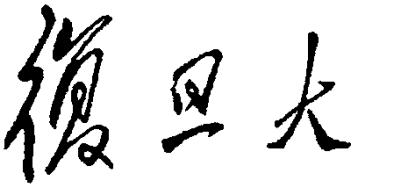

本科毕业论文

论文题目：基于隐模型的财经分析研究和应用

院　　系：计算机科学技术学院

专　　业：计算机科学与技术

姓　　名：朱恬骅

学　　号：09300240004

指导教师：池明旻副教授

2013 年 6 月 15 日

摘要

在本文中，研究了一种隐模型——异构的双视图主题模型，以利用从不同来源得到的、不同物理意义和特征的数据进行聚类。本文从矩阵分解的角度出发，将对双视图数据的建模和参数求解过程转换为带有线性约束条件的非负矩阵的三元分解问题。双视图下各自由主题到文档的生成过程被认为是独立的，从而将（不同的视图）数据矩阵之间的相关性归结于各自主题之间的联系。我们的方法适用于任意的单视图主题-文档生成模型而不需要规定其所符合的概率分布的形式。该方法另一优点是，我们解决方案本身的复杂程度关于样本数是不变的。这意味着我们可以在数据生成的时候线性地计算预期矩阵 **E**。

本文展现了本模型在中国 **A** 股的经营范围描述（文本）和分时段价格数据上的实验结果，并且也在海外股市上进行了类似的实验。实验结果表明：（1）在两个视图上共生的高相关主题可以通过共现概率矩阵 **A** 进行识别，（2）在同一个部门（主题聚类簇）的股票，可以得到比专家人工标注或传统基于单一视图信息的聚类方法更高的两两之间相关度。

关键词：异构，双视图主题模型，财经数据分析

Abstract

Heterogeneous two-view topic modeling is researched in this thesis, in order to utilize data of different physical meaning or from different sources. Starting from the perspective of matrix factorization, we transformed the problem of two-view topic modeling to a nonnegative matrix tri-factorization problem with linear constraints. We regard the topic-document generation process in each view as independent, so that the linkage between data from two views is considered as the result of correlation between topics in both views. Our method is suitable for arbitrary topic-document generation models, regardless of specific forms of probability distributions. Furthermore, the complicity of our solution is constant with regard to the number of samples.

The thesis also showed the results of experiments conducted on data from Chinese A-Share Stock market as well as several stock markets of global vitality, using the business scope descriptions and price data as the two views. The results showed that: (1) the matrix **A** in our factorization identifies pairs of topics of high correlation, and (2) the price trend of stocks in same cluster is expected to enjoy higher correlation in long term prediction than those with same labels tagged by human experts or traditional single-view topic model clustering method.

Keywords: Heterogeneous, two-view topic model, financial data analysis

目录

1. 概述.....1
1.1. 大数据与财经领域的关联.....1
1.2. 主题模型与财经数据分析.....2
1.3. 本文的结构安排.....3
2. 先前的研究进展.....4
2.1. 主题模型（Topic Model）.....4
2.2. 利用异构数据的主题模型.....5
2.3. 双视图下的主题模型.....7
2.4. 主题模型与矩阵分解.....8
2.5. 股票聚类方面的研究.....9
3. 数据的采集与预处理.....11
3.1. 价格信息的采集与预处理.....11
3.2. 文本信息的采集.....15
3.3. 财经专业词汇的提取.....18
4. 双视图主题模型的构建.....28
4.1. 基于矩阵分解的主题模型.....28
4.2. 三元分解求解的局部最优解算法.....31
4.3. 三元分解求解的线性规划算法.....33
4.4. 该模型的验证.....35
5. 模型的应用.....41
5.1. 基于股价与经营范围描述的上市公司聚类.....41
5.2. 在小规模数据集（Stock50）上的运行结果及分析.....41
5.3. 在海外证券市场的运行结果及分析.....43
5.4. 在大规模数据集（Stock2209）上短时段的运行结果及分析.....45
5.5. 在大规模数据集（Stock2209）上长时段的运行结果及分析.....47
6. 讨论与总结.....50
参考文献.....51
致谢.....54

第1章 概述

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。本课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。

1.1. 大数据与财经领域的关联

随着信息和通信技术的发展，尤其是互联网的兴盛，存在于计算机数码空间（cyber-space）的数据量已达到了海量。随之兴起了大数据的概念。

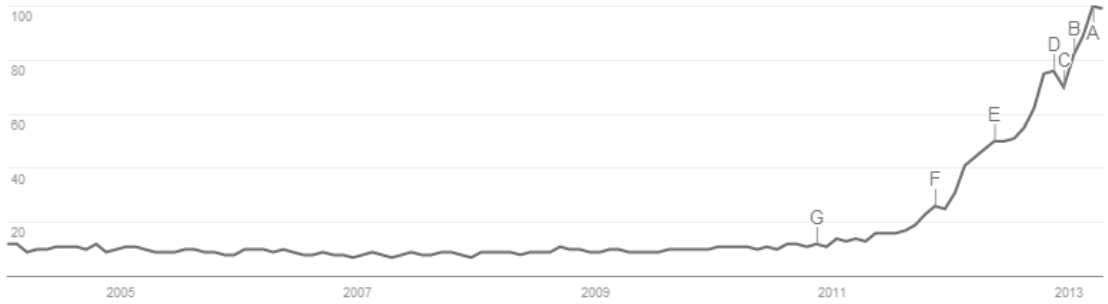

图 1 Google Trends 关于“大数据”的趋势图[1]

2011 年，Facebook 进行首次公开募股（Initial Public Offerings，IPO）。相关的研究发现，Twitter 微博上关注 Facebook IPO 事件的用户，所发微博中的情绪信息能够预测 Facebook 的价格走势。[2] 与之相关地，不少企业已经开始了大数据分析的产业化。例如，IBM 公司发布了 InfoSphere Streams 软件，能够分析和共享运行中的数据，在需要每秒做出百万次决定的环境中，以亚毫秒的速度做出

[1] Google Trends 结果，网址：http://trends.google.com/trends/explore#q=big%20data，获取日期：2013-04-04
[2] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2011, 2(1): 1-8.

决策。这一软件能以每天分析 PB 量级的速度，持续分析海量数据。[3] 商业咨询机构德勤在一份名为《大数据 2.0 时代的新商务战略》的报告中认为，大数据具有三种应用类型和商业价值："回答现有业务的已知问题，专注于提升业绩和运营效率；回答现有业务中的新问题，专注于业务增长机会；回答新业务的新问题，目标是改写竞争格局。"[4] 大数据在财经领域的重要性已经毋庸置疑。

1.2. 主题模型与财经数据分析

传统的经济分析将经济运行看作一个过于简化的由大型经济实体（企业等）驱动的系统，其中，小公司要么被忽略，要么只是将它们集合起来统一考虑。[5] 重要原因在于，传统的经济分析无法应对如今成倍增长的公司数据。各种经济实体之间的联系日益增长，如果只是依靠计算机进行简单的数据预处理，然后将分析工作完全依赖于人工，将越来越无法应对如此大量的相关数据，从而影响了对整体局势的判断，也影响了对个别企业的深入研究。

对此，经济学者已经提出了一系列强调关联性的经济模型，例如网络模型[5] 、拓扑模型[6] 、随机游走模型[7] 等。我们认为，主题模型以其对于"隐形参数"的关注，可以作为传统经济和网络理论之间的桥梁，从而获得关于宏观情况更加准确、有事实基础的了解。先前的研究已经证明（参见第第 2 章 章），运用主题模型分类的股票生成了有意义的行业，并且从行业、公司类型等各方面提供了参考。

然而，先前的研究都只能处理同一种类型的数据，如股价数据或文本数据等，而不能将两者结合起来进行考虑。简单的数据表示上的拼装（例如，将文本和股价都表示为一个向量，然后将二者简单连接起来组成一个新的特征向量，参与到聚类过程中去）将大大增加数据的维度，从而进一步加剧"维数灾难"问题，得不到更好的结果（相关数据见 4.4.5 节）。这是由于两种数据的异质特征导致的。如何正确利用更多信息，而尽量避免其带来的负面影响，这是本文所要进行探索的。

[3] IBM：全面贯彻大数据能力 助力各行业赢得时代挑战，网址：http://www.ibm.com/news/cn/zh/2012/11/30/d163904c58088v20.html，获取日期：2013-04-04

[4] Gopalkrishnan V, Steier D, Lewis H, et al. BIG DATA 2.0: New business strategies from big data[J]. Deloitte Review, 2013, 12(1): 56

[5] Schweitzer F, Fagiolo G, Sornette D, et al. Economic networks: The new challenges[J]. Science, 2009, 325(5939): 422.

[6] Serrano M A, Bogun á M. Topology of the world trade web[J]. Physical Review E, 2003, 68(1): 015101.

[7] Reyes J, Schiavo S, Fagiolo G. Assessing the evolution of international economic integration using random walk betweenness centrality: The cases of East Asia and Latin America[J]. Advances in Complex Systems, 2008, 11(05): 685-702.

本文的目的正是在于，提出一种可以适用于大量数据场景的算法，能够较快地对不同性质和来源的数据进行分析，对不同股票之间的相关性进行挖掘和发现。

1.3. 本文的结构安排

本文的结构如下。第 1 章，介绍选题的背景和目的，说明本文的总体思路。

第 2 章，介绍与本文相关的各项主要工作领域当前的研究状况。首先回顾了主题模型的发展历史，并探讨了其中涉及异构数据与多（双）视图聚类问题的工作，然后通过揭示主题模型和矩阵分解的内在关系，说明本文的思路，介绍相关的研究成果。然后，本文还考察了先前的一些利用数据挖掘方法进行股票聚类的研究。

第 3 章描述了考察的各数据集，并给出预处理的具体方法。其中，介绍了一种基于信息熵的中文文本中词语自动抽取的方法，并用其生成了词典。介绍了相关的实验，就其参数的选择做出了说明，并检验了其效果。

第 4 章，引入本文所采用的基于非负矩阵分解的异构主题模型。说明了验证其有效性的实验数据集（toy data）的生成方法，分析了实验结果。

第 5 章，将第 4 章中提出的模型应用于第 3 章给出的数据集上，给出分析的结果及解释。

最后，在第 6 章中总结该异构主题模型的优缺点，并对其未来的应用和改进提出展望。

第2章 先前的研究进展

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[8]。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[9] 。

统计语言模型[10] 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，主题模型得以产生，为词和文档之间增加了一个新的层面。这也就是主题（topic）的层面。

2.1. 主题模型（Topic Model）

其中，潜在语义分析（latent semantic analysis，LSA）[11] 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）[12] 通过引入概率模型解决了这一问题。在 LSA 中，每个语义

[8] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.
[9] Deza M M, Deza E. Encyclopedia of distances[M]. Springer Berlin Heidelberg, 2009.
[10] Ponte J M, Croft W B. A language modeling approach to information retrieval[C]//Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998: 275-281.
[11] Landauer T K, Foltz P W, Laham D. An introduction to latent semantic analysis[J]. Discourse processes, 1998, 25(2-3): 259-284.
[12] Hofmann T. Probabilistic latent semantic indexing[C]//Proceedings of the 22nd annual international ACM

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－主题分布和主题－词组分布被看作是参数而非随机变量。

针对这一点，Blei 等人[13] 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主题（topics），每个主题z表示为词典V上的一元语言模型θ_z，即一个多项式分布。每个文档d对应这T个主题有一个关于文档的多项式分布ϕ_d。LDA 假定，$\phi_d \sim Dir(\alpha)$，对于文档d中的每一个词w，有主题$z \sim Multi(\phi_d)$，$w \sim Multi(z)$。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，$\theta_z \sim Dir(\beta)$。 LDA 的图模型表示如图 2。

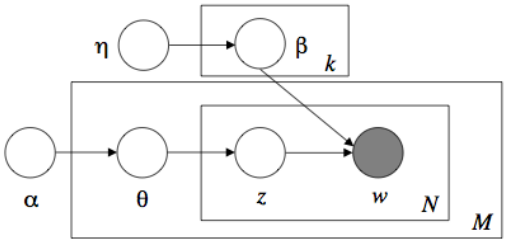

图 2 LDA 的图模型表示[13]

2.2. 利用异构数据的主题模型

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－主题模型（Author-Topic Model）、作者－接收者－主题模型（Author-Recipient-Topic Model），以及考虑了各种要素的 cFTM 模型[14]（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

SIGIR conference on Research and development in information retrieval. ACM, 1999: 50-57.

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

[14] Chen X, Zhou M, Carin L. The contextual focused topic model[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 96-104.

间关系上进行的聚类，是多视图聚类的一个重要侧面。

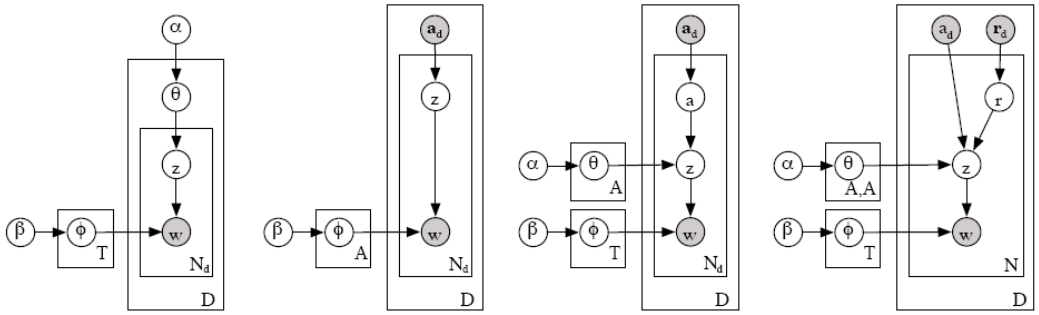

图 3 LDA 及其几种拓展[15]

从左至右：LDA[13]、Author Model (Multi-label Mixture Model)[16] 、Author-Topic Model[17] 、Author-Recipient-Topic Model[15]

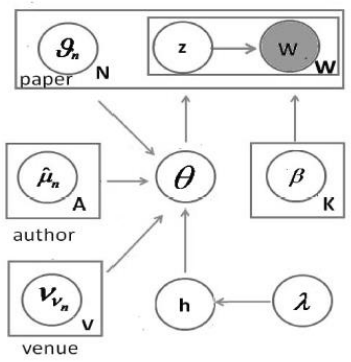

图 4 cFTM 的图模型表示[14]

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）。

[15] McCallum A, Wang X, Corrada-Emmanuel A. Topic and role discovery in social networks with experiments on enron and academic email[J]. Journal of Artificial Intelligence Research, 2007, 30(1): 249-272.
[16] McCallum A. Multi-label text classification with a mixture model trained by EM[C]//AAAI'99 Workshop on Text Learning. 1999: 1-7.
[17] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, 2004: 487-494.

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。

2.3. 双视图下的主题模型

在双视图的研究方面，Jordan 等人[18] 提出了用主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。

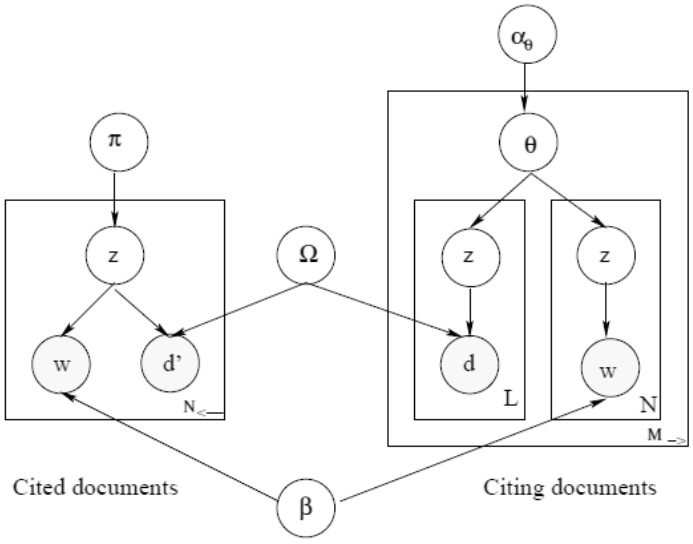

图 5 Link-PLSA-LDA 模型的图模型表示[19]

Nallapati 等人[19] 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

[18] Boyd-Graber J, Blei D M. Multilingual topic models for unaligned text[C]//Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2009: 75-82.
[19] Nallapati R, Cohen W. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs[C]//International Conference for Weblogs and Social Media. 2008: 84-92.

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。Link-PLSA-LDA 模型的结构可以简化为如下图：

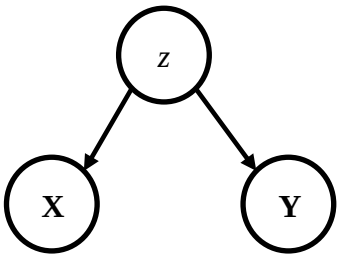

图 6 Link-PLSA-LDA 模型简化示意图

Link-PLSA-LDA 模型将两个视图中的词都视作同一组主题产生。然而这两个视图中的数据实际有着不同的物理含义和重要性，这样的混杂将可能导致主题之间的交叉过多而影响聚类的效果。

2.4. 主题模型与矩阵分解

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（主题）发现过程。如果我们将整个主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主题，而不须知道或假设其所符合的先验分布。早期 Lee 等人[20] 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[21] 、Xu 等[22] 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主题的有效性。

[20] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755): 788-791.
[21] Berry M W, Browne M, Langville A N, et al. Algorithms and applications for approximate nonnegative matrix factorization[J]. Computational Statistics & Data Analysis, 2007, 52(1): 155-173.
[22] Xu W, Liu X, Gong Y. Document clustering based on non-negative matrix factorization[C]//Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003: 267-273.

作为类比，在我们期望提出的模型中，主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[23] 提出了一种处理对称阵三元分解的聚类算法，可以处理异构的数据。不过，该方法关于对称阵的限制拘束了我们的应用范围。

2.5. 股票聚类方面的研究

在关于股票聚类的研究方面，Doyle 等人[24] 较早地利用主题模型对股价进行聚类，该文提出了采用主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，但是由于主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，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 3.1 节。简言之，Doyle 等人得到的话题直接表示了股票之间的同涨跌关系；而我们的方法得到的话题，其在股价的维度上可以表现为一种“走势”。

在通过文本对股票进行聚类这一研究方面，R. Schumaker 等人[25] 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。其训练方式为：输入财经新闻的文本数据，将其用一组布尔值表示，每个布尔值表示某个专有名词在文本中是否存在；同时输入提及的这一股票在新闻发布时的价格和新闻发布 20 分钟之后的价格。对这些输入进行建模、训练，由此达到使用新闻文本预测股价的目的。

[23] Wang H, Huang H, Ding C. Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization[C]//Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011: 279-284.
[24] Doyle G, Elkan C. Financial topic models[C]//Working Notes of the NIPS-2009 Workshop on” Applications for Topic Models: Text and Beyond Workshop. 2009.
[25] Schumaker R P, Chen H. A quantitative stock prediction system based on financial news[J]. Information Processing & Management, 2009, 45(5): 571-583.

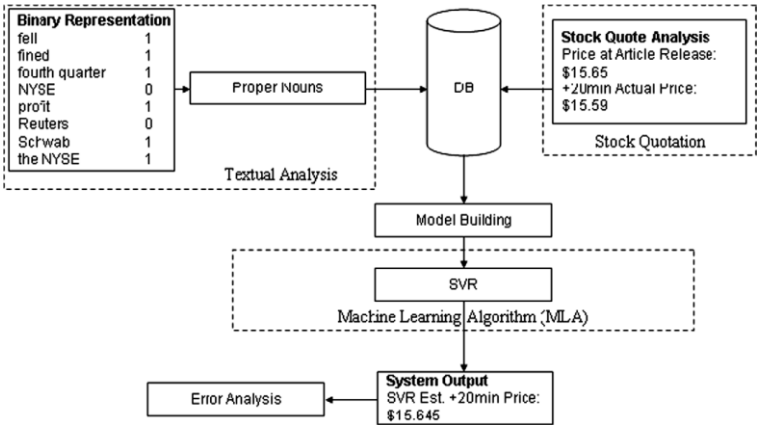

图 7 AZFinText 的工作原理[25]

第3章 数据的采集与预处理

数据的采集和预处理是所有后续工作的基础。在本章中，我们将首先交待数据来源和考察范围。然后，针对较为复杂的中文文本处理，说明本文所采用的一些预处理手段，并就其有效性和参数选择予以论证。

3.1. 价格信息的采集与预处理

3.1.1. 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：

表 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。

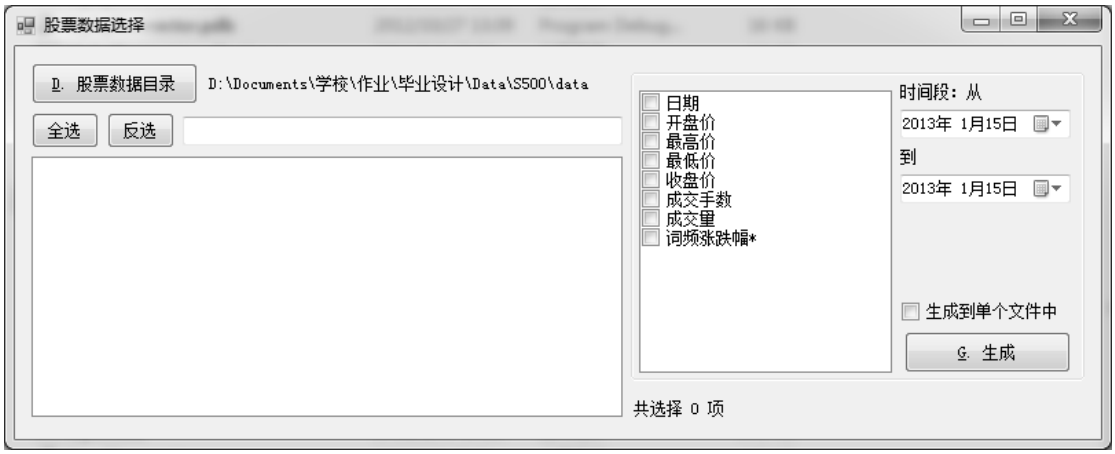

图 8 股票数据选择图形界面

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（business scope descriptions，BSD）文本和行业分类数据来源于各交易所网站。

3.1.2. 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 $T+1$ 天的股价信息，我们将这 T 天的股价信息扩展为一个 $2T$ 维的向量，其中第 i 维数值大小，当 i 为偶数时表示的是第 $i/2+1$ 天相对于前一天的跌幅，而对奇数 i，则表示第 $(i+1)/2+1$ 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 $T+1$ 天的价格信息，建立一个大小为 $3T$ 的词汇表，包括了“第 i 天涨”、“第 i 天持平”和“第 i 天跌”，$i=2,3,...,T+1$。将股票的走势表现为这 $3T$ 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：

表 2 K-means 在 12 支股票数上聚类的结果

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。

3.1.3. 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。

3.1.3.1. 小规模数据集：Stock50、RStock50

3.1.3.1.1. 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（Stock50）。这些股票都是所在版块有代表性的股票，我们用其行业名称作为标签。

表 3 Stock50 的股票列表

银行：	000629 攀钢钒钛	600511 国药股份
000001 深发展 A	000709 河北钢铁	600833 第一医药
002142 宁波银行	000717 韶钢松山	600713 南京医药
600000 浦发银行	000898 鞍钢股份	000028 国药一致
600015 华夏银行	000932 华菱钢铁	600085 同仁堂
600016 民生银行	600282 南钢股份	600129 太极集团
600036 招商银行	600019 宝钢股份	600422 昆明制药
601009 南京银行	000959 首钢股份	000423 东阿阿胶
601166 兴业银行	000022 济南钢铁	002589 瑞康医药
601169 北京银行	600569 安阳钢铁	酒：
601288 农业银行	医药：	000568 泸州老窖
钢铁：	601607 上海医药	000858 五粮液

600519 贵州茅台	600559 老白干酒	002065 东华软件
600779 水井坊	软件：	000938 紫光股份
000596 古井贡酒	600570 恒生电子	002073 软控股份
600809 山西汾酒	600756 浪潮软件	002090 金智科技
000799 酒鬼酒	000948 南天信息	002230 科大讯飞
002304 洋河股份	600271 航天信息	
600702 沱牌舍得	002063 远光软件	

3.1.3.1.2. 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）。

表 4 RStock50 的股票列表

钢铁：	601001 大同煤业	000089 深圳机场
000629 攀钢钒钛	601666 平煤股份	600004 白云机场
000709 河北钢铁	601898 中煤能源	600009 上海机场
000717 韶钢松山	汽车：	600151 航天机电
000898 鞍钢股份	000550 江铃汽车	600893 航空动力
000932 华菱钢铁	000572 海马汽车	000901 航天科技
600282 南钢股份	000625 长安汽车	电力：
600019 宝钢股份	000800 一汽轿车	600011 华能国际
000959 首钢股份	000868 安凯客车	600021 上海电力
000022 济南钢铁	000927 一汽夏利	600027 华电国际
600569 安阳钢铁	000951 中国重汽	600644 乐山电力
煤炭：	000957 中通客车	600101 明星电力
000780 平庄能源	002594 比亚迪	600116 三峡水利
000723 美锦能源	600006 东风汽车	600131 岷江水电
002128 露天煤业	航空：	600236 桂冠电力
600188 兖州煤业	600029 南方航空	600292 九龙电力
600348 阳泉煤业	600115 东方航空	600310 桂东电力
600546 山煤国际	600221 海南航空	
600740 山西焦化	600316 洪都航空	

3.1.3.2. 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。

3.1.4. 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，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：

表 5 海外股市的股价基本数据

证交所	NYSE	LSE	SGX	ASX
股票数量	2024	2030	542	1368
行业分类数量	11	46	14	26

3.2. 文本信息的采集

本文中涉及的文本信息主要是上市公司的经营范围描述。在所有 2209 支股票的经营范围描述文本信息中，采用 3.3 中构建关键词词典方法，查找出的前 100 个高频词如下：

表 6 中国 A 股市场上市公司经营范围描述前 100 个高频词

经营	企业	配件
技术	项目	电子
销售	机械	禁止
生产	工程	管理
业务	国家	公司
出口	商品	不含
设备	加工	进出口业
服务	咨询	进出口业务
开发	除外	相关
进出口	制造	许可证
材料	本企业	化工
许可	投资	设计

代理	金属	货物
机械设备	规定	安装
系统	信息	公司经营
仪器	汽车	原辅
建筑	法律	贸易
营本企业	仪表	零配件
计算机	法规	辅材料
经营本企业	范围	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[26] 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 D 中的一篇文档 d ，需要按 $Pr(p|d)$ 对候选关键词 p 进行排序，选择概率最高的 M_d 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 $Pr_{<D,T>}(t|w)$ ，其中 t 是标题中的单词，而 w 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 w 的重要程度分数 $Pr(w|d)$；b. 计算候选关键词 p 的排位分数 $Pr(p|d)=\sum_{t\in p}\sum_{w\in d}Pr_{<D,T>}(t|w)Pr(w|d)$；c. 选出 $Pr(p|d)$ 最高的 M_d 个候选关键词，作为 d 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 中的每一个主题那样，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

[26] Liu Z, Chen X, Zheng Y, et al. Automatic keyphrase extraction by bridging vocabulary gap[C]//Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2011: 135-144.

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：

表 7 中国 A 股市场上市公司股票经营范围描述的关键词

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 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

3.3. 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。

3.3.1. 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[27] 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[28] 就利用信息熵对中文分词进行了实验。Feng[29] 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。[30] 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 X，字母表上的概率分布 P。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 S，在每单位时间中，从一个有限的集合 $X=\{X_1,X_2,\ldots,X_N\}$（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 S 是离散的。在 T 时间中，S 产生的符号可用一个序列表示：$\{x_1,x_2,\ldots,x_T\}$。若这一过程中，事件 $x_t=X_j$ 发生的概率与时间 t 无关，也与前一个符号 x_{t-1} 的取值无关。我们称这个信源 S 是无记忆（memoryless）的。满足上述两个特征的信源 S 即为 DMS。相反地，如果符号 x_t 的取值与 x_{t-1} 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 $X_1,X_2,\ldots$ 满足下列等式：

$$\Pr(X_{n+1}=x\mid X_1=x_1,X_2=x_2,\ldots,X_n=x_n)=\Pr(X_{n+1}=x\mid X_n=x_n) \qquad (1)$$

亦即，$t=n$ 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

[27] Yamamoto M, Church K W. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus[J]. Computational Linguistics, 2001, 27(1): 1-30.
[28] Sun M, Shen D, Tsou B K. Chinese word segmentation without using lexicon and hand-crafted training data[C]//Proceedings of the 17th international conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 1998: 1265-1271.
[29] Feng H, Chen K, Deng X, et al. Accessor variety criteria for Chinese word extraction[J]. Computational Linguistics, 2004, 30(1): 75-93.
[30] Shannon C E. A mathematical theory of communication[J]. ACM SIGMOBILE Mobile Computing and Communications Review, 2001, 5(1): 3-55.

以定义 2 至 m 阶马尔可夫性质：

$$\begin{aligned}&\Pr(X_n = x_n \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1)\\&= \Pr(X_n = x_n \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_{n-m} = x_{n-m}) \quad \text{对于} n > m\end{aligned} \tag{2}$$

具有 m 阶马尔可夫性质的随机变量系列 $X_1, X_2, \ldots, X_N$ 称为 m 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 $\Pr(X_n = x_n \mid \ldots)$，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 m 阶马尔可夫链真包含 $m-1$ 阶马尔可夫链。

3.3.2. 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 X 的信源 S 所产生，其中 X 代表一个有限的现代汉语词汇的子集，S 是一个 DMS。在不引起混淆的情况下，我们用 $P(X_i)$ 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 S 产生的符号序列 $\mathcal{W}$ 来猜测 S 的字母表 X 和概率分布 P。

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 X。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 $W = c_1 c_2 c_3 \ldots c_{n(W)}$，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 $\Gamma = \{\gamma_1, \ldots, \gamma_{N_s}\}$，其中每个句子都是一个字串 $\gamma_i = \{c_1, c_2, \ldots, c_{N_s(i)}\}$，其中 c_i 是一个汉字。我们称之为在信宿视角下的信源 S，记为 S'。由于我们将抽象的层次从词汇下降到了字符，显然 S' 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 m 阶马尔可夫性，故字符集上的概率分布可以用

$Pr(X_i = x_i \mid X_{i-1} = x_{i-1},\ldots,X_{i-m+1} = x_{i-m+1})$来表示。

设 X 是一个随机变量，其可能的取值范围为 $\{x_1,x_2,\ldots,x_n\}$，$n<+\infty$，它们对应的概率分布为 p。X 的信息熵 $H(X)$ 定义为：

$$H(X)=\mathrm{E}(I(X)) \tag{3}$$

其中 $I(X)$ 称为 X 的信息量，其定义为：

$$I(X)=\sum_{i=1}^{n}-p(x_i)\log_b p(x_i) \tag{4}$$

被求和项称为单个符号的信息量，用 $I(x_i)$ 表示。对数函数的基 b 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 9。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。

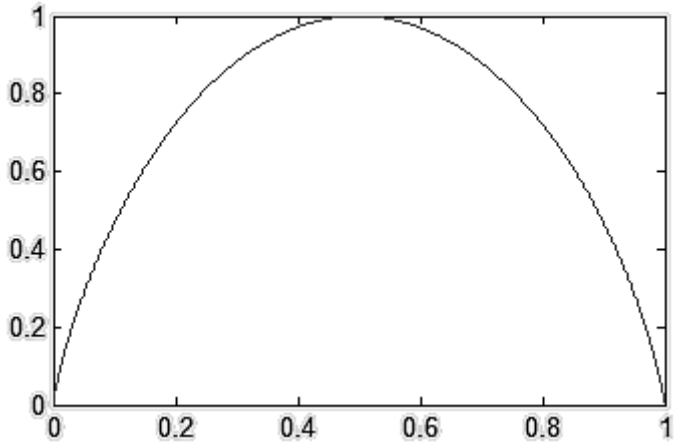

图 9 0-1DMS 的信息熵函数 H(x)，横轴表示 Pr(X=1)

3.3.3. 对词汇的定义

词汇是“语言中能独立运用的最小语法单位”。[31] 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。

[31] 胡裕树. 现代汉语[M]. 上海教育出版社, 2010.

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[32]，“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S'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S'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W对应两个信源，一个是前导字串信源S_W^L，另一个是后继字串信源S_W^R。这两个信源的字母表皆为全体汉字的集合C，显然C是有限而离散的。在前文定义的基础上，我们定义信源S_W^L的概率分布为$\Pr(c_i) = freq(cW)_L P$，信源S_W^R的概率分布为$\Pr(c_i) = freq(Wc_i) = P_R(c_i)$。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$H(S_W^L)$、$H(S_W^R)$。参考图 9，只有当$P_x(x = L, R)$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$free(W)$为：

$$free(W) = \min\{H(S_W^L), H(S_W^R)\} \tag{5}$$

为了用统计的方法能够衡量词汇内在的凝固性，我们考虑 3.3.1 中定义的信源S。假设字串$W = c_1c_2c_3\ldots c_n$是一个词汇，显然有

[32] 顾森. SNS 中的文本数据挖掘[J]. 程序员, 2012 (8): 113-115.

$$\hat{p}(W) = \max_{a} \prod_{i=1}^{\|a\|-1} (freq(c_{a_i} c_{a_i+1} \ldots c_{a_{i+1}-1})) >> freq(W) \quad (6)$$

其中a是对数列$\{1,2,3,\ldots,n\}$的一个划分，函数$freq(W)$表示一个字串W在观测到的符号串中出现的频率。因此，我们定义字串S的凝固度$coh(W)$为：

$$coh(W) = \frac{freq(W)}{\hat{p}(W)} \quad (7)$$

我们引入三个参数θ_f，θ_c和m。其中m是信源S'所满足的马尔可夫性质的最大阶数，亦即词的最大长度。$\theta_f < free(W)$且$\theta_c < coh(W)$时，字串W才能够被认为可能是一个词。

3.3.4. 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》[33] 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

[33] http://www.icl.pku.edu.cn/icl_res/

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 8 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					
正确词汇数量	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）》[34] ），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θ_c和θ_f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$\theta_c = 100$不变，更改θ_f的值，召回率和正确率随θ_f的变化如图 10 所示。

[34]《现代汉语常用词表》课题组. 现代汉语常用词表（草案）[M]. 商务印书馆, 2008.

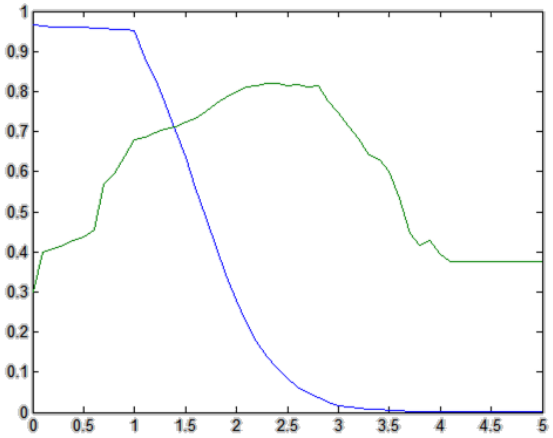

图 10 召回率和正确率随θ_f的变化（蓝色表示召回率，绿色表示正确率）

在保持$\theta_f = 1.3$不变的情况下，更改θ_c的大小。图 11 展示了二者随θ_c变化的曲线。

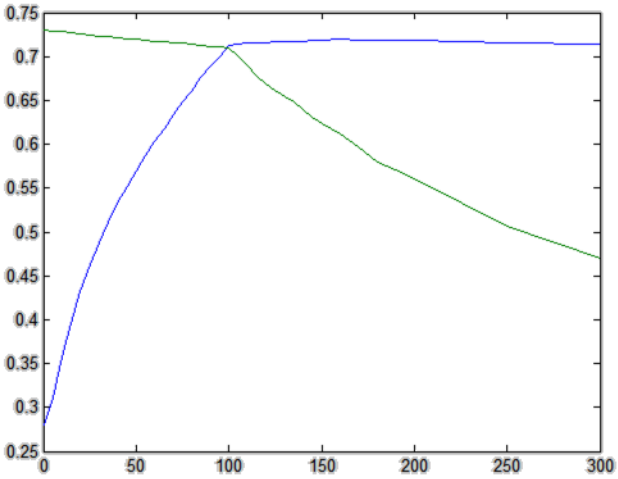

图 11 召回率和正确率随θ_c的变化（蓝色表示召回率，绿色表示正确率）

最终选定的参数为$\theta_c = 100$，$\theta_f = 1.3$。

3.3.5. 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 9 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 10 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

第4章 双视图主题模型的构建

在基于贝叶斯模型的主题模型中，我们需要计算前后验概率分布。在通常情况下，我们用采样（如 Gibbs 采样）来解决积分的问题，这会带来计算复杂度的提高和准确性的下降，需要在二者之间凭借经验进行平衡。此外，如前所述，对数据所符合的分布进行先验假设可能无法有效应对我们的问题，因为股价的波动实际是不可知的。因此，我们引入基于非负矩阵分解的双视图主题模型。

4.1. 基于矩阵分解的主题模型

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主题空间，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。

主题模型通过将文档视作若干主题的混合以应对聚类问题，已得到广泛的应用。然而，目前大多数的主题模型都是基于概率隐主题分析建立的。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，主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[35] 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，我们提出了异构主题模型，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，“主题”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主题

异构主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

[35] Griffiths T, Steyvers M. A probabilistic approach to semantic representation[C]//Proceedings of the 24th annual conference of the cognitive science society. 2002: 381-386.

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主题之间的关系，而 **A** 表示了两组主题的相互关系。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）[36] 。但与 Ding 等人的工作所不同的是，每个视图中，我们并不要求主题－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主题之间的相关性。同时，**A** 矩阵包含了不同视图中不同主题之间的共现概率。从而，我们期望可以从两组主题之间的相关性，利用异构数据，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（异构）数据的主题模型；

（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主题间的相关性。

考虑两个相互有关联的高维随机向量 $\mathbf{X}\in R^p$、$\mathbf{Y}\in R^q$，在双视图异构矩阵分解中，我们的目标是寻找到矩阵 $\mathbf{P}\in R^{p\times s}$、$\mathbf{Q}\in R^{q\times t}$ 及 $\mathbf{A}\in R^{s\times t}$，使得：

$$\mathrm{E}[XY^\top]\approx \mathbf{PAQ}^\top \tag{8}$$

这一问题便转化为了对下列目标函数的优化问题：

$$L_0(\mathbf{PAQ},\mathrm{E}\mathbf{XY}^\top)+R_0(\mathbf{P},\mathbf{A},\mathbf{Q}) \tag{9}$$

其中 L_0 是误差函数，而 R_0 是正规化函数。给定如是的矩阵 **P**、**Q**，我们能够将 $X\in R^p$ 从 p 维缩减到 $\bar{X}=P^\top X\in R^s$ 的 s 维。类似地，也就将 $Y\in R^q$ 从 q 维缩减到 $\bar{Y}=Q^\top Y\in R^t$ 的 t 维。

假定数据集中有 n 个对象，在第一个视图中这 n 个对象的特征表示为 $\mathbf{X}=(\mathbf{x}_i)_{i=1}^n$，$\mathbf{x}_i\in R^p$，在第二个视图中表示为 $\mathbf{Y}=(\mathbf{y}_i)_{i=1}^n$，$\mathbf{y}_i\in R^q$。同传统的主题模型一样，我们引入两个隐随机变量 $(w;z)$。这两个变量都是不可直接观察到的，用于表示隐藏的主题层面，其中 $w\in R^s$、$z\in R^t$，分别为文档在两个视图中各主题的权重。相应地，在第一个视图中，一个输入向量 x 是 s 个主题的混合。设若 $p_i^{(1)}(x)$ 表示第一个视图中主题 i 的概率，则有：

[36] Ding C, Li T, Peng W, et al. Orthogonal nonnegative matrix t-factorizations for clustering[C]//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006: 126-135.

$$p(\mathbf{x}\,|\,\mathbf{w})=\sum_{i=1}^{s}w_i\,p_i^{(1)}(x) \tag{10}$$

相似地， y 是 t 个主题的混合，令 $p_j^{(2)}(y)$ 表示第二个视图中主题 j 的概率，则有：

$$p(\mathbf{y}\,|\,\mathbf{z})=\sum_{j=1}^{t}z_j\,p_j^{(2)}(y) \tag{11}$$

结合我们的目标(8)，有如下推导：

$$\begin{aligned}\mathrm{E}\,[\mathbf{X}\mathbf{Y}^{\top}]&=\int\!\!\int \boldsymbol{x}\boldsymbol{y}^{\top}p(x,y)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}\\&=\int\!\!\int \boldsymbol{x}\boldsymbol{y}^{\top}\sum_{w,z}p(\boldsymbol{x}\,|\,\boldsymbol{w})\,p(\boldsymbol{y}\,|\,\boldsymbol{z})\,p(\boldsymbol{w},\boldsymbol{z})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}\\&=\sum_{i,j}\mathrm{E}w_i z_j\int \boldsymbol{x}p_i^{(1)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\int \boldsymbol{y}^{\top}p_j^{(2)}(\boldsymbol{y})\mathrm{d}\boldsymbol{y}\end{aligned} \tag{12}$$

令 $P\in\mathbb{R}^{p\times s}$ 为一矩阵，其第 i 列 $P[:,i]=\int xp_i^{(1)}(x)dx$ ； $Q\in R^{q\times t}$ 为一矩阵，其第 j 列 $Q[:,j]=\int yp_j^{(2)}(y)dy$ ； $A\in R^{s\times t}$ ，其中元素 $A_{ij}=Ew_i z_j$ 。这样，我们就能改写式(12)为：

$$\mathrm{E}[xy^{\top}]=\mathbf{PAQ}^{\top} \tag{13}$$

如果说 x 、 y 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主题之间的相关性，也就是 $(w;z)$ 之间的相关性，如图 12 所示。

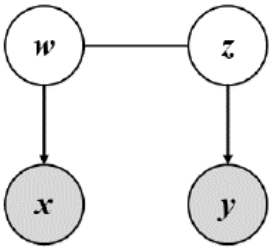

图 12 相关性上推

我们模型的工作流程可以以下图表示：

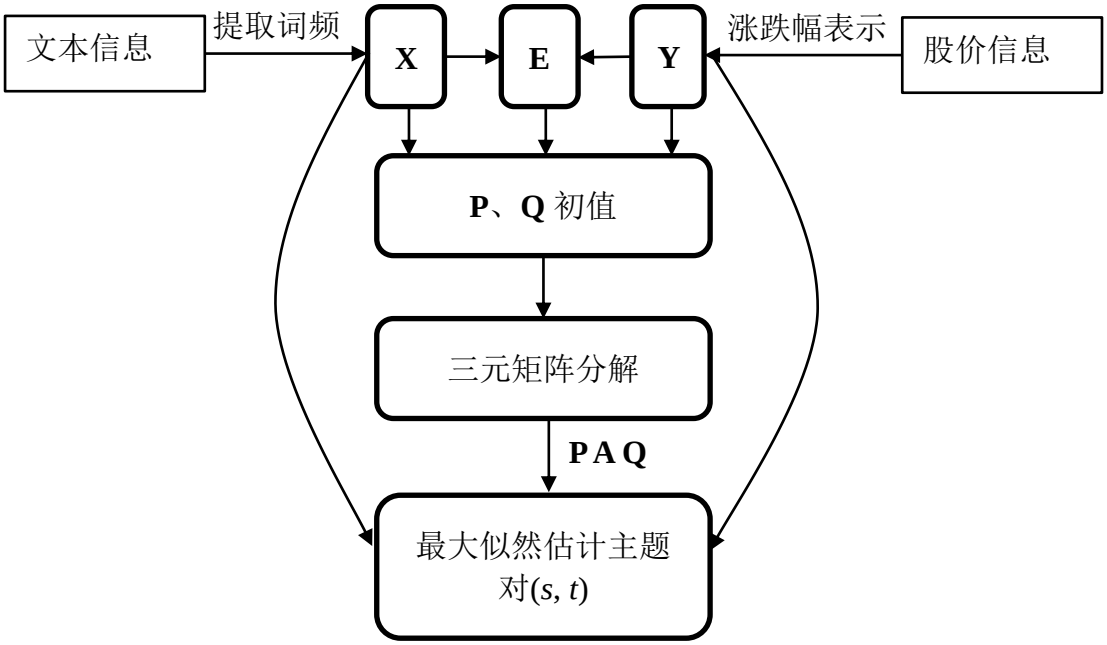

图 13 用 HTM 进行股票聚类的工作流程

4.2. 三元分解求解的局部最优解算法

对二元的非负矩阵分解，如 0 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：

$$\begin{aligned} x_i \geq 0, \sum_i x_i = 1 \\ y_i \geq 0, \sum_i y_i = 1 \end{aligned} \tag{14}$$

亦即 x、y 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：

$$\begin{aligned} &\mathbf{P}, \mathbf{A}, \mathbf{Q} \geq 0 \\ &\left\| \mathbf{P}[:, i] \right\|_1 = 1, 1 \leq i \leq s \\ &\left\| \mathbf{Q}[:, i] \right\|_1 = 1, 1 \leq i \leq t \\ &\left\| \mathbf{A} \right\|_1 = 1 \end{aligned} \tag{15}$$

考虑我们的目标函数(9)，对于 L_0，我们使用最为常用的最小二乘法。首先

不考虑 R_0，我们的目标函数如下：

$$\arg\min_{\mathbf{P},\mathbf{A},\mathbf{Q}} \sum_{i,j} (\mathbf{P}_i^\top \mathbf{A}\mathbf{Q}_j - \hat{E}_{\mathrm{i,j}})^2 \tag{16}$$

其中 $\hat{E}$ 是根据输入的文档所做出的对 $\mathrm{E}[XY^\top]$ 的估计：

$$\hat{E} = \frac{1}{n}\sum_i \mathbf{x}^{(i)}\mathbf{y}^{(i)\top} \tag{17}$$

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(16)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[37] 。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。实际求解时，我们可以将(15)中的三个等式约束代入待求式。设待求的矩阵为 $\mathbf{X}_{m\times n}$，我们将其转换为一个向量 $\vec{x}$，其中 $\mathbf{X}_{ij} = x_{i+m(j-1)}$。这样，目标函数 $\sum_{i,j}(\mathbf{P}_i^\top \mathbf{A}\mathbf{Q}_j - \hat{E}_{\mathrm{i,j}})^2 = \left\|\mathbf{P}\mathbf{A}\mathbf{Q}^\top - \hat{E}_{\mathrm{i,j}}\right\|_2$ 可以表示为如下形式：

$$\left\|\mathbf{C}\vec{x} - \vec{d}\right\|_2 \tag{18}$$

一般地，对于矩阵乘积 $C = A_{m\times n}B_{n\times r}$，我们有 $C_{ij} = \sum_k A_{ik}B_{kj}$，因而如果向量化矩阵 A，其在式(18)中对应的矩阵 C 就为：$\mathbf{C}_{<i,\ j>,\ <i,\ k>} = B$，其中 $<i,\ j> = i + (j-1)$，$<i,k> = i + (k-1)m$。对于 **P** 矩阵，我们可以得到这样的矩阵 **C**，并且更改其中对应于 $<p,\ j>, <i,\ k>$ 的值为 $-B_{kj}$，对应的 d 向量的值为 $D_{pj} - \sum_k B_{k,:}$。求解矩阵 **Q** 时类似。对于矩阵 **A**，只需考虑其对应向量的最后一个分量。可以得到，在 x 对应于 **A** 时，有：

$$\begin{aligned} & B\vec{x} - d \\ &= B_{:,1:n-1}\vec{x}_{1:n-1} + B_{:,n} - B_{:,n}[1...1]\vec{x}_{1:n-1} - d \\ &= (B_{:,1:n-1} - B_{:,n}[1...1])\vec{x}_{1:n-1} - (d - B_{:,n}) \end{aligned} \tag{19}$$

这样，我们就把原问题转换为非负限制下的最小二乘法问题。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：

[37] Lawson C L, Hanson R J. Solving least squares problems[M]. Englewood Cliffs, NJ: Prentice-hall, 1974.

$$\begin{aligned}&\hat{E}=PAQ^{\top},P^{*}\hat{E}=P^{*}PAQ^{\top}=AQ^{\top}\\&P^{*}\hat{E}Q^{*\top}=AQ^{\top}Q^{*\top}=A\\&(PA)^{*}\hat{E}=Q^{\top}\\&\hat{E}(AQ^{\top})^{*}=P\end{aligned}\tag{20}$$

其中 $(\bullet)^{*}$ 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 L_1：

$$L_1=I(\mathbf{A})=\sum_{w,z}p(w,z)\log\frac{p(w|z)}{p(z)}\tag{21}$$

其中 $p(w,z)=A_{wz}$。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 L_1 的一个极大值。为了不影响 **P**、**Q**，这时 **A** 的约束条件除了式(15)给出的之外，还要保持 $p(z)=\sum_{w}A_{wz},p(w)=\sum_{z}A_{wz}$ 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主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。

4.3. 三元分解求解的线性规划算法

在 4.2 节中，我们已经给出了三元分解求解的局部最优解算法。但在随后的各项实验中可以看出，该解法对初始值过分敏感，以至于初始化的方法会极大地影响实验结果。为了得到一种对初始值更加稳定的算法，本节中我们将对问题进行必要的等价转化，使其可以运用线性规划（Linear Programming）的方法进行求解。

我们已经明确了求解问题的约束条件，亦即 **P**、**A**、**Q** 三个矩阵的限制条件。

为了能够运用线性规划进行问题求解，我们需要对矩阵进行必要的向量化。我们用 vec(**X**)表示对 **X** = $(x_{ij})_{n\times m}$ 矩阵的向量化操作，其结果是一个向量 **v**，其中 $v_i = x_{im+j}$。下文中，我们记 **p** = vec(**P**)，**a** = vec(**A**)，**q** = vec(**Q**)，**e** = vec(**E**)。

对于我们的目标函数，由于其意义在于使得残差尽可能小，同时我们已经不再需要考虑求导问题，我们可以方便地使用 1-norm 来表示这一残差的大小。亦即我们的目标是：

$$\arg\min_{\mathbf{P},\mathbf{A},\mathbf{Q}} \|E - PAQ\|_1 = \arg\min_{\mathbf{P},\mathbf{A},\mathbf{Q}} \sum_{ij} |E_{ij} - (PAQ)_{ij}| \tag{22}$$

这可以转化为如下的形式：

最小化：$\sum_{ij} T_{ij}$

其中限制条件为：$-T_{ij} \le E_{ij} - (PAQ^T)_{ij} \le T_{ij}$，$T_{ij} \ge 0$。

很显然我们也可以将此处定义的 T_{ij} 向量化为 $\mathbf{t} = \mathrm{vec}(T)$。现在让我们来检查 $(PAQ)_{ij}$，容易得到：$(\mathrm{PAQ}^T)_{ij} = \sum_k \sum_l P_{ik} A_{kl} Q_{jl}$。将上式用向量的形式表示，可以转化为：

最小化：$\sum_i \mathbf{t}_i$

限制条件：$-\mathbf{t}_{iq+j} \le \mathbf{e}_{iq+j} - \sum_{kl} \mathbf{p}_{is+k} \mathbf{a}_{kt+l} \mathbf{q}_{jt+l} \le \mathbf{t}_{iq+j}$，$\mathbf{t}_i \ge 0$。

而关于 **P**、**A**、**Q** 三矩阵的限制条件可以表示为：

$$\mathbf{p}, \mathbf{a}, \mathbf{q} \ge 0$$

$$\sum_{j=(i-1)p+1}^{ip} \mathbf{p}_j = 1, 1 \le i \le s$$

$$\sum_{j=(i-1)q+1}^{iq} \mathbf{q}_j = 1, 1 \le i \le t$$

$$\sum_i \mathbf{a}_i = 1$$

在使用现行求解的过程中我们仍然用顺次迭代的形式进行，即仍先对 **a** 求解、再 **q**、再 **p** 的顺序进行。每次迭代中未知的向量与 **t** 共同组成了未知向量 **x**。这样，每次我们求解的线性规划问题中，涉及的未知向量实际上也就只有一个。我们所需要求解的问题规模最大为 $\max\{ps, qt\}+pq$。对于我们的数据，这一规模达到了 10^{12} 的数量级，这使得通常的线性规划求解方法无效。为此，我们需要对原问题再进行一定的变换。

考虑求解矩阵 **P** 的过程。此时，矩阵 **A**、**Q** 均视作常量。矩阵 **E** 是常量。对于目前的最优解 **P***，令 t(**P**) = $\|\mathbf{E} - \mathbf{PAQ}^T\|_1$，显然函数 t 在 **P*** 处取得最小值。

现在我们考虑矩阵 **B***，是矩阵 $\mathbf{AQ}^{\mathrm{T}}$ 的伪逆。令函数 t'(P) = || **EB*** - $\mathbf{PAQ}^{\mathrm{T}}$**B*** $\|_1$ = || **EB*** - **P** $\|_1$。显然函数 t'在 **P***处也取得最小值。这样，上述的求解过程中，我们关于 **t** 的限制条件就转变为 $-\mathbf{t}_{is+j}\mathbf{B}^*_{ij} \le (\mathbf{EB}^*)_{ij} - \mathbf{p}_{is+j} \le \mathbf{t}_{is+j}\mathbf{B}^*_{ij}$，隐变量 **t** 的维数便从原先的 *pq* 降低到了 *sp*。类似地，我们也就可以将求解 **Q** 时的隐变量维数降低到 *tq*，求解 **A** 时的维数降低到 *st*。

进一步，通过引入随机投影（Random Projection），可以使维度降得更低。

4.4. 该模型的验证

4.4.1. 评价方法

参照 Gu 等人的文献[38]，我们选择了准确率和 NMI 对聚类的准确率进行评价。设 l 表示这些数据实际所具有的标号，取值范围为 1 到 n 的正整数。用 l' 是聚类得到的聚类簇标号，取值范围为 1 到 m。设 ϕ 是一个从 $1,2,...,m$ 到 $1,2,...,n$ 的映射，定义下式为聚类错误率：

$$ClusteringError = \min_{\phi} \frac{\#\{j \mid l_j \neq \phi(l'_j)\}}{\#\{\text{total objects}\}} \tag{23}$$

并称 $1-ClusteringError$ 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 ϕ 以得到准确率。用简单的贪心方法就可以得到这样的映射 ϕ。

- 建立一个记数矩阵 $C=(c_{ij})_{n\times m}$，其中 c_{ij} =#{在 l 中被标为类 i 而在 l' 中被标为 j 的文档}。令 $k=1$。
- 选择 C_k 中最大值的下标 y，并令 $\phi(k)=y$。
- 将记数矩阵中 C'_y 和 C_k 设为 0，并令 k 自增 1。
- 若 $k \le m$，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号的细分，反

[38] Gu Q, Zhou J. Local learning regularized nonnegative matrix factorization[C]//Twenty-First International Joint Conference on Artificial Intelligence. 2009.

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号，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：

$$RecallRatio = \max_{\psi} \frac{\#\{j \mid l'_j = \psi(l_j)\}}{\#\{\text{total objects}\}} \tag{24}$$

其中，当 $m \geq n$ 时，ψ 是一个从 $1,2,...,n$ 到 $1,2,...,m$ （的一个子集）的双射；对于 $m < n$ 的情况，我们不妨定义这时的 $RecallRatio(l,l')$ 为 $RecallRatio(l',l)$。此时，我们着眼于在 l 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 ψ 的过程与贪心法求 ϕ 的方法相似，只是在第三步中不将 C_k 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：

$$F = \frac{2(1-ClusteringError)RecallRatio}{1-ClusteringError+RecallRatio} \tag{25}$$

为了评价聚类的有效性，我们还引入了 NMI（Normalized Mutual Information）。[39] 其定义如下：

$$NMI = \frac{\sum_{k=1}^{C_1}\sum_{m=1}^{C_2} n_{k,m}\log\frac{nn_{k,m}}{n_k\hat{n}_m}}{\sqrt{(\sum_{k=1}^{C_1} n_k\log\frac{n_k}{n})(\sum_{m=1}^{C_2}\hat{n}_m\log\frac{\hat{n}_m}{n})}} \tag{26}$$

其中 n_k 表示聚类簇 $D_k(1 \leq k \leq C_1)$ 中文档的数量，n_m 表示从属于 $L_m(1 \leq m \leq C_2)$ 的文档数量，$n_{k,m}$ 表示 D_k 与 L_m 的交集中文档的数量。NMI 越大（接近于 1），则聚类的结果越好。

4.4.2. 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 A 中给定的联合分布，选择一对主题 (i, j)。
- 选择文档的长度 W_d，在此被设定为一个常数，因为我们的模型只与文档中

[39] Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics, 2009, 11(3): 033015.

词的频率有关，与频数无关。

- 依据 P_i 给出的多项式分布，在第一个视图下产生 W_d 个词。
- 依据 Q_j 给出的多项式分布，在第二个视图下产生 W_d 个词。

因而，我们需要事先准备好 **P**、**Q** 和 **A**。我们以 s 记第一个视图下主题的数量，t 为第二个，W_t 为每个主题中的词数。考虑到大量的词，其出现与主题的关系不大，以 W_u 记这些词的数量。这些词在所有的主题中都以相同的分布出现。我们同时定义一个参数 p_u，表示了与主题有关的那些词与主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 P 的生成方式为例，对于每个主题 $i=1,2,..s$，将 P_i 设为一个行向量，其中下标 $(i-1)*W_t+$ 到 $i*W_t$ 设为 $1/(W_t+p_u*W_u)$，最后 W_u 维设为 $p_u/(W_t+p_u*W_u)$，其它维的值均置零。Q 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 P 的生成为例，首先我们生成一个矩阵 $\hat{\mathbf{P}}$，其中的每列以如下方式生成：对于下标 $(i-1)*W_t+1$ 到 $i*W_t$（从 1 开始）的元素，我们从 $U[p_u,1)$ 分布中随机取一个数值；对于最后的 W_u 个元素，我们从 $U[0,p_u)$ 分布中随机取一个数值。通过将 $\hat{\mathbf{P}}$ 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 $\eta=1-\frac{1}{mu}=1-\frac{1}{st}$，$\sigma_m=\sqrt{\frac{1}{st}\left(\frac{st-1}{(st)^2}+\left(\frac{1}{st}-1\right)^2\right)}\approx\sqrt{1/st}$。我们选取 Beta 分布的参数 (α,β)，以使得其期望为 $\frac{1}{st}$，而标准差为 $\sigma=c\sigma_m$，其中 $c\in[0,1)$。解方程，可以得到：

$$\alpha=\frac{\frac{\eta}{\sigma^2(1+\eta^2)}-1}{1+\eta}$$
$$\beta=\eta\alpha(c\neq 0) \tag{27}$$

对于 $c=0$，直接生成一个元素全为 $\frac{1}{st}$ 的矩阵即可。

显然，当 c 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：

$$I(X;Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{28}$$

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 c 的函数。其图像如下所示。（c 以 0.1 的步长变化）

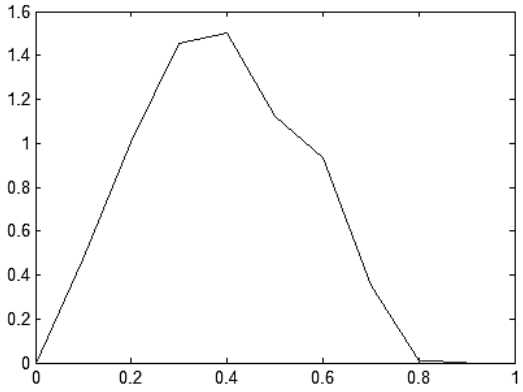

图 14 c 的取值与生成的主题联合分布的互信息熵变化关系图

4.4.3. 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 L_1 的结果。可见对各种初始化方法均有小幅度提高。

表 11 TOY1 中运用二级优化函数后的效果提高

	PLSA	RP[40]	K-means	两次分解[41]
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

[40] 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 E, Mannila H. Random projection in dimensionality reduction: applications to image and text data[C]//Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2001: 245-250.

[41] 见 4.2 节。

F 值提高 在所有测试用例中占比	43%	48%	50%	43%

4.4.4. TOY2 上的实验结果

TOY2 中我们涉及到的参数有无关词数量W_u和c。

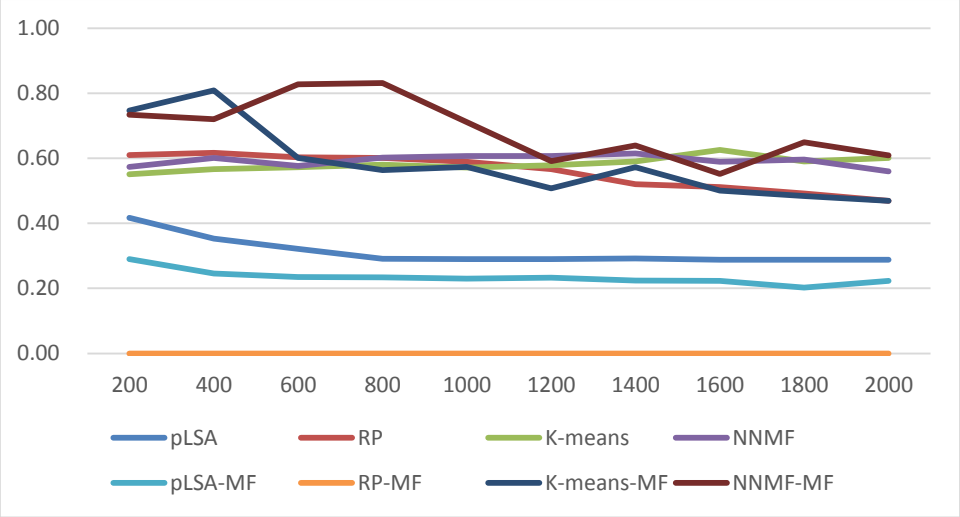

图 15 不同方法聚类结果 NMI 随W_u的变化

图上，标有“-MF”的表示采用某种方法初始化两个视图，然后进行基于矩阵分解的异构主题模型。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W_u增加而效果下降的趋势比原方法更好。

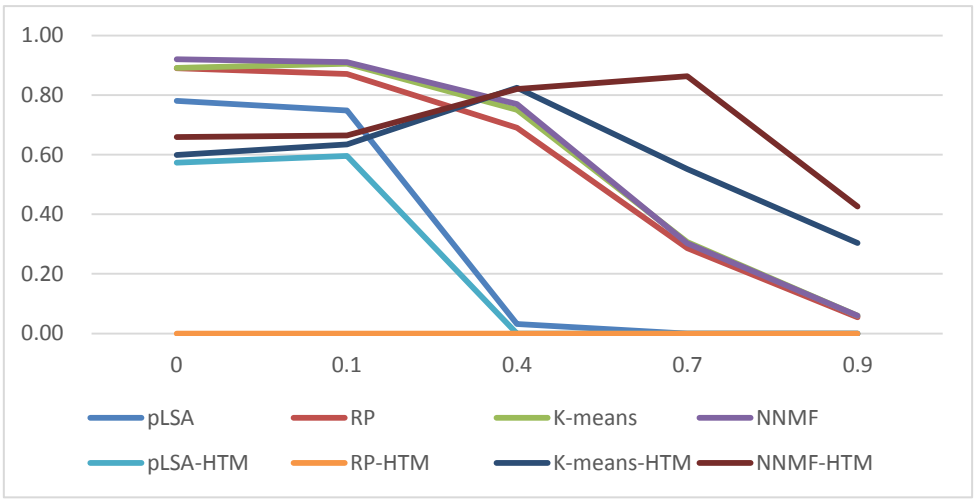

图 16 不同方法聚类结果 NMI 随c的变化

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 NMF 两种初始化方法，所得到的结果，在C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$(w;z)$共现分布。

4.4.5. 在 Stock50 上的验证

先前的研究都只能处理同一种类型的数据，如股价数据或文本数据等，而不能将两者结合起来进行考虑。简单的数据表示上的拼装（例如，将文本和股价都表示为一个向量，然后将二者简单连接起来组成一个新的特征向量，参与到聚类过程中去）将大大增加数据的维度，从而进一步加剧“维数灾难”问题，得不到更好的结果。这是由于两种数据的异质特征导致的。为说明这一问题，我们可以用文本＋股价直接拼合成一个长向量来表示一只股票，组成一个新的特征向量，并进行数据分析。得到的结果如下。

表 12 在 Stock50 上的验证

	NMI
LDA（长向量）	0.4972
Kmeans（长向量）	0.7082
Link-pLSA-LDA	0.4916
我们的方法（双视图 LDA 初始化）	**0.7511**

与直觉相一致，更长的向量带来了聚类质量下降（NMI 下降）的问题。同时，这一结果也提示了 Link-pLSA-LDA 在股价方面具有较低的实用价值。

第5章 模型的应用

基于之前几章的工作，我们已经建立了一个基于非负矩阵分解的异质话题模型。在本章中，我们将把这一模型放置于实际的金融数据场合中，以考察其有效性。我们所使用的主要数据是上市公司的经营范围描述信息（以下简称 BSD）和股价数据（每日收盘价，以下简称 PD）。参照的专家分类结果，是根据上市公司信息中的所在行业和板块来确定的。

5.1. 基于股价与经营范围描述的上市公司聚类

Stock2209 是从中国 A 股市场上获得的股票数据。一个视图是经营范围描述信息（BSD），另一个视图是每日股价的涨跌幅表示。BSD 信息经过分词、标注、去除停用词和罕见词之后得到其词袋假设下的词频向量。在选择每日股价时，参考了上证综指（如图 17）的走势。第一个时间段（记为 T1）从 2012 年 5 月 4 日到 2012 年 9 月 26 日为止，这一段时间内上证综指先持续两个月保持下跌趋势（如图中的“1”所示）。此后出现两周的上扬，然后下跌一个月，最后出现了明显的上扬。这三段（图中的 2～4）组成了第二个时间段（记为 T2）。
Stock50 是该数据集的一个子集。

5.2. 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 17 上证综指在 2012 年 5 月 4 日至 2012 年 12 月 31 日之间的走势

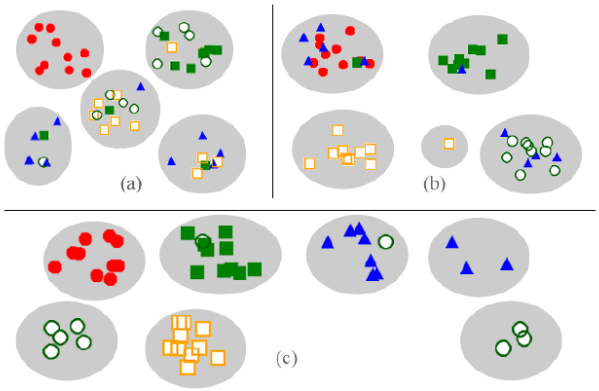

图 18 Stock50 上聚类簇示意图

（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）为我们的异构主题模型聚类的结果，两个视图分别以 LDA 和 K-means 初始化。实心圆：银行；实心正方形：钢铁；三角形：制药；空心正方形：酒；空心圆：信息技术。

根据半强式有效市场假设（semi-strong-form efficient market hypothesis，EMH）[42]，股价最终反映了关于该股价的所有信息。然而需要注意的是，股价同时受到时间因素的影响，因此我们仅仅通过一时的股价数据不一定能够得到有意义的结果。为此，我们通过检查分类的预测情况来检验模型的有效性。

相应地，我们利用皮尔森系数（Pearson correlation coefficient，PCC）[43] 以反映一段时间内一对股票价格波动的相关性，并检验其显著程度（p 值）。为评价一个聚类簇中股价的相关程度，我们以所有不同股票对的皮尔森系数之平均值及其对应 p 值的平均值来评价。

Stock50 的可视化结果如图 18 所示。其中包含了由专家指定的五个行业（话题簇）。图 18（a）中显示了 LDA 在 BSD 单个视图上的聚类结果，（b）显示了使用 K-means 在价格数据视图上的聚类结果。（c）是运用我们的异构话题模型得到的聚类结果，同时利用了 BSD 和价格信息。容易看出，我们提出的方法得到了更好的聚类结果，因为它同时考虑了所属行业和相似的股价趋势。如果只有一个视图，例如 BSD，用来生成分类信息，那么所产生的聚类簇将不能反映股价的上涨下跌趋势行为，因为 BSD 通常对一上市公司而言是不会变动的。我们以其中一个聚类为例，观察它的股价走势图：

[42] Malkiel B G, Fama E F. Efficient capital markets: a review of theory and empirical work [J]. The Journal of Finance, 1970, 25(2): 383-417.
[43] Stigler S M. Francis Galton's account of the invention of correlation[J]. Statistical Science, 1989, 4(2): 73-79.

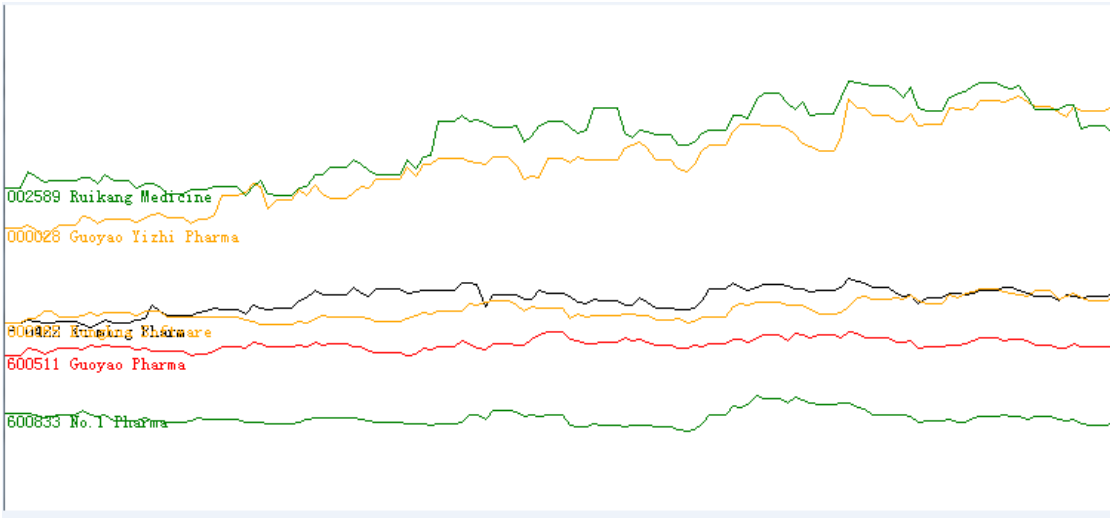

图 19 Stock50 中一个聚类簇的股价走势

可以看出，它们的涨跌保持相对的一致性，这种一致性并不是统计上的多数，而是在一些时间段上的完全同步。

如果同时利用 BSD 和股价数据来定义上市公司的标签，我们将不仅容易将相似产业的上市公司归于一类，同时在这一类中，这些上市公司的股价表现也会显得相似。这时，我们就可以利用聚类信息来反映“板块轮动”：一旦其中的某支股票呈现出剧烈的上涨或下跌趋势，在同一聚类簇中的其它股票就可能出现相似的波动。

5.3. 在海外证券市场的运行结果及分析

下面的结果中，训练时所采用的时间段为 2010 年 1 月 1 日至 2012 年 12 月 31 日。

表 13 在四个海外证券市场上的分析结果

NYSE	行业分类	LDA-BSD	LDA-PD	KM-BSD	KM-PD	LDA-MF	KM-MF	NMF-MF	LDA-KM-MF
ACC		0.677	0.581	0.431	0.431	0.502	0.484	0.500	0.502
NMI		0.434	0.326	0.221	0.038	0.355	0.293	0.306	0.355
COR	0.433	0.438	0.539	0.346	**0.552**	0.433	0.388	0.537	0.433
STD.	0.230	0.175	0.281	0.147	0.206	0.199	0.245	0.251	0.199
LSE									

ACC		0.210	0.103	0.147	0.147	0.141	0.103	0.148	0.141
NMI		0.238	0.122	0.180	0.128	0.187	0.127	0.197	0.187
COR	0.101	0.105	0.234	0.102	**0.519**	0.249	0.320	0.310	0.249
STD.	0.159	0.081	0.184	0.060	0.253	0.231	0.267	0.263	0.231
ASX									
ACC		0.621	0.621	0.535	0.535	0.448	0.466	0.490	0.448
NMI		0.453	0.094	0.346	0.100	0.221	0.303	0.296	0.221
COR	0.203	0.204	0.285	0.206	**0.495**	0.243	0.267	0.244	0.243
STD.	0.140	0.090	0.150	0.161	0.292	0.188	0.194	0.152	0.188
SGX									
ACC		0.495	0.495	0.411	0.411	0.467	0.439	0.448	0.467
NMI		0.276	0.078	0.153	0.127	0.254	0.215	0.214	0.254
COR	0.150	0.281	0.334	0.184	0.321	**0.344**	0.215	0.310	**0.344**
STD.	0.247	0.180	0.259	0.084	0.151	0.249	0.107	0.238	0.249

因为最后我们用股价数据的相关性来衡量结果的优劣，所以单独用股价进行聚类，从相关度上说，理论上应该是最好的；正如在经营范围描述（BSD）上的聚类方法都能得到较高的 ACC 值。但是这提示“过拟合”的可能性，也就是说，仅仅基于股票价格（PD）数据进行的聚类，只能够对训练样本这一时间段负责，而可能缺乏良好的预测能力。这同时也就意味着，单单在股价上聚类，得到的聚类簇随着时间的变化是不稳定的。下列的实验说明了这一点。

以下结果中，训练时采用的时间段为 2010 年 1 月 1 日至 2011 年 12 月 31 日。测试时使用的时间段为 2012 年 1 月 1 日至 2012 年 12 月 31 日。我们保留两个只使用 BSD 信息的方法，同专家的标签一样用作参考。

表 14 不同聚类方法的预测效果。表中数值为测试时间段中的平均皮尔森系数。

	行业分类	LDA-BSD	LDA-PD	KM-BSD	KM-PD	LDA-MF	KM-MF	NMF-MF	LDA-KM-MF
NYSE	0.433	0.438	0.372	0.346	0.169	0.433	0.448	0.429	**0.474**
LSE	0.101	0.105	0.204	0.102	0.133	0.249	0.193	**0.304**	0.287
AS	0.203	0.204	0.157	0.206	0.14	0.243	**0.267**	0.248	0.246

X					3				
SG	0.150	0.281	0.172	0.184	0.26	**0.344**	0.181	0.284	0.220
X					8				

可以看到，加入 BSD 信息之后的聚类结果，在相关度上都不输于在单一视图上进行聚类的结果。

5.4. 在大规模数据集（Stock2209）上短时段的运行结果及分析

为了检验我们的主题模型产生的聚类簇的有效性，我们计算了不同窗口大小下的平均皮尔森系数，亦即对于同一聚类簇中的一对股票，我们对其股价数据先进行加窗处理，窗长分别为 3、5、7……19。此处的平均皮尔森系数，是对每对股票加窗后皮尔森系数的平均值，再进行平均得到的。图 20 给出了不同方法得到的聚类结果中每类的平均皮尔森系数随加窗大小的变化曲线。

表 15 Stock2209 上不同方法利用单一数据分类得到的聚类结果的平均皮尔森系数（aPCC）和平均 p 值。T1 和 T2 的结果表示用 T1 或 T2 的数据训练并在相同数据上进行测试。T1+T2 为使用 T1 的数据训练而在 T1+T2 的数据上进行测试的结果，反映了聚类的预测能力，亦即对时间变化的稳定性。

	T1		**T2**		**T1+T2**	
方法	**aPCC**	**p 值**	**aPCC**	**p 值**	**aPCC**	**p 值**
行业分类	0.3481	0.0521	0.4892	0.0548	0.3934	0.0428
LDA（BSD）	0.3288	0.0533	0.4658	0.0599	0.3723	0.0437
Kmeans（价格）	0.3410	0.0452	0.5360	0.0473	0.4008	0.0371
NMF*（BSD）	0.3702	0.0458	0.5298	0.0478	0.4218	0.0378

注*：NMF 表示用两元的非负矩阵分解作初始化，NMF 在单个视图上的聚类结果是经过NMF对原数据先降维再Kmeans聚类得到的，NMF-HTM是用NMF得到 P 和 Q 矩阵的初值。在第一个视图是 BSD，第二个视图是股价的情况下，就是将 BSD 的数据矩阵分解成 $\mathbf{PR}_1$，price 的分解成 $\mathbf{QR}_2$，其中 **P**、**Q** 是两个视图中主题-词的概率分布，$\mathbf{R}_1$、$\mathbf{R}_2$ 是 NMF 分解余下来的矩阵，可以看作原数据中各文档的主题的组成。

表 16 Stock2209 上不同初始化方法运用异构话题模型所得到之聚类结果的平均皮尔森系数和平均 p 值。与表 15 对比可发现本文的方法具有更好的预测能力。

初始化方法		T1		T2		T1+T2	
BSD	**价格**	**aPCC**	**p 值**	**aPCC**	**p 值**	**aPCC**	**p 值**
LDA	LDA	0.5418	0.0349	0.6394	0.0404	0.5794	0.0285
LDA	Kmeans	0.5448	0.0340	**0.6523**	**0.0442**	**0.5876**	**0.0249**
LDA	NMF	**0.5555**	**0.0288**	0.6072	0.0509	0.5626	0.0334
Kmeans	LDA	0.2889	0.0526	0.4358	0.0637	0.3407	0.0455
Kmeans	Kmeans	0.5788	0.0320	0.5836	0.0602	0.5720	0.0242
Kmeans	NMF	0.4192	0.0386	0.4301	0.0664	0.4006	0.0404
NMF	LDA	0.4496	0.0652	0.5823	0.0458	0.4810	0.0796
NMF	Kmeans	0.4962	0.0273	0.5365	0.0508	0.5114	0.0303
NMF	NMF	0.5471	0.0320	0.6131	0.0416	0.5577	0.0258

除了 LDA 的结果，利用第一时间段 T1 的单一视图数据，用不同的方法训练得到的聚类结果中，同一个聚类簇中的线性相关性稍优于那些由专家标注的标签。同样，预测的结果的平均皮尔森系数（aPCC）和 p 值在第二时间段（T2）和整个时间段（T1+ T2）上都要优于专家。我们所提出的方法的平均皮尔森系数和 p 值如表 16。第一段时间 T1 内的股票数据与 BSD 文本都被用于训练我们的异构主题模型。通过在模型中纳入价格数据的相关性，股票在同一聚类簇中的相关性，比那些由专家和由一个单一的视图数据的聚类簇更好，拥有更高的平均皮尔森相关系数及更低的 p 值。例如，两个视图均用 K-means 初始化得到的结果中，aPCCs 和 p 值分别为 0.5788 和 0.0320，远远好于专家的 0.3481 和 0.0521，其中低 p 值是指在统计上两个变量之间关系的显著性。

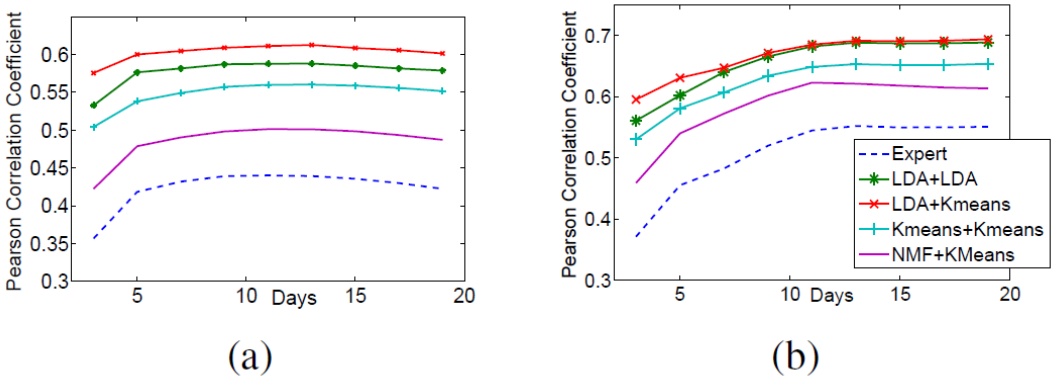

图 20 不同方法得到的聚类结果中每类的平均皮尔森系数随加窗大小的变化曲线

两视图的主题模型可以用不同的方法以初始化矩阵 **P** 和 **Q**，如 LDA+LDA，LDA+Kmeans，Kmeans+Kmeans，NMF+Kmeans。图 20（a）表示时间周期 T1 的结果，而图 20（b）是在时间周期 T2 上的预测结果。结果显示，我们的方法具有比专家更好的预测能力。

5.5. 在大规模数据集（Stock2209）上长时段的运行结果及分析

为进一步检验双视图异构主题模型聚类结果的有效性和对于时间的有效性，我们选择了三个时间段作考察。这三个时间段分别是：

表 17 Stock2209 长时段数据集参数

编号	起始日期（含）	结束日期（含）
P1	2005 年 6 月 6 日	2007 年 10 月 16 日
P2	2007 年 10 月 16 日	2008 年 10 月 28 日
P3	2008 年 10 月 28 日	2012 年 6 月 29 日

我们将以此结果与 Corr-LDA 进行比较分析。这三个时间段的上证指数如下图所示。

图 21 2005 年 6 月 6 日～2012 年 6 月 29 日上证指数走势图

我们用和上述相似的方法，对不同时段的数据进行分析，并对股票进行聚类。其结果如下。可以看到，我们的方法在训练时间段上得出的结果是次好的，但在后一时间段上的预测表现则好于其它方法。这也正是因为我们的方法不只考虑了股票价格的信息，还将其经营范围描述亦即行业信息考虑了进去。

表 18 在长时段上的聚类评价结果

	P1 训练，预测 P2		P2 训练，预测 P3	
方法	**相关度**	**预测相关度**	**相关度**	**预测相关度**
Corr-LDA	0.3087	0.6584	0.7078	0.4665
LDA（BSD）	0.2951	0.6351	0.6351	0.4554
LDA-BSD 自相关*	0.1152	0.3087	0.3087	0.4742
KM（股价）	**0.4322**	0.7683	0.6381	0.4447
KM-股价自相关*	0.3152	0.7209	0.6170	0.4711
LDA-KM-HTM	0.4074	**0.8481**	0.7341	0.4942
NMF（BSD）	0.3847	0.6363	0.6363	0.5114
NMF（股价）	0.3835	0.7614	0.7523	0.4719
NMF-HTM	0.2367	0.8029	**0.9098**	**0.5187**

注*：“自相关”是指，采用 HTM 方法，但令两个视图均为相同的 BSD 或股价信息，要求 **P** = **Q**，进行近似的矩阵分解后聚类。另外，使用 BSD 作为唯一特征的方法，由于不受训练时间段影响，因此“P1 训练时 P2 预测”的结果和“P2 训练时”的结果相同。“自相关”方法解释了主题的数量，例如对于 P1、P2 时段股价上的 Kmeans-自相关，我们得到如下的 **A** 矩阵：

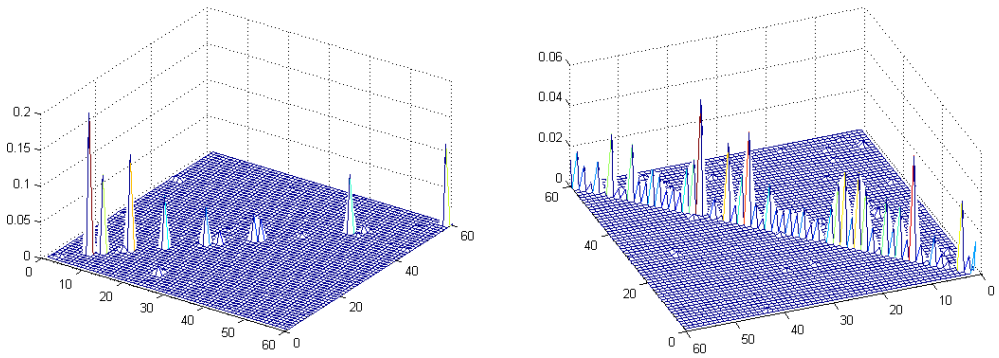

图 22 Kmeans（股价）-自相关的 **A** 矩阵，左：P1 时段股价；右：P2 时段股价

可见，在股价视图上，在 P1 时段发挥作用的只有 8 个主题，而在 P2 时段则所有 60 个主题都发挥了作用。

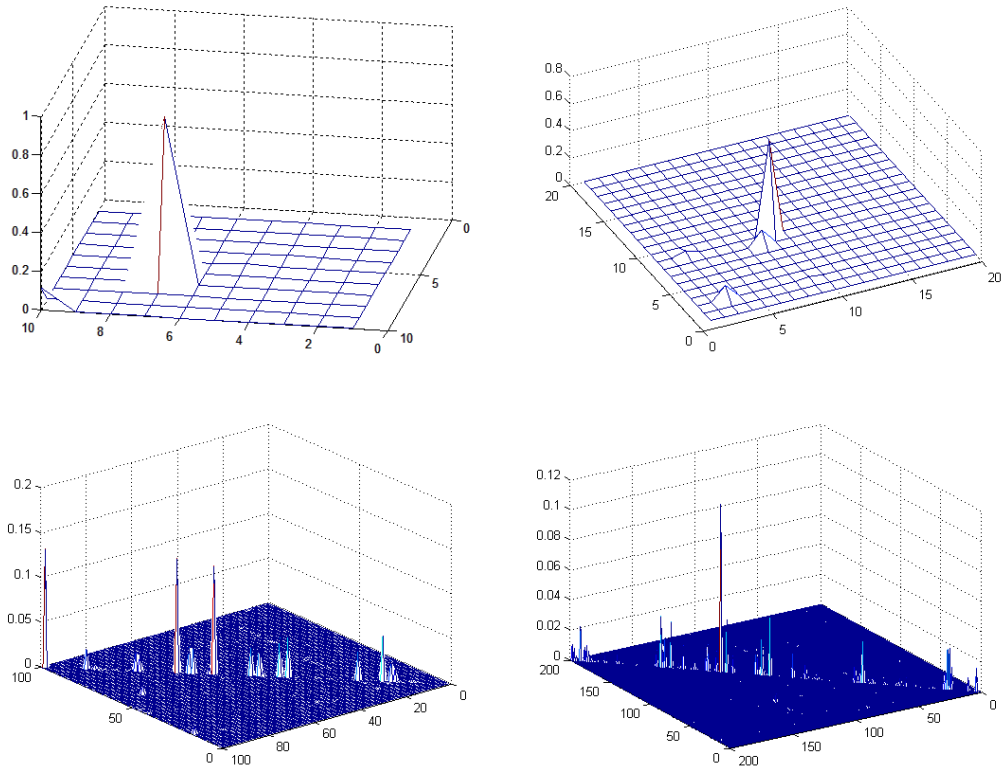

图 23 不同主题数量下 P1 时间 Kmeans（股价）-自相关所得 **A** 矩阵。左上：主题数量为 10，右上 20，左下 100，右下 200。

通过这一方法，我们能够选择较好的主题数量，以尽量减少图中的“毛刺”，在速度和质量之间得到平衡，同时也避免拟合程度过度或不够（over-fitting and under-fitting）。

第6章 讨论与总结

在本文中，研究了异构的双视图主题模型，以利用从不同来源得到的、不同物理意义和特征的数据进行聚类。本文从矩阵分解的角度出发，将对双视图数据的建模和参数求解过程转换为带有线性约束条件的非负矩阵的三元分解问题。双视图下各自由主题到文档的生成过程被认为是独立的，从而将（不同的视图）数据矩阵之间的相关性归结于各自主题之间的联系。由此分解得到三个矩阵 **P**、**A**、**Q**，其中 **P** 表示在一个视图中“词”在某一主题的中出现的比例（概率），**Q** 表示“词”在另一视图下某一主题中的比例。此外，在两个矩阵的元素满足概率分布的约束。矩阵 **A** 为两个视图的主题之间的联合概率。

因此，对我们所提出的方法而言，没有必要知道概率模型 $\mathbf{P}_i(\mathbf{x})$和 $\mathbf{Q}_j(\mathbf{y})$的具体形式。换句话说，该方法适用于任意的模型而不需要知道（事先规定、假设）其所符合的概率分布的形式。相比较而言，贝叶斯主题模型需要知道特定的概率分布，例如 Dirichlet。我们的模型另一优点是，我们解决方案本身的复杂程度关于样本数是不变的。当样本数趋于无穷大时，如在当今的大数据时代，这一点尤其重要。这意味着我们可以在数据生成的时候线性地计算预期矩阵 **E**。而类似于 LDA 的模型是不一致的，因为在整个隐变量空间上进行精确推断是不可行的。

本文展现了本模型在中国 A 股的经营范围描述（文本）和分时段价格数据上的实验结果，并且也在海外股市上进行了类似的实验。实验结果表明：（1）在两个视图上共生的高相关主题可以通过共现概率矩阵 **A** 进行识别，（2）在同一个部门（主题聚类簇）的股票，可以得到比专家人工标注或传统基于单一视图信息的聚类方法更高的两两之间相关度。

在我们提出的模型中，初始化矩阵对主题聚类结果有明显的影响。这说明在我们的优化求解过程中存在一定的问题。作为未来的发展，应进一步调查。此外，在实际应用中，数据通常包含超过两类，如一篇论文、链接信息、作者、发表场所（期刊、会议）等。进一步研究将本模型扩展到两个以上的视图，这是另一个未来的发展路径。

参考文献

1. 《现代汉语常用词表》课题组. 现代汉语常用词表（草案）[M]. 商务印书馆, 2008.
2. 顾森. SNS 中的文本数据挖掘[J]. 程序员, 2012 (8): 113-115.
3. 胡裕树. 现代汉语[M]. 上海教育出版社, 2010.
4. Berry M W, Browne M, Langville A N, et al. Algorithms and applications for approximate nonnegative matrix factorization[J]. Computational Statistics & Data Analysis, 2007, 52(1): 155-173.
5. Bingham E, Mannila H. Random projection in dimensionality reduction: applications to image and text data[C]//Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2001: 245-250.
6. Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. the Journal of machine Learning research, 2003, 3: 993-1022.
7. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2011, 2(1): 1-8.
8. Boyd-Graber J, Blei D M. Multilingual topic models for unaligned text[C]//Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2009: 75-82.
9. Chen X, Zhou M, Carin L. The contextual focused topic model[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 96-104.
10. Deza M M, Deza E. Encyclopedia of distances[M]. Springer Berlin Heidelberg, 2009.
11. Ding C, Li T, Peng W, et al. Orthogonal nonnegative matrix t-factorizations for clustering[C]//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006: 126-135.
12. Doyle G, Elkan C. Financial topic models[C]//Working Notes of the NIPS-2009 Workshop on" Applications for Topic Models: Text and Beyond Workshop. 2009.
13. Feng H, Chen K, Deng X, et al. Accessor variety criteria for Chinese word extraction[J]. Computational Linguistics, 2004, 30(1): 75-93.
14. Gopalkrishnan V, Steier D, Lewis H, et al. BIG DATA 2.0: New business strategies from big data[J]. Deloitte Review, 2013, 12(1): 56

15. Griffiths T, Steyvers M. A probabilistic approach to semantic representation[C]//Proceedings of the 24th annual conference of the cognitive science society. 2002: 381-386.
16. Gu Q, Zhou J. Local learning regularized nonnegative matrix factorization[C]//Twenty-First International Joint Conference on Artificial Intelligence. 2009.
17. Hofmann T. Probabilistic latent semantic indexing[C]//Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999: 50-57.
18. Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics, 2009, 11(3): 033015.
19. Landauer T K, Foltz P W, Laham D. An introduction to latent semantic analysis[J]. Discourse processes, 1998, 25(2-3): 259-284.
20. Lawson C L, Hanson R J. Solving least squares problems[M]. Englewood Cliffs, NJ: Prentice-hall, 1974.
21. Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755): 788-791.
22. Liu Z, Chen X, Zheng Y, et al. Automatic keyphrase extraction by bridging vocabulary gap[C]//Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2011: 135-144.
23. Malkiel B G, Fama E F. Efficient capital markets: a review of theory and empirical work [J]. The Journal of Finance, 1970, 25(2): 383-417.
24. McCallum A, Wang X, Corrada-Emmanuel A. Topic and role discovery in social networks with experiments on enron and academic email[J]. Journal of Artificial Intelligence Research, 2007, 30(1): 249-272.
25. McCallum A. Multi-label text classification with a mixture model trained by EM[C]//AAAI'99 Workshop on Text Learning. 1999: 1-7.
26. Nallapati R, Cohen W. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs[C]//International Conference for Weblogs and Social Media. 2008: 84-92.
27. Ponte J M, Croft W B. A language modeling approach to information retrieval[C]//Proceedings of the 21st annual international ACM SIGIR conference

on Research and development in information retrieval. ACM, 1998: 275-281.

28. Reyes J, Schiavo S, Fagiolo G. Assessing the evolution of international economic integration using random walk betweenness centrality: The cases of East Asia and Latin America[J]. Advances in Complex Systems, 2008, 11(05): 685-702.
29. Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, 2004: 487-494.
30. Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.
31. Schumaker R P, Chen H. A quantitative stock prediction system based on financial news[J]. Information Processing & Management, 2009, 45(5): 571-583.
32. Schweitzer F, Fagiolo G, Sornette D, et al. Economic networks: The new challenges[J]. Science, 2009, 325(5939): 422.
33. Serrano M A, Bogun áM. Topology of the world trade web[J]. Physical Review E, 2003, 68(1): 015101.
34. Shannon C E. A mathematical theory of communication[J]. ACM SIGMOBILE Mobile Computing and Communications Review, 2001, 5(1): 3-55.
35. Stigler S M. Francis Galton's account of the invention of correlation[J]. Statistical Science, 1989, 4(2): 73-79.
36. Sun M, Shen D, Tsou B K. Chinese word segmentation without using lexicon and hand-crafted training data[C]//Proceedings of the 17th international conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 1998: 1265-1271.
37. Wang H, Huang H, Ding C. Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization[C]//Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011: 279-284.
38. Xu W, Liu X, Gong Y. Document clustering based on non-negative matrix factorization[C]//Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003: 267-273.
39. Yamamoto M, Church K W. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus[J]. Computational Linguistics, 2001, 27(1): 1-30.

致谢

首先感谢池明旻老师的指导，本文的选题、理论、实验等各方面都是在池老师的指导下完成的。池老师对待学术的严谨审慎态度和力求完善的作风都给我留下了深刻的印象，教会了我尤其讲求规范明晰的做学风格，希望在日后从事学术研究的过程中能继承和保持这样严肃的学风。

本文的完成也同样离不开实验室鲍江峰、刘隽等学长的指点，以及奚奇学长、陈鑫涛、邹杨修、黄季业、邵可嘉、陈涵洋同学的帮助。同时，要感谢本科四年过程中给予我诸多教益的老师、学长和同学，感谢计算机学院的课程使我的逻辑思维能力得到了锻炼。感谢张江校区图书馆、校车司机与后勤人员为张江生活抹上的色彩。

此外，感谢郑元者、侯体健等老师，孔雪莹同学、陈一星学姐、席越学姐、江睿杰学长，是他们的帮助使我能够接近理想的远方。感谢命运交错的所有人，你们都是我前行的动力。

最后，感谢父母一直以来无怨无悔的支持。且请允许我用骆老师的一句话作结：“关于过去所做的事情除了不满没有什么想说的，唯一的愿望是终有一天能够做好。”