上海证券交易所博士后出站报告

证券市场文本挖掘技术研究与应用

报告人：白雪

所内导师：白硕

所外导师：朱扬勇

联系人：卢文莹

日期：2015.6

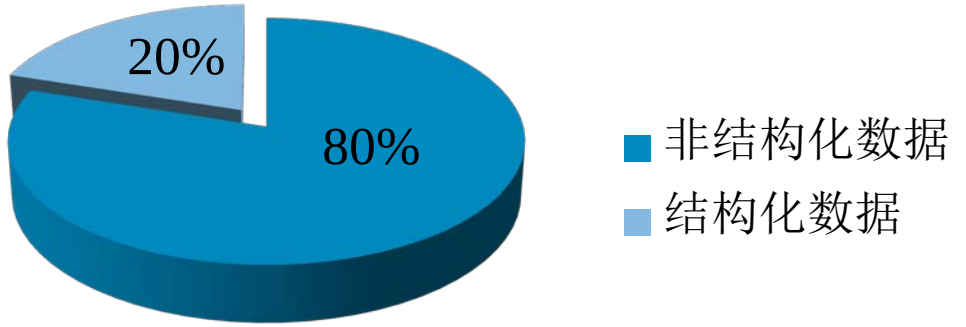
上海證券交易所 SHANGHAI STOCK EXCHANGE

汇 报 提 纲

1. 研究背景和意义
2. 报告主要章节
3. 主要研究成果
4. 启示与建议
5. 工作总结

研究背景和意义

- 互联网数据大爆炸时代
  - 海量非结构化数据尚未得到有效应用
- 证券行业
  - 相关的网络数据资源越来越多
  - 对于信息的快速加工、精准反应需求也与日俱增
    - 市场情绪网络留痕——为度量、量化情绪提供了素材
    - 主题投资、事件套利
    - 舆情监控、信息追踪



对证券行业文本挖掘产生了迫切的需求

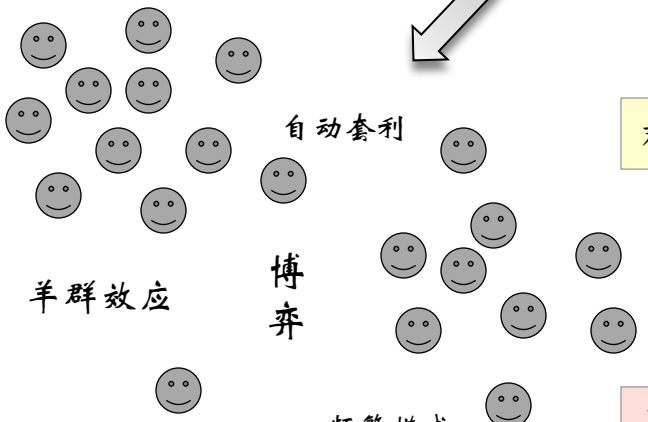
上海證券交易所 SHANGHAI STOCK EXCHANGE

文本对市场的影响

文本

主题、观点、情绪等

消息影响投资者行为

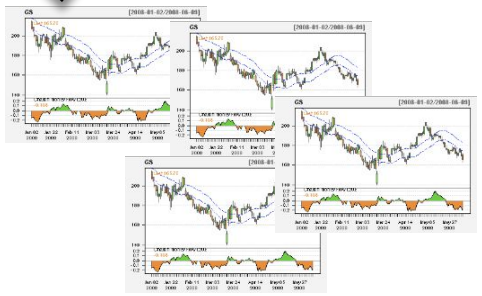


群体行为导致股价波动

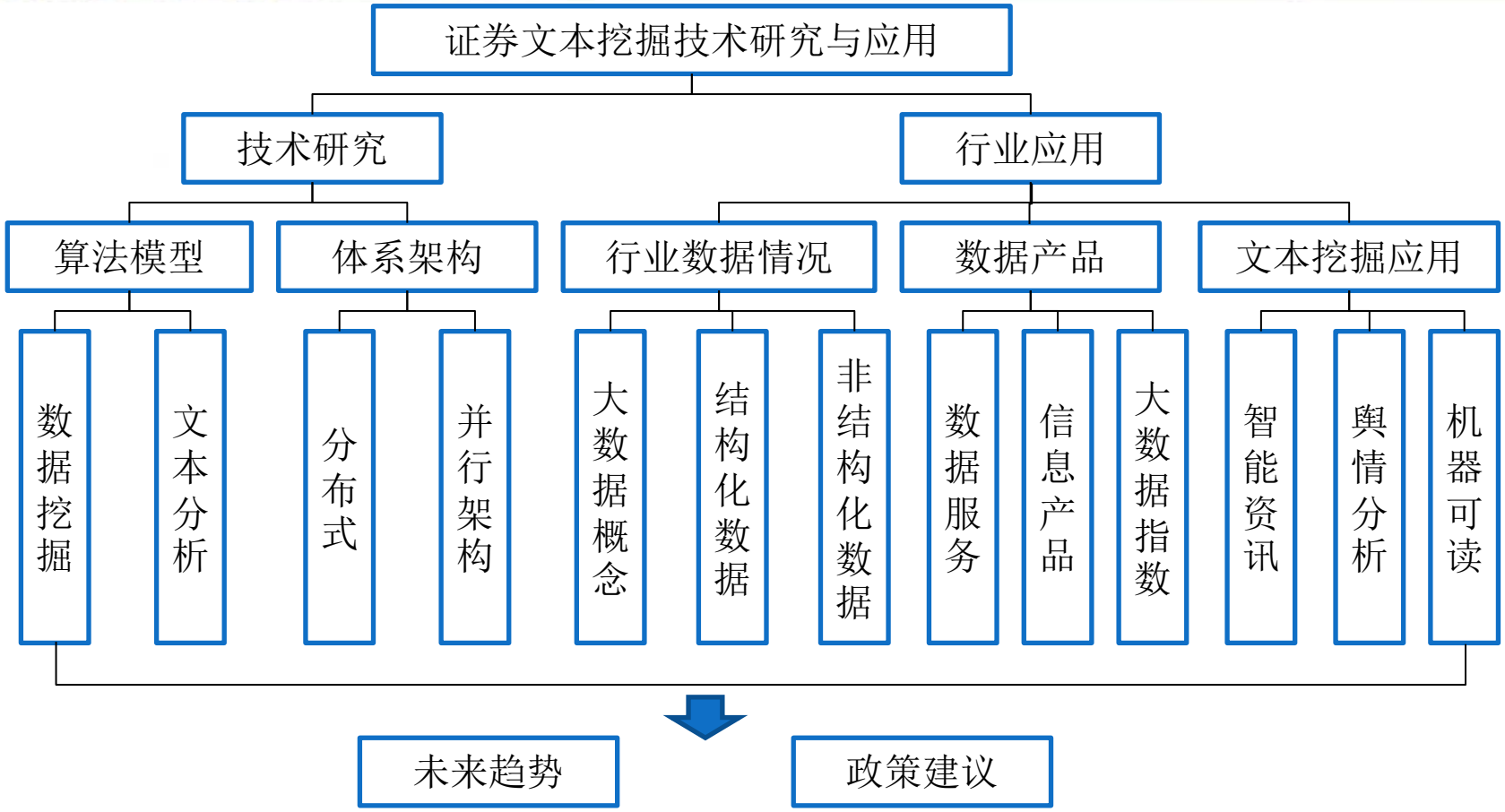
股价波动影响投资者行为

文本关联相应股票

文本反映市场波动



研究内容



上海證券交易所 SHANGHAI STOCK EXCHANGE

报告章节目录

上海證券交易所 SHANGHAI STOCK EXCHANGE

数据、算法和架构

证券行业数据现状

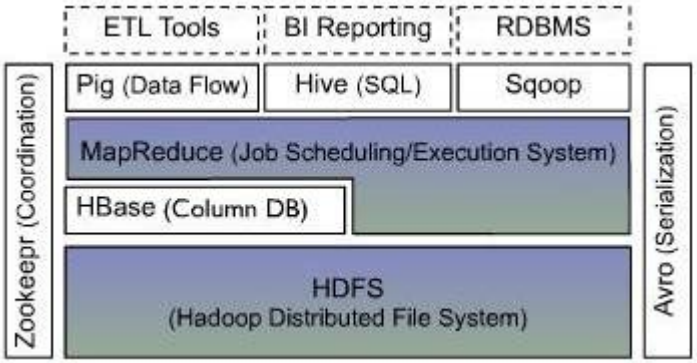
- 结构化数据
  - 行情数据
  - 成交数据
  - 指标数据
- 非结构化数据
  - 公司数据
  - 网络数据
  - 研报数据

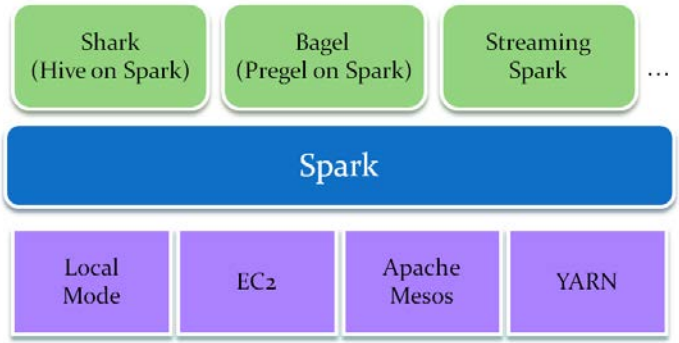
数据挖掘算法

- 文本挖掘
- 情感分析
- 关联分析
- 聚类分析
- ……

大数据基础架构

Hadoop　Yarn　Spark　Storm





行业文本挖掘现状调研

- 证券行业文本挖掘应用可归纳为三大类
  - 投资综合性社区 StockTwits 雪球
    - 各种重要消息发布、频繁互动、思想碰撞的交流平台
  - 文本信息资讯服务 THOMSON REUTERS Bloomberg 朝阳永续 SUNTIME
    - 金融资讯端文本分类、机器可读等服务
  - 专业文本挖掘服务 The Stock Sonar See through the current SmogFarm
    - 互联网专业情感分析服务
- 大数据指数
  - 南方-新浪大数据 i指数
  - 百度百发 百发指数 Baifa Index

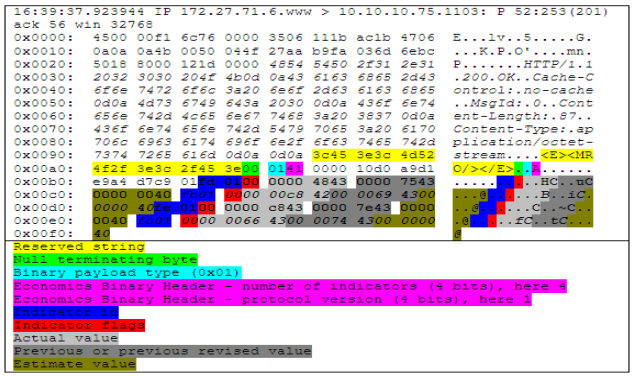
上海證券交易所 SHANGHAI STOCK EXCHANGE

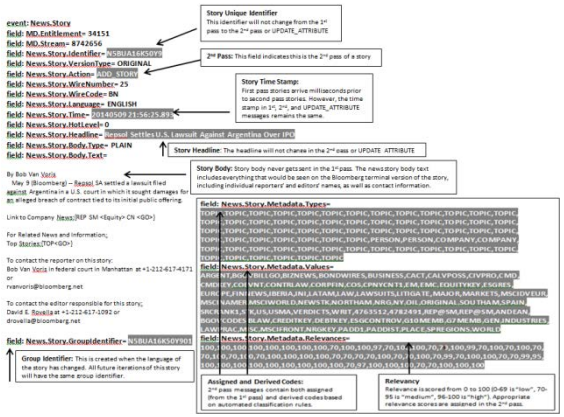
机器可读新闻技术调研

- 机器可读新闻

THOMSON REUTERS　　Bloomberg

  - 基于信息编码技术、文本挖掘、自然语言处理、情感分析、统计学等算法技术；
  - 以电脑可读的语言编写，**低延迟**瞬时传递关键信息；
  - 服务于高频交易、程序化交易、实时指数发布、套利交易、风险管理等。





上海證券交易所 SHANGHAI STOCK EXCHANGE

文本挖掘应用调研启示

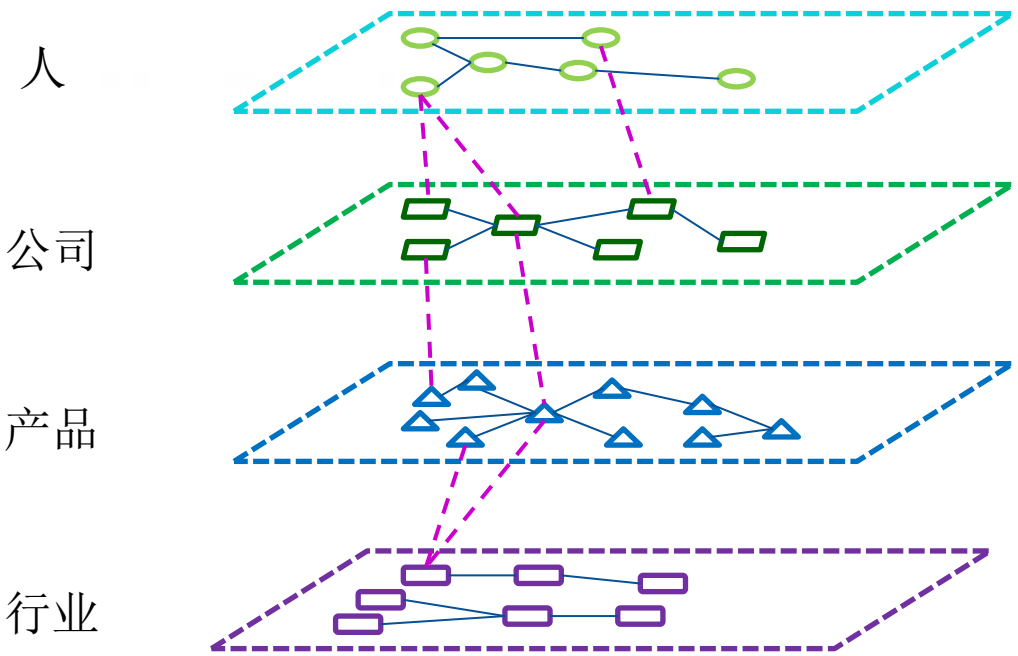
- 证券行业文本挖掘有着广泛的需求及市场前景
- 仍存在很多问题：
  - 证券行业深度依赖信息精准度 VS 文本挖掘？
  - 网络舆情的准确度、可信度
  - 数据孤岛、作用链传导
  - 大数据交互印证

文本挖掘应用调研启示

- 建立证券行业知识图谱是证券文本语义理解和语义搜索的关键基础技术，为证券领域文本分析、舆情监控、知识发现、模式挖掘等提供了坚实支撑
  - 行业专业中文词典
  - 业务知识分析及行业特性储备
  - 证券行业实体对象及属性的定义与提取
  - 行业内实体关联关系网络拓扑图的建立
  - 行业语义分析等等

发现潜在关联、行业推理、传导链、量化影响程度、提高分析精准度.....

基础工作之——证券行业知识图谱



万物互联

蝴蝶效应

E.g.，突发洪水，导致一类原料短缺，这与所持股票有何关系？又将预示着哪些投资套利机会？

证券行业知识图谱

- 终极目标：建立一个全谱系的上市公司关联图谱

证券行业
知识图谱

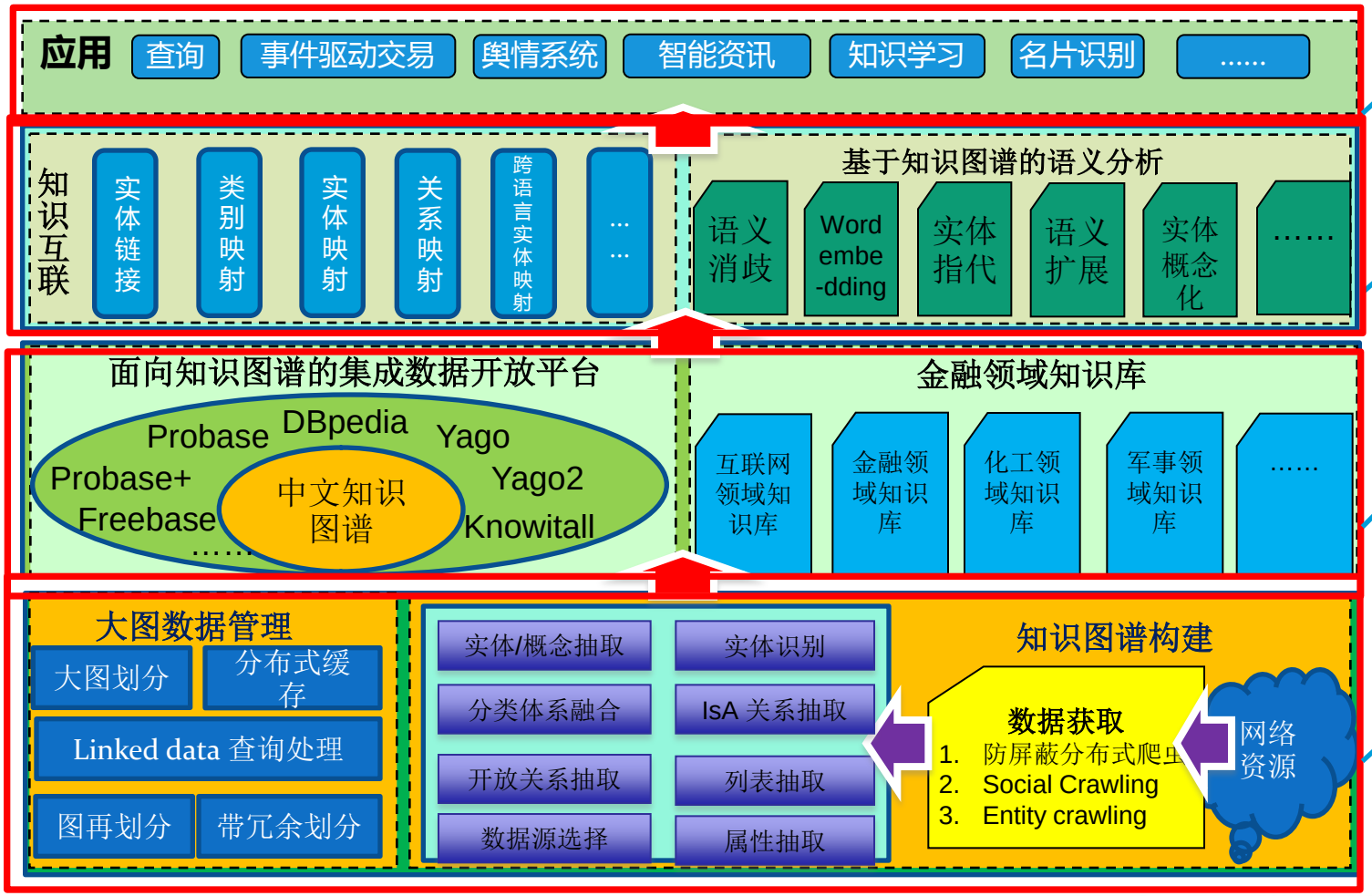
监管
- 追溯异动产生根源
- ……
- 发掘打击内幕交易

市场
- 关联事件套利
- ……
- 个性化智能化资讯

- 相关联合研究项目：证券知识图谱关键技术研究

上海證券交易所
SHANGHAI STOCK EXCHANGE

知识图谱系统架构设计



应用

服务

集成

数据

总结与建议

- 给监管部门提供了全新的**风险评估与管理**手段
- 提供了高效、便捷、准确的**信息情报分析**方式
- 有助于还原相关金融市场事件的**因果链条**
- 大数据的体量、维度均为分析的**准确性**提供了**保障**

舆情监控

打击内幕交易

搭建大数据服务体系

投资者资格认定

加强对拟上市公司的审查

建立大数据指标指数

工作总结

- 发表文章：
  - 《证券行业文本挖掘技术应用现状与探讨》 交易技术前沿
  - 《证券行业数据及大数据服务探讨与展望》 交易技术前沿
  - 《机器可读新闻技术与服务》 交易技术前沿
  - 《金融时间序列数据挖掘实例分析》 交易技术前沿
  - 《时间序列显著模式挖掘及其在证券市场的应用展望》交易技术前沿等。
- 翻译：
  - 《MESA：异地可复制、近实时、可扩展数据仓库》
  - 《知识图谱的探索性搜索》等
- 书籍：
  - 《证券市场国际化技术实践》

上海證券交易所 SHANGHAI STOCK EXCHANGE

工作总结

- 参与课题：
  - “基于知识图谱的沪港两地情绪指数研究”，上证联合研究课题，课题协调人。
- 所内工作参与：
  - 参与证监会《行业信息安全规划》的制定；
  - 参与所内舆情系统的试用、问题反馈，并参与证监会互联网信息采集及舆情监测系统需求统筹工作会议；
  - 负责《交易技术前沿》的编辑、征稿、审稿、出版等工作；
  - 参与沪港通检查小组，对券商沪港通技术方面的准备情况进行了解、调研与学习；
  - 负责每周《金融技术动态》的资料采集，编辑和发布工作。

请各位老师批评指正

谢谢！

上海證券交易所
SHANGHAI STOCK EXCHANGE