BDTC
2014 中国大数据技术大会
BIG DATA TECHNOLOGY CONFERENCE
暨第二届CCF大数据学术会议

金融投资大数据实践分享

龙白滔 博士

2014年12月14日

目录

- 金融大数据 vs （消费）互联网大数据
- 金融数据生产
- 金融大数据存储
- 金融大数据分析和挖掘
- 在线交互式金融编程分析研究平台

金融大数据 vs （消费）互联网大数据

	（消费）互联网	金融
研究对象	比较偏重研究个体的行为体征	比较偏重研究群体行为和趋势
数据相关性	与个体强相关的数据比较容易获得（例如浏览器cookie），数据噪音小	与群体行为强相关的数据比较难获得，数据噪音大
算法复杂度	因为数据质量高，所以算法可以相对较简单	因为数据噪音大，因此对算法要求很高
数据容量	大	更大，互联网大数据 + 金融专门的大数据（例如行情数据、行业数据、分析师报告等）
数据类型	多种结构化和非结构化数据	更多，互联网的数据类型 + 金融特别的数据类型，例如时间序列数据
数据速度	一般数据处理速度要求不高	对数据处理速度要求比较高，例如量化交易、动态风险定价、反信用卡欺诈、实时新闻分析和处理等

目录

- 金融大数据 vs （消费）互联网大数据
- 金融数据生产
- 金融大数据存储
- 金融大数据分析和挖掘
- 在线交互式金融编程分析研究平台

金融数据生产

数据采集

- 分布式爬虫系统
- 基于主题关注度的数据采集
- 优化每日增量爬取策略
- 实现代理机制
- 支持动态网站爬取

数据抽取和解析

- 自动抽取OFFICE、PDF等文档中的数字、文字、表格、公式等
- 利用OCR和图片处理技术抽取图像数据
- 行业分类：国泰君安和证监会的分类标准
- 财务数据支持中国的会计准则
- 人工定义和机器学习抽取/解析规则
- 自动化数据抽取和解析的流程

数据质量检查

- 支持多vendor数据质量平台
- 添加自动数据质量校验
- 人工校验结果作为反馈输入到机器学习模型，进一步提升自动化处理的质量
- 众包平台来提升数据质量检查的效果和降低成本

结果：几乎完全自动化地采集、抽取、解析和质检传统的金融数据，包括上市公司基本信息、财务信息、公司事件和公告等，包括历史数据，质量和效率全面超越了传统的金融信息服务提供商。

目录

- 金融大数据 vs （消费）互联网大数据
- 金融数据生产
- 金融大数据存储
- 金融大数据分析和挖掘
- 在线交互式金融编程分析研究平台

金融大数据的存储

新闻数据和社交媒体数据（文本类型）

- 财经类新闻，每天8000篇左右
- 过去10年所有财经类新闻，1000万篇左右
- 元数据和处理过后的数据，例如新闻分类、故事（新闻聚类）、事件和标签等
- 暂存：Cassandra vs MongoDB
- 历史数据存储：HDFS

行业数据和宏观经济数据（RMDB的结构化数据）

- 数据量不大，目前我们用MySQL
- Cassandra在逐渐代替传统RMDB（包括MySQL和Oracle）在企业内部的作用，作为大容量实时或者近实时存储和分析平台，例如全球最大的云应用Netflix（95%的数据从O->C，拥有50个C集群共750个节点）、纽交所、Splunk和Barracuda Networks（MySQL->C）

金融大数据的存储（续）

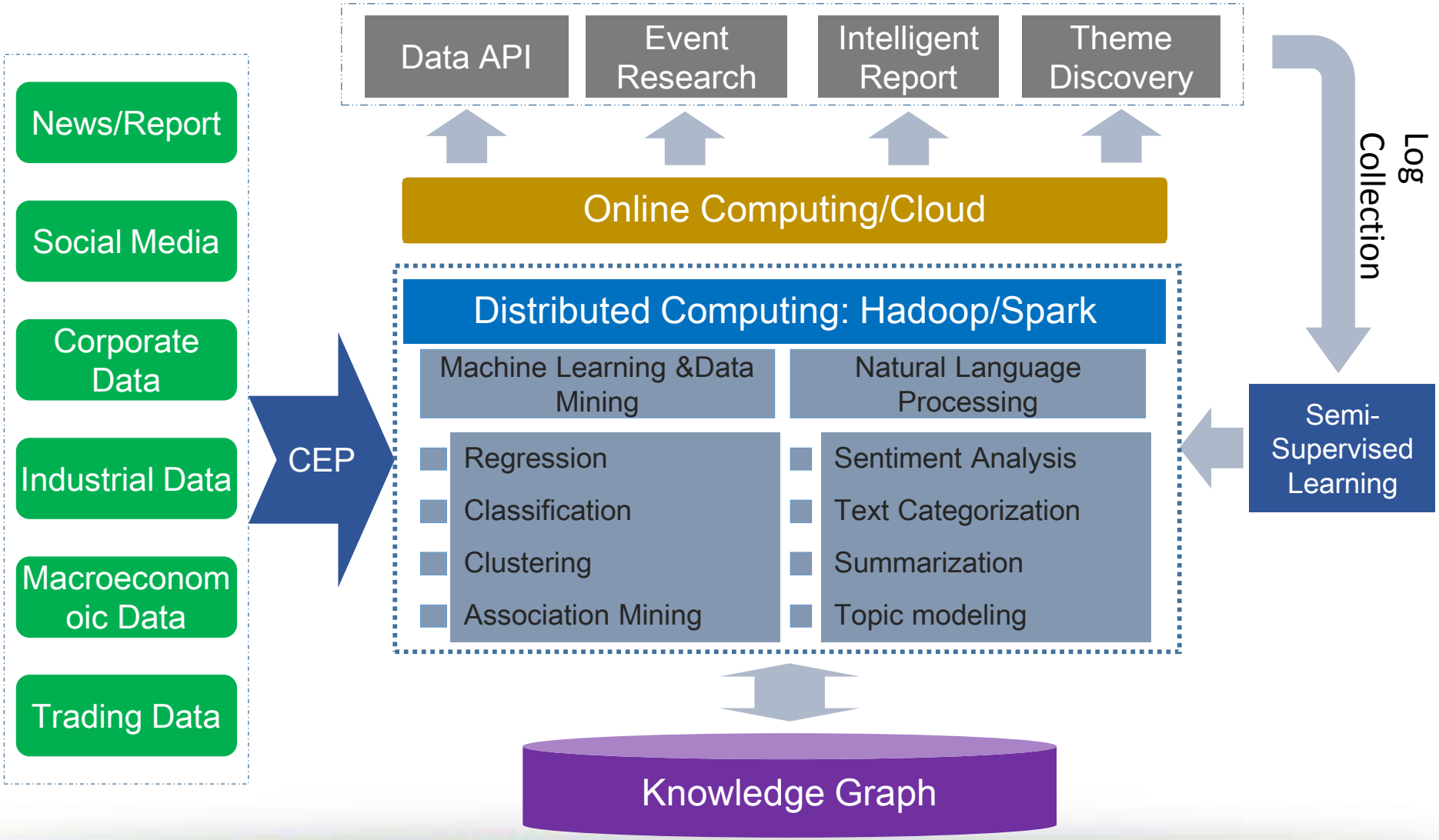
市场行情数据（实时+历史）（时间序列数据）

- 商用：
  - ✓ KDB，传统金融机构标配，高富帅，专用开发语言q（复杂但高效）
- 开源：
  - ✓ Cassandra在国外已经得到比较成功的应用（物联网和能源数据）
    - ➢ row key的设计非常适合将时间序列数据分散到集群各个节点进行存储
    - ➢ 提供类SQL的查询语言CQL
    - ➢ 分布式集群提供卓越的水平扩展性和较好的查询性能（典型查询100ms级，集群处理70请求/s）
    - ➢ NASA（安全数据）,Tendril（目前5T/月，未来20T/月能源时间序列数据），Agentis Energy（150亿个时间序列记录，Cassandra集群跨越2个数据中心）
  - ✓ 我们目前的选择-InfoBright
    - ➢ 列存数据库，高数据压缩率（5年高频股票历史数据2.7T->140GB，期货和其它历史数据5-6T->250G）
    - ➢ Partition-index：快速实现对数据某个区域的查询
    - ➢ SQL兼容，提供较好的查询性能；（典型查询50ms级别，单机300处理300请求/s）
    - ➢ 开源版本支持单机和单核，扩展性有限

目录

- 金融大数据 vs （消费）互联网大数据
- 金融数据生产
- 金融大数据存储
- 金融大数据分析和挖掘
- 在线交互式金融编程分析研究平台

金融大数据分析和挖掘


Data API
Event Research
Intelligent Report
Theme Discovery
News/Report
Social Media
Corporate Data
Industrial Data
Macroeconomoic Data
Trading Data
CEP
Online Computing/Cloud
Distributed Computing: Hadoop/Spark
Machine Learning &Data Mining
Natural Language Processing
Regression
Classification
Clustering
Association Mining
Sentiment Analysis
Text Categorization
Summarization
Topic modeling
Log Collection
Semi-Supervised Learning
Knowledge Graph

算法应用

- 投资研究
  - ✓事件研究
  - ✓主题发现和跟踪
  - ✓量化分析
- 数据聚合
  - ✓个股/主题新闻聚合
  - ✓智能研报，情感指数，热度统计

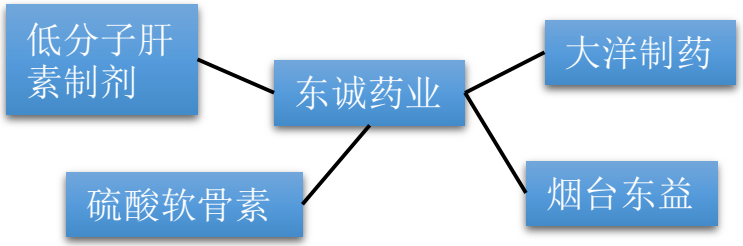
基础算法

- 金融领域文本分词服务
- 财经新闻搜索引擎
- 新闻类型分析
- 新闻聚类
- 情感分析
- 知识图谱

金融新词发现

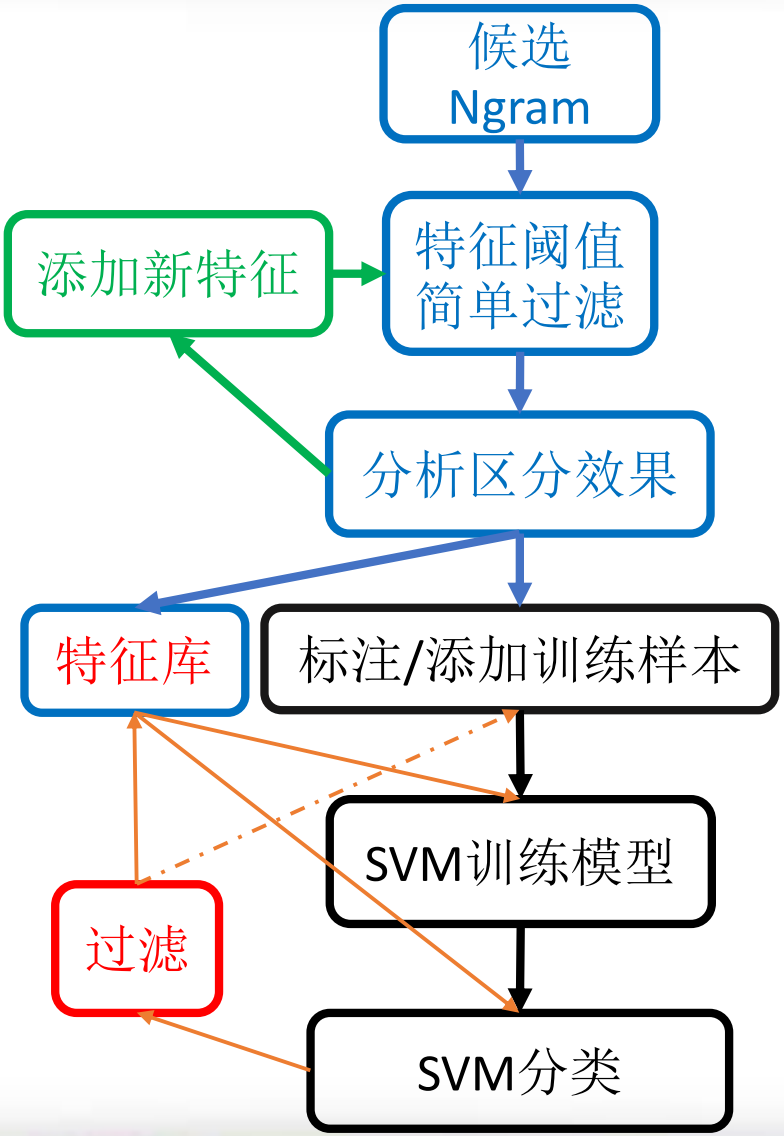
金融类文本切词

金融实体关联



产业热点实时跟踪

新词发现的流程框架


候选
Ngram
特征阈值
简单过滤
添加新特征
分析区分效果
特征库
标注/添加训练样本
SVM训练模型
过滤
SVM分类

新闻聚类

应用

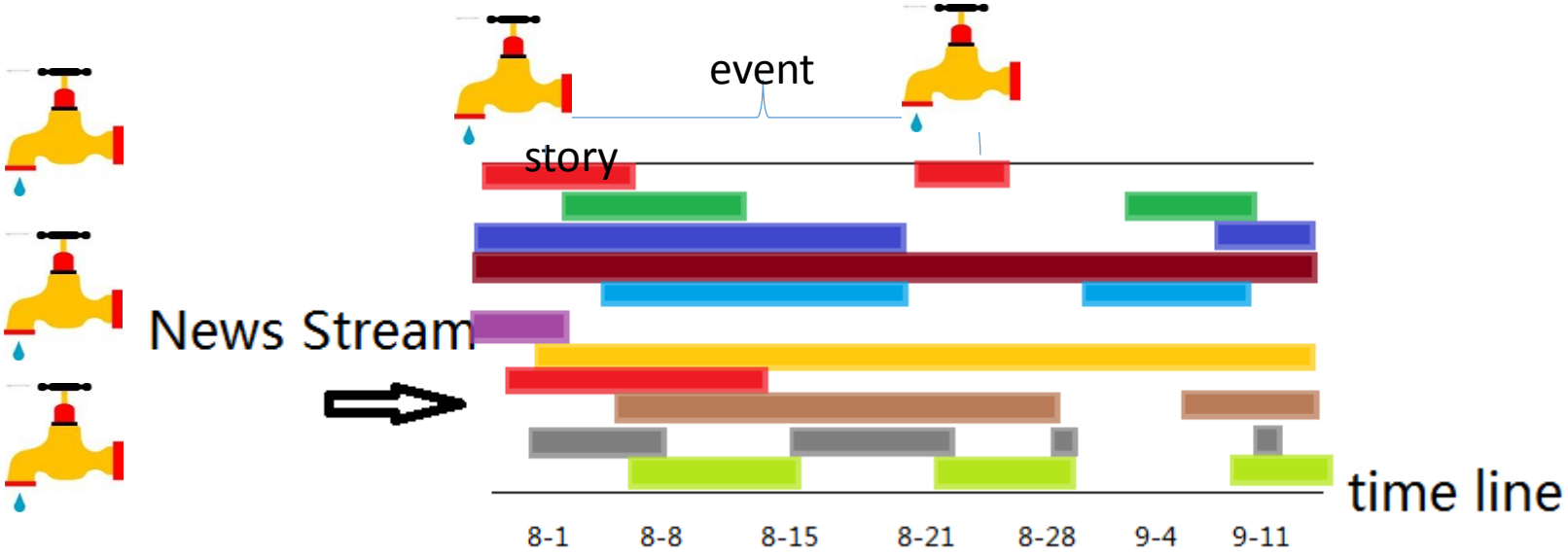
- 新闻聚合，选取代表新闻
- 发现事件，追踪事件

算法

- 个股新闻聚类，采用近邻传播算法，对个股新闻流聚类，可以跟踪近期热点事件，动态生成新类
- 全网事件发现，基于spark框架实现的分布式层次聚类

事件发现

- 事件发现等价于TDT (Topic Detection & Tracking)
  - ✓ Detection: identifying the new events
  - ✓ Tracking: associating stories with old events
- News → stories → Events



聚类结果实例

唱歌跑调 粉丝离场
张曼玉承认走调
张曼玉献声草莓音乐节 唱歌跑调致粉丝离场
张曼玉北京亮嗓遭狂风被叫停 承认上海首唱演砸
张曼玉一开口 粉A丝忙离场
周迅、杨幂等被媒体评为“唱歌不好听的女神”
张曼玉唱歌跑调被批
张曼玉草莓音乐节：请给我20次机会

挂面爷爷 《舌尖2》“挂面爷爷”因病去世
走进《舌尖2》幕后：不仅仅是“美食”那么简单
挂面爷爷患癌去世 悼念：张家山手工挂面巅峰人物
“挂面爷爷”患癌去世,导演撰文悼念
《舌尖2》吴堡挂面主人公患骨癌去世
《舌尖2》挂面爷爷去世
“舌尖2”挂面爷爷患癌去世
【挂面爷爷】“舌尖上的中国”挂面爷爷患癌去世

情感分析

- 判断新闻和研报对个股是看涨看跌
  - ✓综合基于词典的方法和基于机器学习的算法

新闻热度

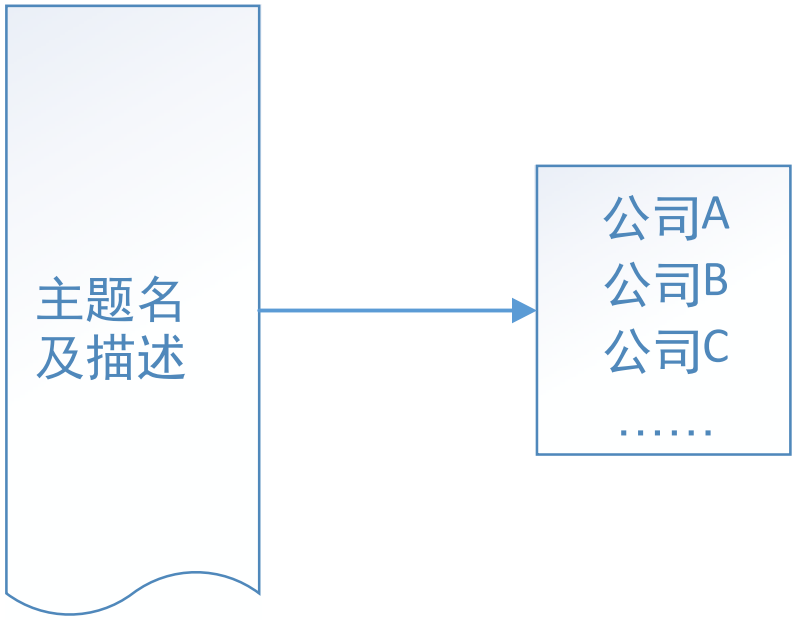
综合新闻及社交媒体判断上市公司的实时关注度。

知识图谱

- 实体及关系查询
- 量化指标关联
  - ✓ 自动找出影响个股股价的行业和宏观指标
- 主题成分股推荐(公司深度关联)
  - ✓ 给定一个主题及描述，自动生成该主题的成分股

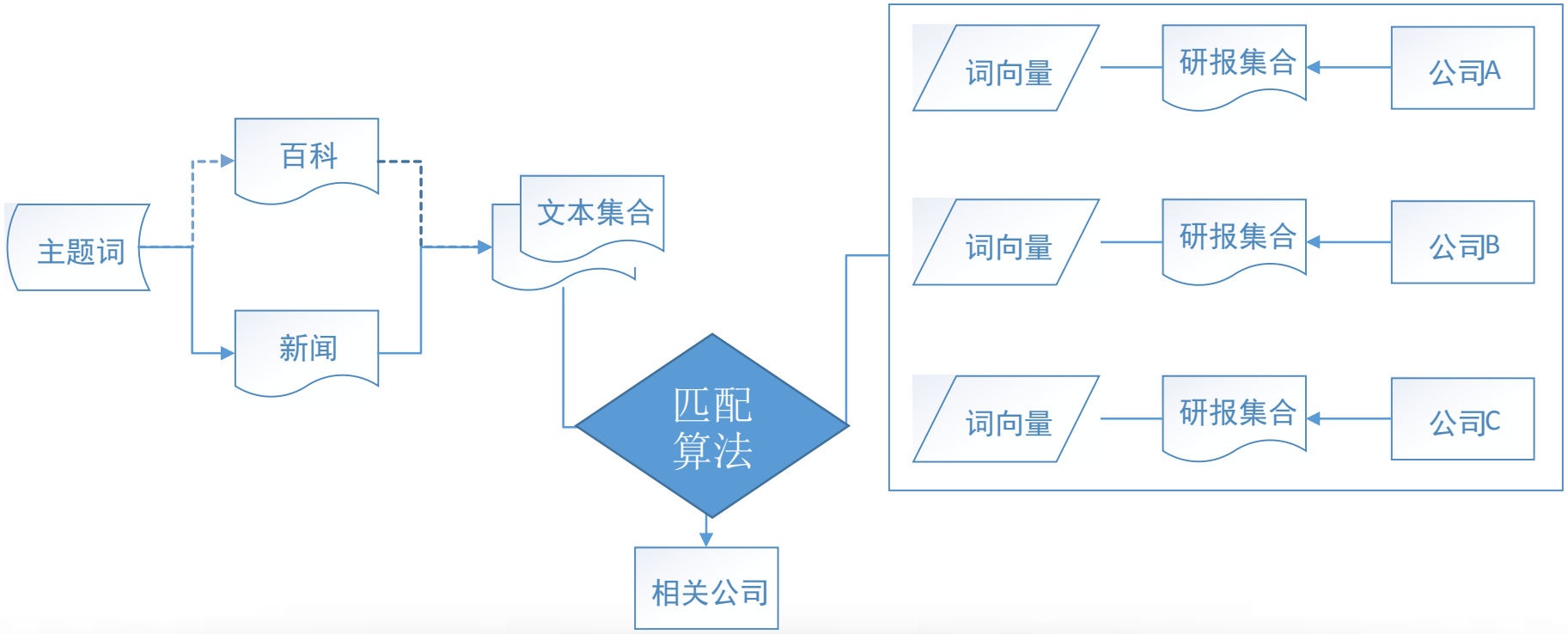
主题成分股推荐

- 输入
  - ✓一个长文本（主题描述）
  - ✓一个词
- 输出
  - ✓一篮子股票
  - ✓关联分数
  - ✓关联原因



主题成分股推荐

- 数据
  - ✓百科：根据关键词扩展主题描述
  - ✓新闻：根据关键词扩展相关新闻
  - ✓研报：产生公司相关的文本/词语描述



目录

- 金融大数据 vs （消费）互联网大数据
- 金融数据生产
- 金融大数据存储
- 金融大数据分析和挖掘
- 在线交互式金融编程分析研究平台

在线交互式金融编程分析研究平台

涵盖

金融分析

大数据分析

衍生品

固定收益

算法交易

一个云端、多租户、在线交互式的策略开发、回测、部署和分享平台。

基本功能

金融工程库：

主要提供金融领域常用的日期处理函数，Vanilla期权定价函数，债券估值功能以及收益率曲线构造能力。其中：

期权模型包括:
- Black - Scholes - Merton
- Bachelier
- Displaced Diffusion
- CEV
- SABR/Hagan
- Heston

债券类型包括：
- 零息利率国债
- 固定利率债券

收益率曲线构造，可以基于：
- 贴现因子
- 零息利率
- 远期利率

金融数据API

主要提供股票/期货市场行情,基本面数据和宏观数据：

市场行情数据包括:
- getTickRTSnapshot --获取最新市场信息快照
- getTickRTSnapshotIndex --获取指数成份股的最新市场信息快照
- getFutureTickRTSnapshot --获取期货最新市场信息快照
- 等。。。

基本面数据包括：
- getBalanceSheetOnePeriod --获取资产负债表信息
- getCashFlowOnePeriod --获取现金流表信息
- 等。。。

宏观数据包括：
- getMacroKPIDataByCode --根据指标的datayescode获取kpi数据
- getMacroKPIDataByName --根据指标的中文名获取kpi数据
- 等。。。

社区功能

- 一键分享，社区或URL
- 一键克隆
- 支持移动终端

其它功能

- 接入交易系统
- 策略跟踪和交易
- 接入其它类型的数据API，例如新闻，以及用户自定义数据API等

后台其实是一个完整的PaaS服务平台

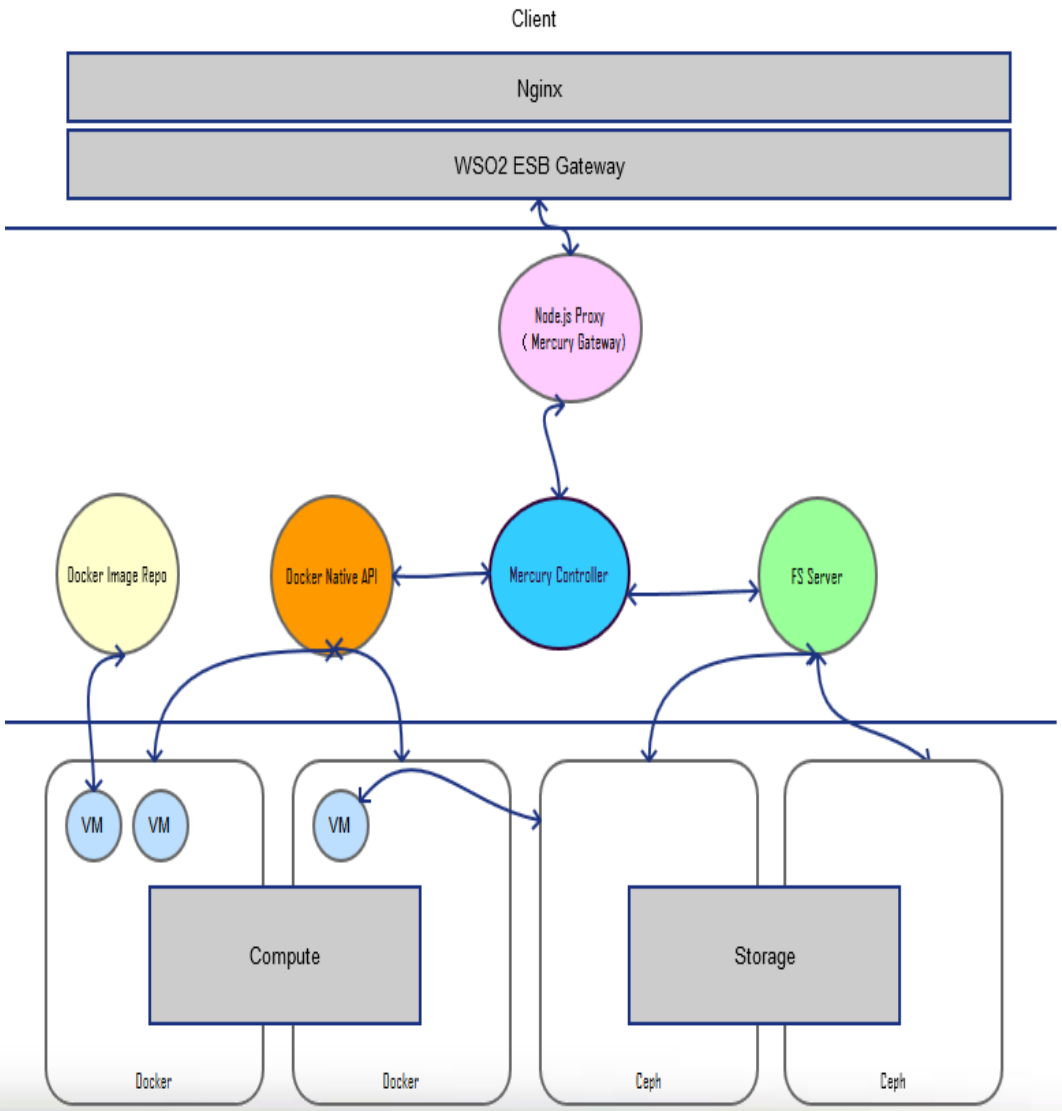
云平台

Mercury Gateway

Mercury Controller

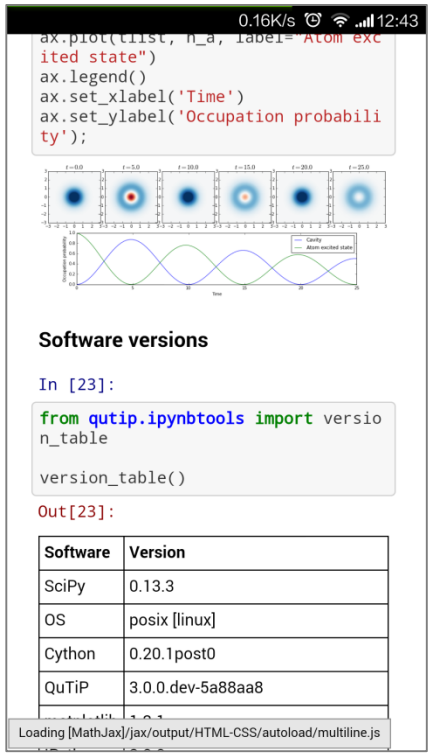
Mercury Cloud

Mercury Runtime



在线策略编程-甚至在移动端

notebook demo



谢谢！