搜披露

技术、产品、商业化的发展思考

2016.11

搜披露：起点

- 本质：用户从信息披露中获取有用信息
  - 领域特定：
    - 信息披露相关，跨金融、法律等领域
  - 有用信息：
    - 对谁有用？不同用户群体，如投行从业人员、二级市场投资者
    - 为什么有用？可以通过信息降低业务中的不确定性，如了解前例、了解公司经营信息
    - 什么样的组织形式有用？非结构化/结构化、遵循业务逻辑的数据之间的关联
  - 获取：
    - 快速：搜索 vs 逐篇阅览
    - 全面：全文搜索 vs 标题搜索，对PDF的高程度解析
    - 可长期贮存：拷贝到本地 vs 浏览

搜披露：发展规划

- 技术
  - 非结构化数据的处理
  - PDF解析
  - ElasticSearch
  - 技术架构调整
- 产品
  - 订阅功能
  - 数据源种类
  - 数据关联
- 商业化应用场景
  - 资本市场中介服务
  - 一二级市场投资
  - 券商经纪业务
- Synergy
  - 技术层面
  - 产品层面
  - 商务层面

技术：非结构化数据的处理

- 处理非结构化数据
  - 信息丰富、逻辑性强、易于理解
  - 篇幅长，难以归类、量化
  - 如何兼顾？
    - 从非结构化数据中提取key，通过key的结构（树形、网状）实现对前者的归并
    - 以特定Pattern（关键词、词组、匹配数量、日期等）建立文本到量化指标的映射
- 分离结构化数据
  - 公告往往是文字和表格混排，表格信息可以抽取
    - 股权结构、客户/供应商、销售数据、三大表及各种财务附注
    - 难点：不同公司的同类表格格式不同，难以对齐

技术：非结构化数据的处理

- 词向量/词典
  - 检索中的拓词（同义近义）
    - “投资人”、“出资人”、“股东”
  - 计算文档之间的相关性
    - 词向量 => 句子向量 => 文档向量
  - 正负面、利好利空
- 自动摘要
  - 提取公告/文档中的核心内容
    - CopyNet
- 特征提取
  - 命名实体：企业名称、人名等
  - 抽象：主题、事件、概念等

技术：PDF解析

- iTextSharp 转向 PDFBox
  - 基于Java，Apache许可
  - 同 iText 一样支持元素的坐标抽取，可以移植现有算法
  - 具有渲染PDF的能力，可以对无法解析的复杂表格进行渲染、截图
  - 面临问题：
    - PDFBox的文档和示例不够详细
    - 对中文字体的支持情况待验证（iTextSharp对繁体字支持较差）

技术：PDF解析

- PDF的版面识别
  - 分栏、表格（特别是无边/少边表格）、图表区域的区分和识别
  - 主要应用场景：
    - 港股公告的解析，涉及分栏、无边表格
    - 研究报告的解析，涉及分栏、表格、图表混排
- 扫描件的解析
  - 引用第三方插件：Abbyy OCR SDK
  - 神经网络图像识别/语言模型
    - 版面和表格识别是难点，初期可仅识别文字用于校对
  - 主要应用场景：A股公告10万、三板公告20万、债券公告全部

技术：ElasticSearch

- ElasticSearch 5.X版本升级
  - 支持Lucene 6.0
    - 磁盘空间少一半；索引时间少一半；查询性能提升25%
  - Java原生的REST客户端SDK
- ES分词器
  - 自定义分词器，平衡Recall和Precision
    - ik的recall相对较低，依赖字典补全，但分词准确率较高
    - nGram的recall较高，不需要字典，但容易有误命中
    - 停止词和标点的问题
- Kibana：数据可视化

技术：技术架构调整

- 分布式
  - 分布式文件系统/数据库，支持更多节点同时对一组文档/数据进行操作，提升性能
- 消息队列
  - 形成爬取、解析、索引、展现的业务pipeline，提升业务即时性
  - 为订阅推送功能做准备

产品：订阅功能

- 用户有订阅推送功能的需求
  - 对未来不确定时间信息的接收
    - 被动推送、及时、内容可定制
  - 待考虑问题：
    - 订阅主题是否有数量限制（考虑服务器负担）
    - 订阅主题是否完全由用户设置，是否会产生过多信息，如用户订阅“股东”
    - 产生过多信息时如何推送，如排序、筛选、折叠、时间段聚合

产品：数据源种类

- 其他披露类数据源
  - 港股公告、债券公告
- 其他相关数据源
  - 历史财务报表、交易数据
  - 研究报告
  - 财经新闻
  - 社交媒体（微博、微信公众号）
  - 裁判文书
  - 统计数据等.....

产品：数据关联

- 多维数据能够支撑复杂的关联，单一搜索无法做到
  - 目的：提示用户有价值的关联关系/推理链条
  - A公司和B公司可能因多种原因相互关联：
    - 母子公司、同一实际控制人、合资经营
    - 相同的高管或重要员工
    - 行业、板块、概念
    - 原材料/供应商、产品/客户
    - 诉讼相对方
    - ……

商业化应用场景

- 搜披露目前主要有两类使用人群
  - 资本市场中介服务，如投行、律所、会计师事务所、上市公司
    - 需求：为准备材料而参考前例
  - 一二级市场投资，如券商行研、基金
    - 需求：捕捉与公司经营相关的信息
- 此外，潜在的B2C场景
  - 券商经纪业务

商业化应用场景

- 资本市场中介服务领域
  - 核心诉求：更高效率的撮合交易
  - 商业化方向：
    - 内部的资源集成平台（KB、搜索引擎）
    - 电子化申报OA系统
    - 交易机会的发现：帮助上市公司寻找收购标的、搭建交易方案
    - 可充分利用中介机构拓展客户资源（找到上市公司），作交叉产品销售（舆情、竞争对手监测等产品）

商业化应用场景

- 一二级市场投资领域
  - 核心诉求：获取与资产价格变动相关的信息
  - 商业化方向：
    - 财经新闻/公告垂直搜索平台（对标Factiva，但国内付费意愿不强），用于铺设渠道
    - 付费的另类数据集成（自有或第三方），类似通联数据的思路
    - 如果相关性足够强，可以募集基金

商业化应用场景

- 券商经纪业务领域
  - 核心诉求：拉取个人开户、促进交易
  - 商业化方向：
    - 将搜披露包装为散户易接受的炒概念的版本，付费嵌入券商的终端中

Synergy

- 搜披露与天眼查的协同效应
  - 技术层面
    - 搜索
    - 自然语言处理技术
    - 知识图谱
  - 产品层面
    - 搜披露可为天眼查产品提供丰富的公司经营信息，保守估计信披中总共提及了超过100万家（Distinct）公司，这些公司与上市公司因各种关系共现
    - 天眼查可为搜披露提供公司/人关联的线索
  - 商务层面
    - 拓展券商、公募和私募基金等领域的应用场景
    - 为客户提供市场上最完整的产品服务

谢谢！