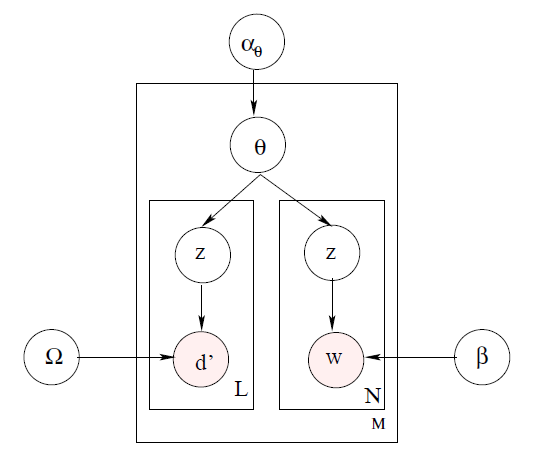


图 5 Link-PLSA-LDA模型的图模型表示

Nallapati等人[[[11]](#footnote-11)] 提出了Link­-PLSA-LDA这一模型，其图模型表示如图 5所示。同样是为了利用两个视图中的文本信息进行聚类。他们对于不同特性的文本采用不同的方法进行分析，最后又统一到同一个框架中进行聚类，这对我们来说提供了一个值得参考的经验，启发我们在解决股价聚类问题的时候，可以对两个视图采用不同的模型进行解释。这也对我们的框架提出了新的要求，亦即要求它能够灵活应用不同模型的估计结果。

虽然LDA有了很好的数学基础，Link­-PLSA-LDA也为我们提供了一种整合不同分布假设，分别应用于不同视图的解决方案，对于我们的应用而言仍然显得力不从心。其主要原因仍然在于股价的特殊性，决定了现有的以文本为重心的聚类解决方法对数据的解释能力总显牵强。这使得我们重新回到pLSA，回到LSA，重新审视作为矩阵分解的隐变量（话题）发现过程。如果我们将整个话题发现的过程视作在一定限制条件下（如非负、和为1等概率原则所要求的条件）进行矩阵分解的过程，那么我们有可能通过矩阵分解发现有概率意义的话题，而不须知道或假设其所符合的先验分布。早期Lee等人[[[12]](#footnote-12)] 的工作为非负矩阵分解的物理意义提供了解释，即“整体作为各个部分一定比例的调和”，并且通过非负矩阵分解重新发现那些组成整体的“部分”，就像五官之于人脸。Berry等人[[[13]](#footnote-13)] 、Xu等[[[14]](#footnote-14)] 的工作证明了单一视图条件下，运用非负矩阵分解的方法发现话题的有效性。

Arora等人[[[15]](#footnote-15)] 重新考察了SVD的局限性，认为：传统的基于SVD的矩阵分解，要么需要假设每个文档只包含一个话题，要么只能恢复话题向量的共轭（span）而不是话题向量本身。该文研究了一种基于非负矩阵分解（Nonnegative Matrix Factor）的方法，类比于SVD，从而得到了更加泛化的一种有关于话题间相关性的模型，例如相关话题模型（Correlated Topic Model，CTM）和PAM（Pachinko Allocation Model）。

作为一个类比，在我们期望提出的模型中，话题间的相关性是当给出的两个视图完全相同时的一个特例。这也就是说，在我们的模型中涉及矩阵的三元分解（tri-factorization）：两个矩阵分别代表两个视图的特性，剩下的那个矩阵表示两个视图之间的相关性。关于矩阵的三元分解，先前的文献也已有所研究。如Wang等[[[16]](#footnote-16)] 提出了一种处理对称阵三元分解的聚类算法，可以处理异质的数据。不过，该方法关于对称阵的限制拘束了我们的应用范围，使得我们可能仍将面对Arora等人提出的两难境地。

在关于股票聚类的研究方面，Doyle等人[[[17]](#footnote-17)] 较早地利用话题模型对股价进行聚类，该文提出了采用话题模型对股价进行分析的方法，并就一个较小规模的数据给出了实验和解释。为解决股价的表示问题，使之能够适合于LDA的框架，该文作者提出采用正比于涨跌幅度百分数的方式生成“某股涨/某股跌”的文档，每篇文档对应于一个交易日。我们认为，这样的表示方法有其合理性，但是由于话题模型并不考虑文档之间的先后次序，所以这种表示方法丧失了股价信息中非常重要的时序信息。这样的表示法有助于发现就概率而言容易同涨同跌的股票，但不便于对二者的走势之间是否相似进行描述。为避免这种表示法带来的股价的失序性，我们采用“某天涨/某天跌”的表示法。关于股价表示方法的具体讨论，请参见第2.1节。

文本方面，R. Schumaker等人[[[18]](#footnote-18)] 提出了用财经新闻作为特征，来进行股票聚类的方法，以对市场进行预测。该文提出的AZFinText系统和现有的量化基金及股票专家给出的预测结果相比有一定的优势，并发现按照经营范围（sector）进行分类预测有着更好的结果。该系统有更高的投资回报率。

## 本文的结构安排

本文的结构如下。第一章，介绍选题的背景和当前的研究状况。第二章描述考察的各数据集，并给出预处理的具体方法。其中，对于中文文本的词典抽取作了介绍，并检验了其效果。第三章，简介话题模型，并引入本文所采用的基于非负矩阵分解的异质话题模型。第四章，将第三章中提出的模型在一些标准的数据集上进行测试，并应用于第二章给出的数据集上，给出分析的结果及解释。最后，在第五章中总结该异质话题模型的优缺点，并对其未来的应用和发展提出展望。

# 数据的采集与预处理

## 价格信息的采集与预处理

### 价格信息的获得

股市的价格信息在多家网站都可查询得到，炒股软件中也可以方便地下载到历史数据。对于中国A股市场，本文采用的是通达信软件，将沪深股市A股的全部历史数据（自1991年12月19日至2012年12月31日）下载后导出为文本文件，每个文件的格式如下：

代码 1 数据格式

600000 浦发银行 日线

日期 开盘 最高 最低 收盘 成交量 成交额

11/10/1999 4.03 4.08 3.61 3.73 174085000 4859102208.000

11/11/1999 3.71 3.84 3.70 3.73 29403400 821582208.000

11/12/1999 3.75 3.83 3.74 3.78 15007900 421591616.000

11/15/1999 3.81 3.82 3.73 3.73 11921000 332952800.000

……

为方便预处理，使用C#编写了选取指定日期、股票的图形界面程序，其运行时截图如下所示。

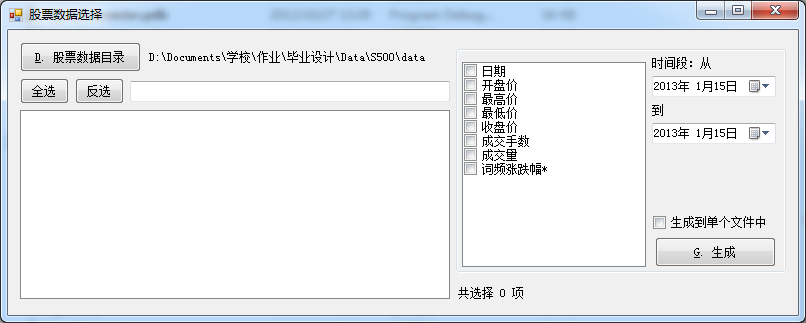


图 6 股票数据选择图形界面

对于国外股市，我们采用的是雅虎财经（finance.yahoo.com）上的数据。我们所选取的证券市场是：美国纽约证交所（NYSE）、英国伦敦证交所（LSE）、新加坡证交所（SGX）、澳大利亚证交所（ASX）。其中，LSE、SGX、ASX各上市公司经营范围描述（BSD）文本和行业分类数据来源于各交易所网站。

### 价格信息的预处理

由于模型的性质，我们将只能处理以“词频”（或“频率”）的形式出现的数据，而股票的价格是连续变化的、范围不定。考虑到股票价格的绝对数值大小对我们分析股市情况的意义不大，我们采用“涨跌幅”的形式进行处理，亦即只考虑一个单位时间的股价相对于前一单位时间的股价涨跌幅度的百分比。因此，设若有天的股价信息，我们将这T天的股价信息扩展为一个维的向量，其中第维数值大小，当为偶数时表示的是第天相对于前一天的跌幅，而对奇数，则表示第天相对于前一天的涨幅。

对于股价涨跌幅的另一表示方式是 “正-负-零收益表示法”。其方法是，将每一支股票的走势看作一篇文档。设每支股票取天的价格信息，建立一个大小为的词汇表，包括了“第天涨”、“第天持平”和“第天跌”，。将股票的走势表现为这个词上的词频。在数据集上的K-means聚类验证表明（参见下表），上述这两种表示方式的聚类结果并无太大差异。由于正-负-零收益表示法的维数太大而且不直观，在下面的讨论中我们采用词频表示法。例如：

采用词频表示法，K-means聚类的结果如下：

1(4) 600000浦发银行 600015华夏银行 600016民生银行 600036招商银行

2(4) 600019宝钢股份 000959首钢股份 000709河北钢铁 600569安阳钢铁

3(4) 600085同仁堂 600129太极集团 600422昆明制药 000423东阿阿胶

而用正-负-零收益表示法：

1(4) 600085同仁堂 600129太极集团 600422昆明制药 000423东阿阿胶

2(4) 600000浦发银行 600015华夏银行 600016民生银行 600036招商银行

3(4) 600019宝钢股份 000959首钢股份 000709河北钢铁600569安阳钢铁

为节省篇幅，在更多股票上的分类结果此处予以省略。

### 中国A股市场不同尺度数据集的选取

对于中国A股市场，参照财经网站的专家分类和“龙头股”信息，选取了含有50支不相关股票、50支相关股票和全部股票的数据集。数据集中具体股票的列表如下。

#### 小规模数据集：Stock50、RStock50

#### 不相关板块的50支股票：Stock50

选择了来自银行、钢铁、医药、酒、软件这五个板块的50支股票，构成不交叉板块50支股票数据集（S50）。

银行：

000001 深发展A

002142 宁波银行

600000 浦发银行

600015 华夏银行

600016 民生银行

600036 招商银行

601009 南京银行

601166 兴业银行

601169 北京银行

601288 农业银行

钢铁：

000629 攀钢钒钛

000709 河北钢铁

000717 韶钢松山

000898 鞍钢股份

000932 华菱钢铁

600282 南钢股份

600019 宝钢股份

000959 首钢股份

000022 济南钢铁

600569 安阳钢铁

医药：

601607 上海医药

600511 国药股份

600833 第一医药

600713 南京医药

000028 国药一致

600085 同仁堂

600129 太极集团

600422 昆明制药

000423 东阿阿胶

002589 瑞康医药

酒：

000568 泸州老窖

000858 五粮液

600519 贵州茅台

600779 水井坊

000596 古井贡酒

600809 山西汾酒

000799 酒鬼酒

002304 洋河股份

600702 沱牌舍得

600559 老白干酒

软件：

600570 恒生电子

600756 浪潮软件

000948 南天信息

600271 航天信息

002063 远光软件

002065 东华软件

000938 紫光股份

002073 软控股份

002090 金智科技

002230 科大讯飞

#### 相关板块的50支股票：RStock50

选择了来自钢铁、煤炭、汽车、航空、电力这5个板块的各10支股票。它们都存在一定的相关性，如钢铁和电力都依赖煤炭，汽车依赖钢铁，航空则与煤炭、电力所代表的能源产业有密切关联。构成交叉板块50支股票数据集（RStock50）。

钢铁：

000629 攀钢钒钛

000709 河北钢铁

000717 韶钢松山

000898 鞍钢股份

000932 华菱钢铁

600282 南钢股份

600019 宝钢股份

000959 首钢股份

000022 济南钢铁

600569 安阳钢铁

煤炭：

000780 平庄能源

000723 美锦能源

002128 露天煤业

600188 兖州煤业

600348 阳泉煤业

600546 山煤国际

600740 山西焦化

601001 大同煤业

601666 平煤股份

601898 中煤能源

汽车：

000550 江铃汽车

000572 海马汽车

000625 长安汽车

000800 一汽轿车

000868 安凯客车

000927 一汽夏利

000951 中国重汽

000957 中通客车

002594 比亚迪

600006 东风汽车

航空：

600029 南方航空

600115 东方航空

600221 海南航空

600316 洪都航空

000089 深圳机场

600004 白云机场

600009 上海机场

600151 航天机电

600893 航空动力

000901 航天科技

电力：

600011 华能国际

600021 上海电力

600027 华电国际

600644 乐山电力

600101 明星电力

600116 三峡水利

600131 岷江水电

600236 桂冠电力

600292 九龙电力

600310 桂东电力

#### 大规模数据集：Stock2209

选择在2010～2012年有挂牌交易的2209支股票，几乎涵盖了中国A股市场正在交易的所有股票。具体的股票名单此处从略。

### 海外股市的股价数据选取

对于纽约证券交易所（NYSE）、伦敦证券交易所（LSE）、新加坡证券交易所（SGX）和澳大利亚证券交易所（ASX），我们首先获取全部股票的一般信息，包括其所属公司和公司描述文本（business scope descriptions，BSD），只保留上述两者信息均齐全的股票。这样，我们最终得到的有效股票数量如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 证交所 | NYSE | LSE | SGX | ASX |
| 股票数量 | 2024 | 2030 | 542 | 1368 |
| 行业分类数量 | 10 | 46 | 14 | 26 |

## 文本信息的采集

本文中涉及的文本信息共有三种：其一是上市公司的经营范围描述；其二是新浪的财经新闻；其三是东方财富网上的股吧帖子。在所有2209支股票的经营范围描述文本信息中，采用2.3中构建关键词词典方法，查找出的前100个高频词如下：

经营

技术

销售

生产

业务

出口

设备

服务

开发

进出口

材料

许可

企业

项目

机械

工程

国家

商品

加工

咨询

除外

制造

本企业

投资

配件

电子

禁止

管理

公司

不含

进出口业

进出口业务

相关

许可证

化工

设计

代理

机械设备

系统

仪器

建筑

营本企业

计算机

经营本企业

金属

规定

信息

汽车

法律

仪表

法规

范围

货物

安装

公司经营

原辅

贸易

零配件

辅材料

所需的

然而词典过大，不利于我们发现其中的关键信息，增加了聚类的难度。这里我们考虑引入一种关键词挖掘的方法。Liu[[[19]](#footnote-19)] 等基于翻译的思想给出了一种关键词抽取的方法。文章关键词抽取中一个基本的挑战是文档与关键词之间的词汇差距（vocabulary gap）。这篇论文提出的一个关于文档及其关键词的观点是：每篇文档和它的关键词是对同一目标的不同描述，只是文档以一种语言描写，而关键词以另外一种语言写成的（这里的语言应该被理解成一个抽象的符号及其规则）。所谓的关键词提取就成为了一个翻译的问题。这也暗示了我们的双视图聚类技术或许可以用于处理一些机器翻译方面的问题。在技术上，Liu等使用的是统计机器翻译（statistical machine translation，SMT）词对齐模型（word alignment models，WAM）。该文使用了标题和摘要，或者几句重要句子，以建立与文档之间的翻译对。形式上说，就是对于文档集合中的一篇文档，需要按对候选关键词进行排序，选择概率最高的个候选关键词作为结果。整个过程被分为三部：1) 准备翻译对（用标题和摘要，构成翻译对<D, T>和<D, S>）；2) 用WAM训练翻译模型，计算，其中是标题中的单词，而是文档中的单词；3) 关键词提取，其中 a. 计算文档d中每一个单词的重要程度分数；b. 计算候选关键词的排位分数；c. 选出最高的个候选关键词，作为的关键词。训练WAM所需要的时间太长，而在“关键词生成”工作中，也就是从文档中提取含义相似而未出现在文档中的词汇作为关键词（就像LDA中的每一个话题那样，不一定是在某个文档中共现的），WAM表现地要比TFIDF、LDA等方法都要好。然而其缺点是，WAM需要大量的训练数据和训练时间才能得到较好的结果，这对我们目前的人力物力构成了挑战。

因而，本文中最后使用的是贝叶斯平均的方法，估计全部经营范围描述中的总体词频，再提取那些在某支股票的描述中出现频率明显高于平均状态的词，将它们标记为关键词。具体方法如下。首先计算出每个词在所有股票的经营范围描述文本中的平均总频数，再计算出它们的平均得分。容易看出，样本越大的词，就越有能力把最终得分拉向自己本来的得分，样本太小的词，最终得分将会与全局平均分非常接近。这种与全局平均取加权平均的思想就是贝叶斯平均，也是一种较为常见的平滑处理方法。

得到的部分结果如下：

000001 监管 人民币 汇款 借款 放款 非贸易 有价证券 汇兑 信托业 外币 见证 资信 承兑 各项 存款 贴现 调查 票据 结算 外汇 代理业 人民 保险 境内 买卖 允许 发行 有关 境外 办理

002142 十一 十三 十二 金融债 中国银 公众 银行卡 信用证 发放 中期 款项 长期 收付 兑付 短期 吸收 债券 监督 保险业 政府 承兑 银行 拆借 存款 担保 中国 贴现 保管 同业 票据

600000 外汇 托管 保险箱 全国 离岸 保障 外币 借款 汇款 兑换 委员会 社会 银行业 见证 资信 中国银 拆借 结汇 股票 存款 担保 公众 贴现 同业 信用证 发放 中期 款项 长期 收付

600015 金融债 委员会 中国银 结汇 公众 银行卡 债券 信用证 发放 款项 中期 长期 收付 兑付 短期 政府 吸收 监督 承兑 拆借 存款 担保 贴现 保管 同业 买卖 票据 结算 贷款 代理业

600016 本行 十四 十一 十三 十二 可以 银行业 结汇 金融债 公众 银行卡 信用证 发放 中期 款项 长期 收付 兑付 短期 吸收 监督 保险业 承兑 银行 拆借 债券 存款 担保 中国 政府

……

000028 医用 区域性 救护车 口腔科 化验 缝合 一次性 灭菌 诊断 同化 第一 器具 激素 手术室 急救室 精神 抗生素 射线 麻醉药 超声 敷料 附属 临床 诊疗 蛋白 毒性 疫苗 分析 消毒 合剂

000423 膏剂 合剂 糖浆 口服液 保健 颗粒剂 胶囊 药品 批准 食品 范围 许可证 进出口业 商品 生产 销售

002589 保存 常温 毒液 罂粟 助听器 隐形眼镜 同化 激素 体外 麻醉药 蛋白 毒性 疫苗 健身器 护理 诊断 抗生素 精神 配送 三类 化学药 饮片 日用品 生化 试剂 中药材 制毒 生物制品 中成药 化妆品

600085 营养液 老年病 乌鸡 作用 妇产科 儿科 梅花鹿 乌骨鸡 外科 冷食品 中医科 内科 马鹿 涂膜剂 同仁 皮肤科 供暖 定型 皮肤 北京 诊疗 其中 股份 动植物 西药 饲养 有限公司 图书 保健 饮片

600129 执业 中草药 旅馆 水产 西药 作业 二级 首饰 前不 副食品 保健 金银 土地 中成药 养殖 以下 工艺美术 维护 种植 经济 印刷 不得 百货 医疗 旅游 器械 出租 自有 化学 包装

## 财经专业词汇的提取

由于中文的书写习惯不同于英语等西方语言，在英语语言学者中非常简单而普遍的词频统计方法应用于中文时需要首先完成对文本的分词。对中文文本进行分词成为进一步文本处理的先决条件。中文文本的分词准确程度将很大程度上影响后续分析的效果。

近年来，随着汉语学者在语言学和计算机科学方面的不断努力，各种新的分词算法正在向更高的分词准确率发起挑战。然而，由于计算机系统自身的局限，尽管在未登录词的识别方面也已得到可喜的成绩，各类分词算法都首先需要一个完备的词典和词频统计数据，以便得到较为准确的结果。

本节实现了一个基于信息熵的中文词汇识别、抽取系统，并给出了此系统在识别财经新闻中专业词汇（词组）的一个应用，以演示该方法在处理以专业词汇为主的文本时所具备的优势。

在本节的实验部分，首先给出了这一词汇抽取系统在《人民日报》1998年1月分词标注语料库上的表现，与人工分词的结果互为比较。实验表明，该方法在《人民日报》语料库上的词汇覆盖率为64%，结合常用词表后覆盖率可以超过96%；在所有找到的词汇中，正确率为 68%～69% ，这是因为系统缺少关于词组的知识，在词汇表中引入了较多的高频搭配所致。

在本文的应用部分，给出了这一系统在新浪网财经频道个股新闻板块84.8万篇新闻中抽取出的词汇词频表，并介绍了利用这一词频表，结合分词算法进行词典迭代精化的方法及其结果。

### 理论与方法

早期的无监督词典建造大多采用信息熵的方法。例如，Yamamoto[[[20]](#footnote-20)] 提出了利用信息熵为无词语边界的语言（该文以日语为例）进行分词的思想。Sun等[[[21]](#footnote-21)] 就利用信息熵对中文分词进行了实验。Feng[[[22]](#footnote-22)] 给出了一个更加清晰的用于中文文本词汇抽取的方法，其主要关注点在于字串的前驱和后继字符，并且主要展现了其在未登录词上的能力。

Shannon在1949年提出了信息熵的概念。[[[23]](#footnote-23)] 他将信息熵定义为对不确定性的测量。熵的概念最早起源于物理学，用于度量一个热力学系统的无序程度。但在信息论中，熵越高，代表信道传输的信息量越大；熵越低，则意味着传输的信息越少。

任何能够产生符号的发送者都可以被认为是一个信道。信道的基本组成有其字母表，字母表上的概率分布。符号的接受者则被称为信宿。信道是信号（符号）传送的渠道。 一个常用的信源模型是离散无记忆信源（discrete memoryless source，DMS）。DMS的定义如下所述。考虑一个信源，在每单位时间中，从一个有限的集合（信源字母表）中独立地产生一个符号。由于集合有限且取值离散，我们称信源是离散的。在时间中，产生的符号可用一个序列表示：。若这一过程中，事件发生的概率与时间无关，也与前一个符号的取值无关。我们称这个信源是无记忆（memoryless）的。满足上述两个特征的信源即为DMS。相反地，如果符号的取值与有关，则该信道就是有记忆的。

为方便建模，这里再简单引入一下（一阶）马尔可夫性质（Markov property）的概念。马尔可夫性是指，一系列的随机变量满足下列等式：



亦即，时刻信源所产生的符号仅与其前一个符号有关。类似地，我们可以定义2至阶马尔可夫性质：



具有阶马尔可夫性质的随机变量系列称为阶马尔可夫链（Markov chain）。显然地，我们可以通过恰当定义概率分布，来使得高阶的马尔可夫链完成低阶的马尔可夫链的行为。尽管在实际中我们并不会这样做，在数学上，我们完全可以认为阶马尔可夫链真包含阶马尔可夫链。

### 文本生成过程的信源建模

在本文中，假设文本信息是由字母表为的信源所产生，其中代表一个有限的现代汉语词汇的子集，是一个DMS。在不引起混淆的情况下，我们用来表示上述式子。我们假设信道是无错的。本文所需要完成的工作即是通过观测产生的符号序列来猜测的字母表和概率分布。

将文本信息建模为DMS并非完全出于简化问题的需要，而是出于下面的考虑：其一，我们主要解决的是词频统计中的问题，而词频统计中并不关注词汇和词汇之间的顺序关系，采用所谓的“词袋”（bag-of-words）模型。其二，关于词汇的前后位置关系，即语法关系，应当由分词算法处理。而本文只关注如何对未分词的文本进行必要的词频统计处理，以利于分词算法发挥作用。如果在此越俎代庖，则会构成逻辑上依赖的循环。

然而在猜测的过程中，由于词汇的集合是待定的，信宿并没有获得完整的。为了能够记录信源发来的信息，信宿采用的字母表是全体汉字的集合。对于发送来的词汇，我们只能观测到它们的字符形式。借助于标点符号，我们将语料表示为一个句子的集合，其中每个句子都是一个字串，其中是一个汉字。我们称之为在信宿视角下的信源，记为。由于我们将抽象的层次从词汇下降到了字符，显然不再是一个DMS（例如，对于字串“力”，*p*(“力”|“巧克”)的取值一定大于*p*(“力”|“上海”)）。我们假设它具有阶马尔可夫性，故字符集上的概率分布可以用来表示。

设  是一个随机变量，其可能的取值范围为 ，，它们对应的概率分布为。的信息熵定义为：



其中称为的信息量，其定义为：



被求和项称为单个符号的信息量，用  表示。对数函数的基决定了信息量的单位，一般取为2，这时信息量的单位称比特（bit）。信源的信息熵就是信源概率函数的信息熵。

假设一个0-1 DMS，其信息熵函数的图像可见于图 7。可见，两个符号的信息量函数是一个凸函数（concave），过高和过低的出现概率都会导致较低的函数值。这一特性可以被用于刻画信源发送符号时的丰富程度。

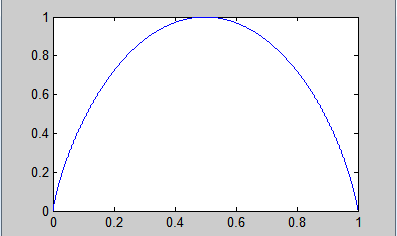


图 7 0-1DMS的信息熵函数 H(x)，横轴表示 Pr(X=1)

### 对词汇的定义

词汇是“语言中能独立运用的最小语法单位”。[[[24]](#footnote-24)] 它有两种主要特性：其一是“独立运用”，这是相对于语素这一概念而提出的。其二是“最小语法单位”，它决定了词汇的抽象层次是处在语法层面上的。由于本文不涉及对词汇的语义分析，这决定了词汇是目前所采取的抽象级别上所能处理的问题。

词汇之能够独立运用，主要体现在下述两种特性：其一，是外部的“自由程度”。这也同时体现了“最小单位”的两个方面。从“大到小”的方向上来看，“谢谢”是一个词而“谢谢你”不是一个词，因为前者参与句子构成时候可能的前后组合要比后者丰富。而从“很小到小”的方向上看，“中国银行”是一个专有名词，而“中国银”却不是词汇，尽管在一篇论述“中国银监会”和“中国银行”的文章中，“中国银”出现的次数比这两个专名的出现次数都要多，但它却不是自由的，因为它在给定的词汇表中只能作为前导的三个字出现。而“中国银行”则可作为名词参与句子，出现在不同的字词之后或之前。

其二是内部的“凝固性”。用顾森提出的例子[[[25]](#footnote-25)] ，“电影院”是一个词而“的电影”不是一个词，是因为“电影院”的出现频率大于“电影”、“院”独立出现的频率之积，亦即假若信源在产生“电影院”这个字串的时候仅仅是出于随机而将它们拼凑在一起，这种情况的概率应当远小于“电影院”是由信源采用一条3阶马尔可夫链生成的情况。

为了量化词语能够独立运用、自由参加句子组成的程度，我们引入上述信息熵的概念。令一个词对应两个信源，一个是前导字串信源，另一个是后继字串信源。这两个信源的字母表皆为全体汉字的集合，显然是有限而离散的。在前文定义的基础上，我们定义信源的概率分布为，信源的概率分布为。显然，这两个信源都是DMS。由此，我们计算这两个信源的信息熵、。参考图 7，只有当的非零项平均分布时，其信息熵才能取得最大值。任何偏离平均分布的情况下都会导致信息熵函数值减小。因而，我们可以用信息熵函数的这种性质定义字串的前导/后继自由度。

考虑到词根（定位语素）的存在，一个字串是词，只有当它前导和后继的自由度都很大的时候才能被认为是词。我们据此定义字串的自由度为：



为了用统计的方法能够衡量词汇内在的凝固性，我们考虑2.3.1中定义的信源。假设字串是一个词汇，显然有



其中是对数列的一个划分，函数表示一个字串在观测到的符号串中出现的频率。因此，我们定义字串的凝固度为：



我们引入三个参数，和。其中是信源所满足的马尔可夫性质的最大阶数，亦即词的最大长度。且时，字串才能够被认为可能是一个词。

### 参数选择：在《人民日报1998年1月标注语料库》上的实验

为量化地评价方法的有效性，我们以北京大学《人民日报1998年1月标注语料库》[[[26]](#footnote-26)] 的分词结果为基准，测试本文的方法，并选择合适的参数。本文方法在先前给定的参数下，检测出候选字串6577个，人工标注的结果为7015个。其中：相同词汇有4457个，占本文方法检出词汇的（覆盖率）67.7%；未检出词汇2558个，占词汇总个数的36.5%，正确率为63.5%。

除去相当数量的错误结果外，有必要说明人工标注规则和自动抽取规则中的显著不同。在人工标注规则中，姓名被分割为两个词，由多个通用或专有名词组成的专有名词被分割为多个名词，并在最外部加括号以示边界。如此，“江泽民”在自动抽取时被认为是一个词，而在手工标注时则不是；“中国共产党”在自动抽取时被认定为一个词，而在手工标注中则为“[中国/共产党]”。同时，由于《人民日报》语体的特殊性，在结果中会出现大量满足长度限制的固定搭配，这些固定搭配中的词并未能够在给定的语料中体现出其独立性，而手工标注由于了解这方面的先验知识，所以能够作出正确的判断。例子有：“由公安机关”、“本报评论员”、“五个一工程”、“级偏北风”等。

一方面，我们需要通过归并词组中的词语来减少词典的大小；另外一方面，寻找到的词组也可能会有益于我们减少数据的维度。例如，将人名视作单个词语，就有助于我们更好地识别语篇句的主题；将套话、固定格式的短语视作词汇，有助于我们方便地去除噪音，对文本进行降维以利后续的处理。

然而，为便于将本文方法得到的词组与语料库中给定的词进行比较，考虑了如下的这些方法。首先，对于长度超过3的字串，程序进行迭代切分，每次迭代中检查词典中现存的字串，如果分拆成两个部分时二者皆已存在于词典中，或一者存在于词典而另一者长度为1，则进行切分并删除原先的词组。进行上述操作后的结果如下。

表 1 进行词组二次切分后的结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 迭代次数 | 0 | 1 | 2 | 3 | 4 | 5 |
| 本文方法所得字串数 | 6577 | 6464 | 6131 | 6065 | 6042 | 6042 |
| 标准（人工）词汇数 | 7015 | 7015 | 7015 | 7015 | 7015 | 7015 |
| 相同词汇数量 | 4457 | 4317 | 4219 | 4171 | 4163 | 4163 |
| 召回率 | 0.6353 | 0.6154 | 0.6014 | 0.5946 | 0.5934 | 0.5934 |
| 正确率 | 0.6776 | 0.6678 | 0.6881 | 0.6877 | 0.6890 | 0.6890 |
| F值 | 0.6557 | 0.6405 | 0.6418 | 0.6377 | 0.6376 | 0.6376 |

可见，分词的结果并没有得到提高。这说明，本方法必须引入少量的先验知识（即常用词的项目）参与词频统计。通过引入1000个高频词（据《现代汉语常用词表（草案）》[[[27]](#footnote-27)] ），正确率提高到95.8%，召回率为66.8%。但是这种方法并不能从根本上改善算法的运行效果。

上述参数选择中和有相当的随意性。为选择一个较好的参数，进行了更加具体的实验。

首先我们保持不变，更改的值，召回率和正确率随的变化如图 8所示。

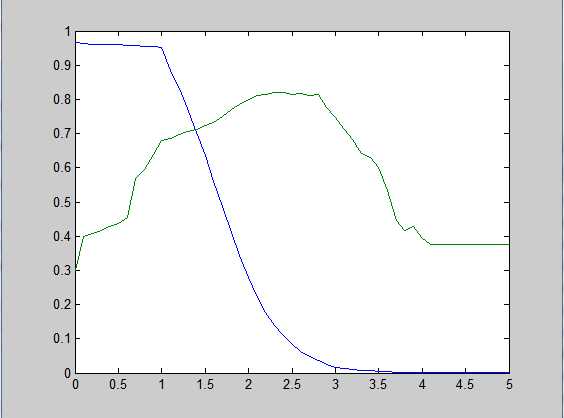


图 8 召回率和正确率随的变化（蓝色表示召回率，绿色表示正确率）

在保持不变的情况下，更改的大小。图 9展示了二者随变化的曲线。

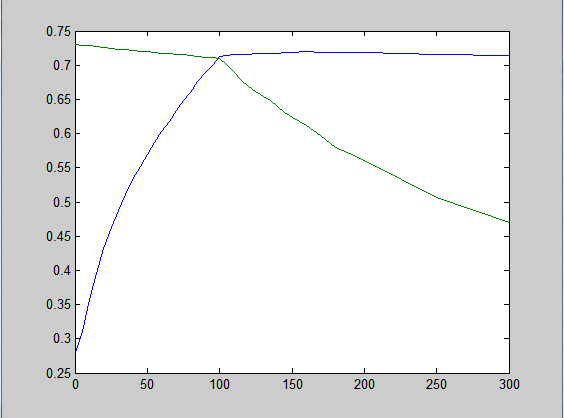


图 9 召回率和正确率随的变化（蓝色表示召回率，绿色表示正确率）

最终选定的参数为，。

### 在新浪财经新闻上得到的结果

在上述参数下，在新浪财经新闻的文本数据上运行得到的结果如下（部分）：

表 2精简前新浪财经新闻的部分词条词频

|  |  |
| --- | --- |
| 公司 | 2096998 |
| 中国 | 725431 |
| 有限公司 | 649869 |
| 股份 | 537047 |
| 亿元 | 499923 |
| 企业 | 476876 |
| 发展 | 466241 |
| 我们 | 465056 |
| 经济 | 425110 |
| 工作 | 413367 |
| 万元 | 406624 |
| 市场 | 402314 |
| 投资 | 393525 |
| 国家 | 387654 |
| 进行 | 382630 |
| 问题 | 355703 |
| 一个 | 326493 |
| 新华社 | 320408 |
| 基金 | 316954 |
| 通过 | 308460 |
| 目前 | 306072 |
| 记者 | 290986 |
| 生产 | 289248 |
| 情况 | 288023 |
| 会议 | 286005 |
| 其中 | 285450 |
| 表示 | 284006 |
| 他们 | 280058 |
| 人民 | 278677 |
| 国际 | 263114 |
| 产品 | 257241 |
| 项目 | 256868 |
| 管理 | 256622 |
| 北京 | 255960 |
| 技术 | 254165 |
| 建设 | 252179 |
| 增长 | 238363 |
| 全国 | 236672 |
| 证券 | 230777 |
| 同时 | 230009 |
| 今年 | 228717 |
| 以及 | 227257 |
| 主要 | 227243 |
| 没有 | 224222 |
| 影响 | 223614 |
| 关于 | 219818 |
| 股东大会 | 219190 |
| 美国 | 216681 |
| 认为 | 216195 |
| 我国 | 214025 |
| 其他 | 212688 |
| 有限 | 208277 |
| 根据 | 207772 |
| 政府 | 204730 |
| 股票 | 202432 |
| 一些 | 197893 |
| 这个 | 193402 |
| 增加 | 191706 |
| 方面 | 189609 |

由于文本数量庞大，产生的词条数量也会很大，不利于后续的处理。许多词语出现的频率过于频繁，如“公司”、“股市”等，对反映文档内容特征的价值不大；还有许多词语出现的频率过低，也没有太大价值。为了选取恰当频率的文档，此处采用了基于贝叶斯平均（Bayesian average）的方法，寻找每天新闻中和那天以前的那些新闻相比，出现频率显著增加的那些词。这样改过之后，一些常用词（比如“公司”）和不是词的短语（比如“中国证”），只可能在一开始的某些天当中作为关键字出现，而随着时间推移它们就不会再成为关键词了。同时，那些专有名词却可以一直保持关键词的地位，因为它们出现频率是很低的。

最后，把所有关键词整理出来作为词典，词典的大小就减小了很多，从177万条多减少到一万多。进一步去掉那些出现次数不超过十次的关键词后，词典的大小可以控制在6400左右。

表 3 新浪财经新闻的部分词条词频

|  |  |
| --- | --- |
| 词条 | 词频 |
| 金额单位 | 3580 |
| 交易单元 | 2581 |
| 债券持有人 | 2006 |
| 华安 | 1963 |
| 业绩比较 | 1956 |
| 家族 | 1905 |
| 占基金资产 | 1823 |
| 净值比例 | 1739 |
| 中国水电 | 1707 |
| 人民币元 | 1666 |
| 汇丰晋信 | 1614 |
| 报告期末按 | 1606 |
| 所属行业 | 1396 |
| 年前三季度 | 1383 |
| 本报告期内 | 1346 |
| 公平交易 | 1259 |
| 万昌科技 | 1253 |
| 股东地位 | 1237 |
| 债券正回购 | 1203 |
| 校车 | 1177 |
| 否如否 | 1176 |
| 公司总部 | 1166 |
| 泰信 | 1163 |
| 年龄 | 1154 |
| 高庆昌 | 1139 |
| 净值增长率 | 1137 |
| 柯达 | 1122 |
| 小排 | 1112 |
| 企业年金 | 1052 |
| 赛马实业 | 1047 |
| 一二年七月 | 1032 |
| 请详 | 1023 |
| 易方达 | 1018 |
| 一二年六月 | 1009 |
| 的前十名 | 907 |
| 反向交易 | 895 |

# 双视图聚类模型的构建

## 基于矩阵分解的话题模型

对数据的聚类问题归根结底是根据输入特征为数据进行自动标注的问题。在这之中，当输入特征的维度特别高时，容易引发维数灾难问题，从而对聚类准确度带来不良影响。为解决“维数灾难”，著名的方法如主元分析（principal component analysis, PCA）通过引入奇异值分解的方法，将输入数据在线性空间中进行一定的变换然后再作投影，从而能够将高维数据投影到低维中。遵循这一思路，后人又提出了隐含语义索引（latent semantic indexing，LSI）、引入了概率的pLSA等模型和方法，将变换投影后的低维空间作为话题空间，从而给予矩阵分解以新的意义。顺着这一思路，本文采用了双视图下矩阵分解分析的思想，以解决利用双视图信息进行聚类这一问题。

话题模型通过将文档视作若干话题的混合以应对聚类问题，已得到广泛的应用。然而，目前大多数的话题模型都是基于概率隐话题分析建立的。它们共同的特点是，为一种对象赋予某一相似度或不相似度的量度。最近几年，考察对象不同种类特征的共现情况的文献开始出现。但是，话题模型中利用到的贝叶斯公式要求知道特定的先验分布，例如LDA中使用的Dirichlet分布。另外，在隐变量（通常是在高维空间中）进行积分，也增加了计算的复杂程度。尽管像Gibbs取样[[[28]](#footnote-28)] 这样的抽样方法能够用来应对这一问题，近似的贝叶斯推断并不总是能够得到正确的分布。为了解决这些问题，我们提出了异质话题模型，以对不同类型特征所描述的对象进行聚类。每一种类型的特征被称作一个“视图”。每个视图中，“话题”是由“词”按一定比例的混合产生的。观察到的词频数据中，两个视图所对应的两组话题

异质话题模型可以通过矩阵分解的方法进行求解。数据矩阵之间的互相关联能够通过三个矩阵**P**、**A**、**Q**表示。其中**P**、**Q**矩阵分别表示每个视图下词与话题之间的关系，而**A**表示了两组话题的相互关系。由于这三个矩阵需要满足概率分布的特征，它们首先是非负的，因此分解的过程相似于非负矩阵的三元分解（Non-negative Matrix Tri-Factorization，NMTF）[[[29]](#footnote-29)] 。但与Ding等人的工作所不同的是，每个视图中，我们并不要求话题－词频的分布（或者“特征向量”）是相互正交的，从而能够保留同一视图内话题之间的相关性。同时，**A**矩阵包含了不同视图中不同话题之间的共现概率。从而，我们期望可以从两组话题之间的相关性，利用异质数据，得到更好的聚类结果。进行矩阵分解的动机总结如下：

1. 采用非负矩阵三元分解的方法处理不同来源（异质）数据的话题模型；
2. 不必实现知道数据所符合的分布，同时也可以加入这方面的先验知识；
3. 该方法能够捕捉到不同视图中话题间的相关性。

考虑两个相互有关联的高维随机向量、，在双视图异质矩阵分解中，我们的目标是寻找到矩阵、及，使得：



这一问题便转化为了对下列目标函数的优化问题：



其中是误差函数，而是正规化函数。给定如是的矩阵**P**、**Q**，我们能够将从*p*维缩减到的*s*维。类似地，也就将从*q*维缩减到的*t*维。

假定数据集中有个对象，在第一个视图中这个对象的特征表示为，，在第二个视图中表示为，。同传统的话题模型一样，我们引入两个隐随机变量。这两个变量都是不可直接观察到的，用于表示隐藏的话题层面，其中、，分别为文档在两个视图中各话题的权重。相应地，在第一个视图中，一个输入向量是个话题的混合。设若表示第一个视图中话题的概率，则有：



相似地，是个话题的混合，令表示第二个视图中话题的概率，则有：



结合我们的目标，有如下推导：



令为一矩阵，其第列；为一矩阵，其第列；，其中元素。这样，我们就能改写式为：



如果说、是我们能够观察到的向量，它们是对同一对象在不同视图下的描述，那么它们之间显然是有相关性的。而上述矩阵分解的过程实际是将这二者之间的相关性上推至产生它们的话题之间的相关性，也就是之间的相关性，如图 10所示。

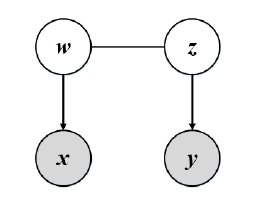


图 10 相关性上推

## 三元分解求解的数值算法

对二元的非负矩阵分解，如1.2节中所述，已有相关文献进行了研究。对于三元非负矩阵的分解，我们首先需要对上一节中隐含的关于概率的假设予以明确。对于输入数据的限制如下：



亦即、是“词频”或其他频率的形式，其每一个维度表示一个“词”。而对于三个矩阵的限制如下：



考虑我们的目标函数，对于，我们使用最为常用的最小二乘法。首先不考虑，我们的目标函数如下：



其中是根据输入的文档所做出的对的估计：



求解策略方面，我们采用交替迭代的方法。首先保持**P**、**Q**不变以求解**A**，然后保持**P**、**A**不变求解**Q**，然后保持**A**、**Q**不变求解**P**。如此循环，直到残差（式中的求和结果）足够小或几乎保持不变为止。对求解最小二乘法的方法，我们采用了Matlab的默认实现，基于Lawson等人的工作[[[30]](#footnote-30)] 。我们在每一次迭代求解后对被求解的矩阵进行正规化以要求它满足对应的约束条件，然后继续进行迭代。

为加快求解，我们使用两次二元矩阵的分解得到算法运行前三个矩阵的初始值，再使用矩阵的伪逆运算得到每一次迭代中最小二乘的猜测值：



其中 表示一个矩阵的伪逆。当矩阵性质较差时，我们用近似伪逆特性的一个矩阵代替。

实验中，我们发现，当矩阵性质较差（这是通常的情况）时，有可能使得**A**矩阵的结果显得非常稀疏，甚至出现只有一个非零元的情况。究其原因，问题在于矩阵性质较差时通过伪逆给出的初值，其误差会随着不断的迭代而放大。为了改进这一情况，同时使得我们的结果中**A**矩阵的物理意义得以明确，我们增加了一个第二级的优化函数：



其中。每一轮迭代在最小二乘的意义上更新了**A**矩阵之后，我们保持**A**的约束条件，求解的一个极大值。为了不影响**P**、**Q**，这时**A**的约束条件除了式给出的之外，还要保持不变。在实验数据（Toy Data）上的结果验证了增加第二级优化函数的有效性。

前文已经提及，我们同样可以利用现有的各种模型，作为先验知识，添加到这个三元分解的过程中。其方法是利用现有模型，分别给出每个视图中话题和词汇的分布情况，以此构建**P**、**Q**矩阵的初值，进行运算。在本文的实验结果部分，我们将讨论不同初始化方法对聚类准确性或有效性的影响。

## 该模型的验证：在NIPS数据集上的结果

### 评价方法

参照Gu等人的文献[[[31]](#footnote-31)] ，我们选择了准确率、F值和NMI对聚类的准确率进行评价。设表示这些数据实际所具有的标签，取值范围为1到的正整数。用是聚类得到的标签，取值范围为1到。设是一个从到的映射，定义下式为聚类错误率：



并称为聚类的准确率（accuracy）。

显然我们无需像定义中的那样遍历所有的映射以得到准确率。用简单的贪心方法就可以得到这样的映射。

* 建立一个记数矩阵，其中#{在中被标为类而在中被标为的文档}。令。
* 选择中最大值的下标，并令。
* 将记数矩阵中和设为0，并令自增1。
* 若，返回第二步；否则，退出。

聚类纯净度的定义告诉我们，它允许聚类的结果是原先分类标签的细分，反之则不允许。然而如果只使用纯净度，可以给每一篇文档设置不同的聚类标签，从而使得纯净度为1，然而这时的聚类没有意义。为此，我们还需要引入召回率。亦即：



其中，当时，是一个从到（的一个子集）的双射；对于的情况，我们不妨定义这时的为。此时，我们着眼于在中属于同一标签的对象，在聚类结果中是否仍在同一聚类中。获得这样一个映射的过程与贪心法求的方法相似，只是在第三步中不将的值置零。

利用纯净率和召回率，我们同样可以计算F值，也就是这二者的调和平均数：



为了评价聚类的有效性，我们还引入了NMI（Normalized Mutual Information）。[[[32]](#footnote-32)] 其定义如下：



其中表示聚类簇中文档的数量，表示从属于的文档数量，表示与的交集中文档的数量。NMI越大（接近于1），则聚类的结果越好。

### 实验数据（Toy Data）的生成

按照模型假设，两个视图（view）下描述同一个对象的两篇文档是以如下方式产生的：

* 按照矩阵中给定的联合分布，选择一对话题。
* 选择文档的长度，在此被设定为一个常数，因为我们的模型只与文档中词的频率有关，与频数无关。
* 依据给出的多项式分布，在第一个视图下产生个词。
* 依据给出的多项式分布，在第二个视图下产生个词。

因而，我们需要事先准备好**P**、**Q**和**A**。我们以记第一个视图下话题的数量，为第二个，为每个话题中的词数。考虑到大量的词，其出现与话题的关系不大，以记这些词的数量。这些词在所有的话题中都以相同的分布出现。我们同时定义一个参数，表示了与话题有关的那些词与话题无关词出现概率的比例。在建立第一个实验数据（TOY1）时，我们使用平均分布。以的生成方式为例，对于每个话题，将设为一个行向量，其中下标到设为，最后维设为，其它维的值均置零。的生成方法与此类似。在TOY1中，**A**是人工设定的。

在构建第二个实验数据（TOY2）时，我们的策略是随机化的。以矩阵的生成为例，首先我们生成一个矩阵，其中的每列以如下方式生成：对于下标到（从1开始）的元素，我们从分布中随机取一个数值；对于最后的个元素，我们从分布中随机取一个数值。通过将正规化以使其每列的和为1，我们就得到了**P**。同时，为了测试在不同A矩阵条件下我们算法的有效性，我们的矩阵**A**通过对一个Beta分布进行采样给出。该Beta分布的参数选取方式如下。

记，。我们选取Beta分布的参数，以使得其期望为，而标准差为，其中。解方程，可以得到：



对于，直接生成一个元素全为的矩阵即可。

显然，当过小或过大，两个视图中的互信息量都会变小。两个视图中的互信息量如下：



因为许多Beta分布的随机变量的和的分布会收敛到一个高斯分布，我们可以近似地认为这一互信息熵的取值是的函数。其图像如下所示。（以0.1的步长变化）

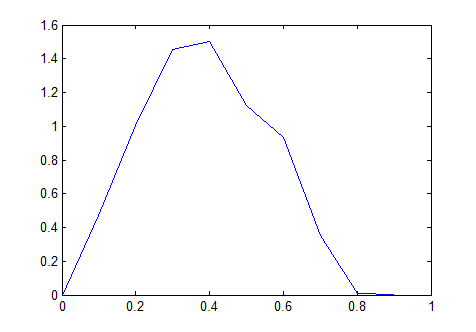


图 11 的取值与生成的话题联合分布的互信息熵变化关系图

### TOY1上的实验结果

在TOY1上得到的*ClusteringError*和F值如下。其中标“（M）”者为添加了第二级优化函数的结果。可见对各种初始化方法均有小幅度提高。

表 4 TOY1中运用二级优化函数后的效果提高

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | PLSA | RP[[[33]](#footnote-33)] | K-means | 两次分解[[[34]](#footnote-34)] |
| *ClusteringError* | 0.25180 | 0.28500 | 0.15590 | 0.05545 |
| F值 | 0.80977 | 0.79213 | 0.89685 | 0.95555 |
| *ClusteringError*（M） | 0.24620 | 0.28125 | 0.15325 | 0.05455 |
| F值（M） | 0.81525 | 0.79469 | 0.89760 | 0.95786 |
| F值提高  在所有测试用例中占比 | 43％ | 48％ | 50％ | 43％ |

### TOY2上的实验结果

TOY2中我们涉及到的参数有无关词数量和。

图 12 不同方法聚类结果NMI随的变化

图上，标有“-MF”的表示采用某种方法初始化两个视图，然后进行矩阵分解。例如“pLSA-MF”表示两个视图先用pLSA的结果初始化，再进行三元分解。可以看到，除了RP-MF因随机投影之后矩阵性质变得更差而不适合矩阵分解之外，其它方法运用矩阵分解分析之后得到的结果均好于原方法，并且其随着噪声词数量增加而效果下降的趋势比原方法更好。

图 13 不同方法聚类结果NMI随的变化

图中，能够看到对于pLSA和 RP两种初始化方法，矩阵分解分析得到的结果不如原方法；而对于K-means和NNMF两种初始化方法，所得到的结果，在较大时都好于原方法。这似乎说明了我们的方法更加偏向于较为集中的共现分布。

### 对科技文献的话题提取验证

NIPS数据集包含了1987至1999年NIPS（Neural Information Processing Systems，神经信息处理系统）会议的论文。这一会议包含了学习算法这一领域下许多不同的研究方向。原始的NIPS数据集包含了1740篇论文，总共有2,301,375个词和符号，其词汇量（词典大小）为13,649个词。为了获得更有意义的评价结果，我们去除了所有标为“未分类”的文档，并去除了出现频数过少的那些词，保留了1641篇文档，词典中包含6417个词。下面的实验中，一个视图包含了论文全文的词频数据，而另一个视图包含的是摘要的词频数据，给定每个视图下各自的话题数量为48个。

# 模型的应用

## 基于股价与经营范围描述的上市公司聚类

在本章中，我们主要考虑

## 在小规模数据集（Stock50）上的运行结果及分析

我们选取的时间段是2012年5月4日至2012年12月31日。在此时间段中，上证综指的走势如下图：



图 14 上证综指在2012年5月4日至2012年12月31日之间的走势

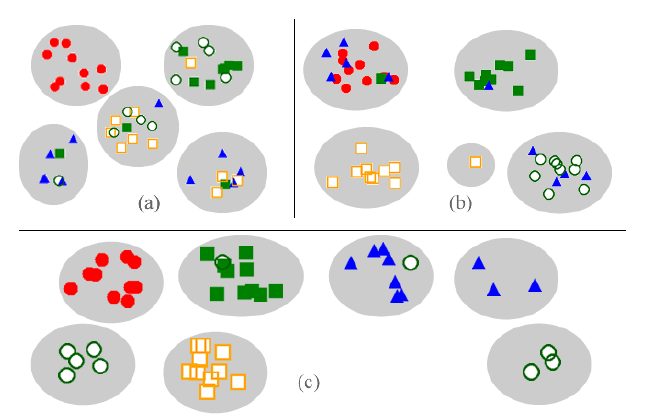


图 15 Stock50上聚类簇示意图  
（a）图为经营范围描述视图上LDA聚类的结果；（b）为在股价视图上K-means聚类的结果；（c）为我们的异质话题模型聚类的结果，两个视图分别以LDA和K-means初始化。

## 在大规模数据集（Stock2209）上的运行结果及分析

首先我们仍然从

## 在不同国家证券市场的运行结果及分析

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | NYSE | LSE | SGX | ASX |
| LDA-BSD (acc) | 0.6774 | 0.2099 | 0.4908 | 0.6352 |
| (nmi) | 0.4337 | 0.2377 | 0.2719 | 0.4671 |
| LDA-PD (acc) | 0.5810 | 0.1025 | 0.3524 | 0.3830 |
| (nmi) | 0.3262 | 0.1218 | 0.0823 | 0.0957 |

# 讨论与总结

（模型的不足、展望）

# 参考文献

# 致谢

1. [] Salton, Gerard, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing." Communications of the ACM 18.11 (1975): 613-620. [↑](#footnote-ref-1)
2. [] Ponte, Jay M., and W. Bruce Croft. "A language modeling approach to information retrieval." *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998. [↑](#footnote-ref-2)
3. [] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An Introduction to Latent Semantic Analysis. *Discourse Processes*, (25):259–284, 1998. [↑](#footnote-ref-3)
4. [] Hofmann, Thomas. "Probabilistic latent semantic indexing." *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999. [↑](#footnote-ref-4)
5. [] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003. [↑](#footnote-ref-5)
6. [] Chen, Xu, Mingyuan Zhou, and Lawrence Carin. "The contextual focused topic model." *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012. [↑](#footnote-ref-6)
7. [] McCallum, Andrew, Andrés Corrada-Emmanuel, and Xuerui Wang. The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. 2005. [↑](#footnote-ref-7)
8. [] A. McCallum. Multi-Label Text Classication with a Mixture Model Trained by EM. *AAAI'99 Workshop on Text Learning*. 1999. [↑](#footnote-ref-8)
9. [] Michal Rosen-Zvi, Tom Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors

   and documents. In *UAI*, 2004. [↑](#footnote-ref-9)
10. [] Boyd-Graber, Jordan, and David M. Blei. "Multilingual topic models for unaligned text." Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2009. [↑](#footnote-ref-10)
11. [] Nallapati, Ramesh, and William Cohen. "Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs." *International Conference for Weblogs and Social Media*. 2008. [↑](#footnote-ref-11)
12. [] Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." *Nature* 401.6755 (1999): 788-791. [↑](#footnote-ref-12)
13. [] Berry, Michael W., et al. "Algorithms and applications for approximate nonnegative matrix factorization." *Computational Statistics & Data Analysis* 52.1 (2007): 155-173. [↑](#footnote-ref-13)
14. [] Xu, Wei, Xin Liu, and Yihong Gong. "Document clustering based on non-negative matrix factorization." *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2003. [↑](#footnote-ref-14)
15. [] Arora, Sanjeev, Rong Ge, and Ankur Moitra. "Learning Topic Models--Going beyond SVD." *Foundations of Computer Science (FOCS)*, 2012 IEEE 53rd Annual Symposium on. IEEE, 2012. [↑](#footnote-ref-15)
16. [] Wang, Hua, Heng Huang, and Chris Ding. "Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization." *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011. [↑](#footnote-ref-16)
17. [] G. Doyle and C. Elkan. Financial topic models. In *Working Notes of the NIPS-2009 Workshop on "Applications for Topic Models: Text and Beyond Workshop"*, 2009. [↑](#footnote-ref-17)
18. [] Schumaker, Robert P., and Hsinchun Chen. "A quantitative stock prediction system based on financial news." *Information Processing & Management* 45(5): 571-583. 2009. [↑](#footnote-ref-18)
19. [] Zhiyuan Liu, Xinxiong Chen, Yabin Zheng, Maosong Sun. Automatic Keyphrase Extraction by Bridging Vocabulary Gap. *The 15th Conference on Computational Natural Language Learning* (CoNLL'11). [↑](#footnote-ref-19)
20. [] M. Yamamoto and K.W. Church, ‘Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus,’ *Computational Linguistics*, 27 (1), 1-30. [↑](#footnote-ref-20)
21. [] Maosong, Sun, Shen Dayang, and Benjamin K. Tsou. "Chinese word segmentation without using lexicon and hand-crafted training data." *Proceedings of the 17th international conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 1998. [↑](#footnote-ref-21)
22. [] H. Feng, ‘Accessor variety criteria for Chinese word extraction,’ *Computational Linguistics*, 30 (3), 75-93. [↑](#footnote-ref-22)
23. [] C.E. Shannon, ‘A Mathematical Theory of Communication,’ *Bell System Technical Journal*, 27 (3): 379-423. [↑](#footnote-ref-23)
24. [] 胡裕树. 现代汉语. 上海教育出版社, 2010. [↑](#footnote-ref-24)
25. [] 顾森. “SNS中的文本数据挖掘.” 程序员 (2012): 113–115. [↑](#footnote-ref-25)
26. [] http://www.icl.pku.edu.cn/icl\_res/ [↑](#footnote-ref-26)
27. []《现代汉语常用词表》课题组. 现代汉语常用词表（草案）. 商务印书馆, 2008. [↑](#footnote-ref-27)
28. [] Griffiths, Thomas. "Gibbs sampling in the generative model of latent Dirichlet allocation." Unpublished note. http://citeseerx. ist. psu. edu/viewdoc/summary (2002).；Griffiths, Tom, and Mark Steyvers. "A probabilistic approach to semantic representation." *Proceedings of the 24th annual conference of the cognitive science society*. 2002. [↑](#footnote-ref-28)
29. [] Ding, Chris, et al. "Orthogonal nonnegative matrix t-factorizations for clustering." *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006. [↑](#footnote-ref-29)
30. [] Lawson, Charles L., and Richard J. Hanson. *Solving least squares problems*. Vol. 161. Englewood Cliffs, NJ: Prentice-hall, 1974. [↑](#footnote-ref-30)
31. [] Gu, Quanquan, and Jie Zhou. "Local learning regularized nonnegative matrix factorization." *Twenty-First International Joint Conference on Artificial Intelligence*. 2009. [↑](#footnote-ref-31)
32. [] Lancichinetti, Andrea, Santo Fortunato, and János Kertész. "Detecting the overlapping and hierarchical community structure in complex networks." *New Journal of Physics* 11.3 (2009): 033015. [↑](#footnote-ref-32)
33. [] 所采用的方法是随机投影（Random Projection）降维后再通过K-means聚类。随机投影的方法参见：Bingham, Ella, and Heikki Mannila. "Random projection in dimensionality reduction: applications to image and text data." *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001. [↑](#footnote-ref-33)
34. [] 见3.2节。 [↑](#footnote-ref-34)