基于文本大数据分析的会计和金融研究综述

马长峰，陈志娟，张顺明 2020管理科学学报

文本大数据分析在经济学和金融学中的应用：一个文献综述

沈艳, 陈赟，黄卓 2019经济学（季刊）

Textual Analysis in Finance

Tim Loughran, Bill McDonald

Presented by: Long Zhen

目录

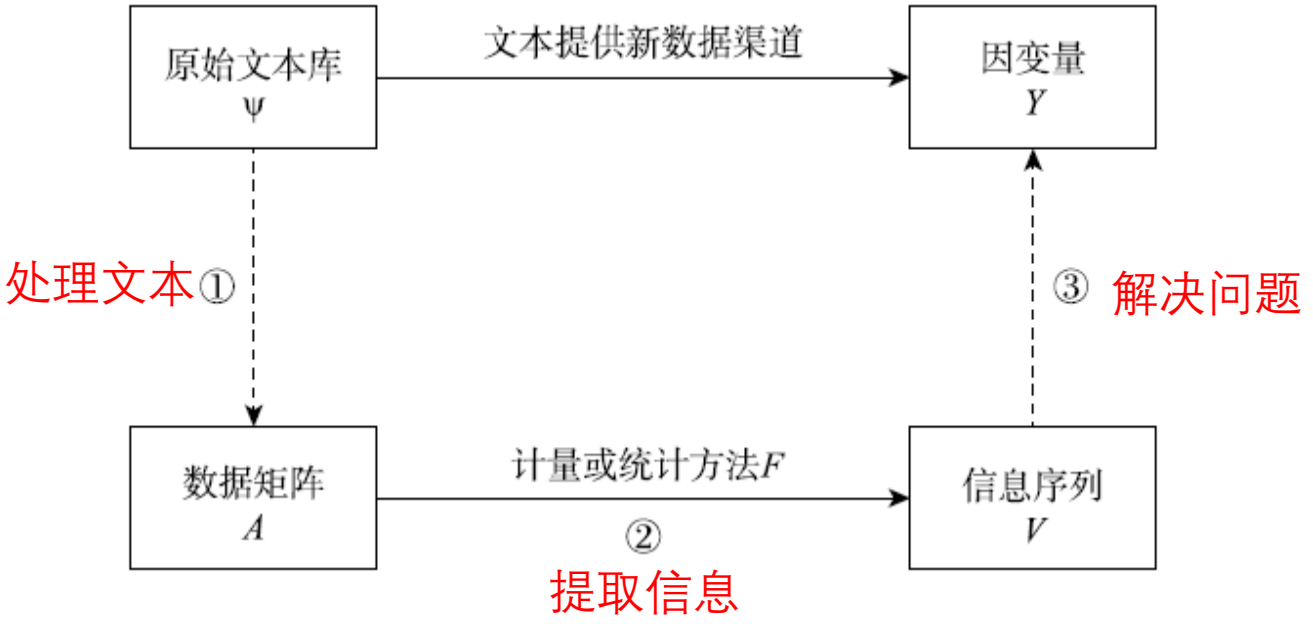
- 引言
- 文本处理
- 信息提取
- 问题应用
- 总结

引言

- 非结构化数据
- 三个特征:
  - 数据来源多样化:
    - 个人、企业、媒体、机构、政府…
    - 推特、微博、论坛、产品评价、公众号、年报、电话录音、招聘广告、招股书、分析师报告、会议纪要、名人演讲、央行公告……
  - 数据体量呈现几何级增长（纸媒→互联网媒介）
  - 时频高
- 核心问题:
  - 如何准确、有效地从文本中提取所需信息?

引言

- 研究：
  - 为经典问题提供新视角 eg. Sentiment
  - 产生新的问题 eg. Readibility
- 步骤：



文本处理

- 通过转换方法把**原始文本库**转换为**结构化矩阵**
- 分词 （文本→词）
  - 英文：空格分词→ 扩展为n元词组（n-gram从一个句子中提取n个连续的字的集合）
  - 中文：
    - 基于字符串匹配
    - 基于理解：加入句法语义分析
    - 基于统计：用ML学习切分规律

    往往是三种方法的结合 eg jieba软件包

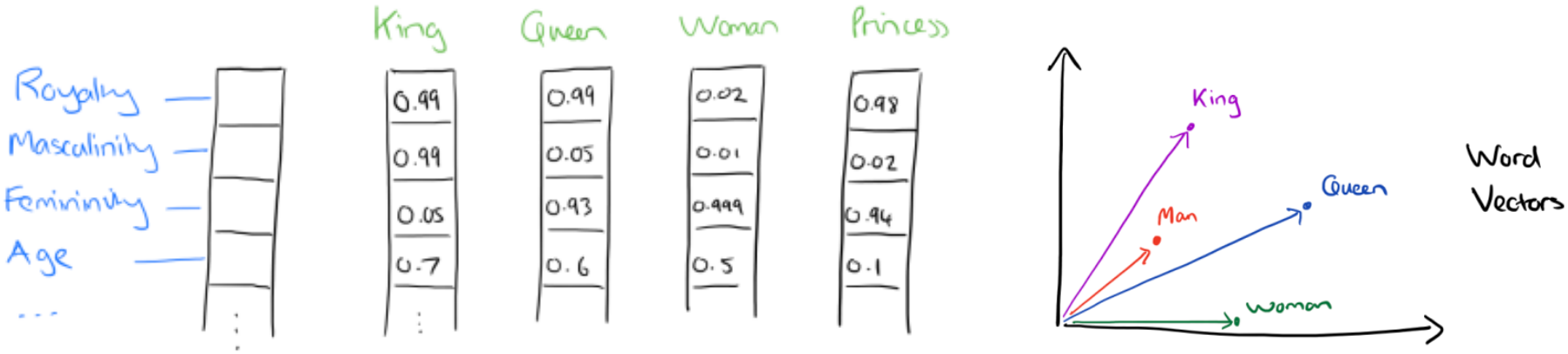
- 词→向量
  - **离散表示**：词袋模型、独热one-hot表示法：忽略语法和语序，把文本看作独立词汇的集合。得到TxN的数字矩阵（T为文本数，N为词）

例如，原始文本库 Ψ 由两条帖子组成。第一条的内容是“明天涨停。后天涨停没戏。”第二条是“玛丽有个小绵羊”。分词后得“明天、涨停、后天、没戏、玛丽、有、个、小、绵羊”九个不同词语，即 $N=9$。用独热表示法则“明天”用向量 $[1, 0, 0, 0, 0, 0, 0, 0, 0]$ 表示，“涨停”为 $[0, 1, 0, 0, 0, 0, 0, 0, 0]$，以此类推。于是第一个帖子可用向量 $[1, 2, 1, 1, 0, 0, 0, 0, 0]$ 表示，第二个帖子即 $[0, 0, 0, 0, 1, 1, 1, 1, 1]$。

→忽略上下文结构而产生歧义

→高维度稀疏矩阵，需要降维

- **分布表示**：词嵌入技术：eg. Word2Vec
  - 通过简单的神经网络训练，将上下文联系起来



$$\overrightarrow{King} - \overrightarrow{Man} + \overrightarrow{Woman} = \overrightarrow{Queen}$$

信息提取

- 从结构化矩阵中提取信息
- 无监督学习
  - 词典法: 统计文本中词语出现次数，通过**加权**提取信息。
    - 核心：选取合适的词典，金融会计领域情绪主要用LM词典；也可以自己构造 eg. Li et al, 用股票论坛帖子手工构建中国股吧金融情绪词典。
    - 权重：等权、TF-IDF (Term Frequency-Inverse Document Frequency)
  - 主题分类模型：
    - Latent Dirichlet Allocation, LDA：预设文档主题个数，LDA模型将每篇文档的主题以概率形式给出
    - 层次狄利克雷过程：不需要设定主题数
    - 动态主题模型：允许主题分布随时间变化

    Eg. Forecasting the Equity Premium: Mind the News！RF,2020

- 有监督学习
  - 经典的有监督机器学习法
    - 朴素贝叶斯、支持向量机、决策树、K近邻……
  - 深度学习
    - 神经网络
    - 拓展：DNN/ CNN/ RNN eg. Li et al，采用SVM和CNN度量中国投资者情绪
    - Transfer learning
    - 可以提供非线性分类，但参数过多

- **小结**：如何选择信息提取方法？
  - 语料来源是什么？文本长短、语言逻辑强弱……
  - 是否有合适的词典？
  - 是否有高质量的标注数据和明确的模型选择标准？
  - ……

解决问题

- 指标分类
  - 可读性
  - 关注度
  - 情绪
  - 隐含波动率
  - 意见分歧
  - 行业关联性
  - 财务欺诈

- 关注度
  - 投资者关注度
    - 传统变量：交易量、超额收益率、广告费用……
    - 文本数据：
      - 散户：网络搜索次数、论坛股民发帖数 eg. Da et al (2011) 用个股的谷歌搜索指数构建关注度指标，发现短期收益率更高
      - 机构投资者：eg. Ben-Rephael et al(2017) 采用Bloomberg终端搜索和浏览记录度量机构投资者关注度
  - 媒体关注度
    - 资产价格影响：eg. Hillert et al(2014) 用220万条新闻数据研究媒体关注与动量效应的关系，发现更受关注的公司可预测性越强
    - 管理层行为影响：eg 周开国等（2016）媒体关注度的提高会降低企业违规概率
  - 分析师关注度
    - eg 周开国（2014）媒体关注度可以影响分析师关注度

- 情绪
  - 媒体：eg Chen et al(2019) 检查StockTwit的情绪及其对加密货币收益率的影响.
  - 管理层：eg Jiang(2019)使用上市公司财报和电话会议记录等文本数据衡量了市场层面的经理人情绪，并发现经理人情绪能显著反向预测未来市场收益率
  - 投资者

文献	数据来源	数据类型	情绪指数类型	主要发现
Tsukioka *et al*.（2018）	2001—2010 年雅虎财经日本股票版块	654 家公司相关的帖子数据	个股层面投资者情绪指数	投资者情绪可用于解释 IPO 抑价现象
Sun *et al*.（2016）	1998—2011 年汤普森路透数据库	标普 500 指数对应的 1 分钟频率情绪数据	市场层面投资者情绪指数	日内半小时情绪变化可预测日内收益率
Renault（2017）	2012—2016 年美国社交媒体平台 Stock Twits	约 6 000 万条帖子	市场层面投资者情绪指数	第一半小时投资者情绪变化能预测标普 500 指数 ETF 最后半小时收益率，但下个交易日反转
Behrendt and Schmidt（2018）	2015—2017 年推特	道琼斯指数成分股情绪数据（1 分钟频率）	个股 Twitter 情绪	情绪与日内波动率存在反馈效应，但经济意义不显著

- 可读性
  - 迷雾指数Fog index

    $Fog = 0.4 * (\text{平均每句单词数} + \text{复杂单词占比} * 100)$

    → 对金融文本难以适用
  - Bog Index：generated by proprietary software like $StyleWriter$

    → not transparent; actually measure writing style, not readability
  - 基于平实英语(plain English)：第一、二人称代词数
  - 对金融文本直接使用年报文件大小
    - 在中国可行吗?
  - 中国文献有调整，如用笔画数度量汉字难度
    - eg丘心颖等（2016）根据年报完整句子占比、基础词汇占比、汉字笔画数等构造指标研究了年报可读性对分析师关注、预测信息含量以及预测准确性的影响

- 新闻隐含波动率指数
  - Eg. Manela and Moreira (2017)使用《华尔街日报》头版数据，采用支持向量回归法将新闻文本数据中出现的词语和市场上的VIX指数相对应，构建了新闻隐含波动率指数
- 投资者分歧
  - 传统：分析师预测分散程度、经济政策不确定性指数、对经济变量预测的分散程度
  - 网络发帖情绪得分标准差
  - Eg 段江娇等（2017）使用东方财富股吧上的帖子数据构建了日度频率的投资者分歧，发现投资者分歧越大，未来两天的交易量也越大

- 财务欺诈
  - Hoberg and Lewis (2017) examine, in the Management Discussion and Analysis (MD&A) section of a 10-K, whether word usage differs between AAER会计和审计强制披露 firms and industry-age-size matched non-AAER firms and find that specific vocabulary choices can actually be used to help predict accounting fraud out-of-sample.
- 企业复杂度
  - Loughran and McDonald (2020) create a list of more than 300 words that are markers for firm complexity.

结论和展望

- 更丰富的数据源
  - 政府工作报告、公众号文章、微博大V观点、专利、法院判决（裁判文书网）、医生处方……
- 更深入和广泛的问题
- 从预测到因果关系的研究
- 方法之间的联系和差异
- 跨学科趋势