金融知识图谱自动构建与应用探索

2017.8

吴雪军、呼大为
wuxuejun@dingfudata.com
hudawei@dingfudata.com

提纲

- 公司背景介绍
- 知识图谱自动构建
- 知识图谱架构探索
- 知识图谱金融应用探索

公司背景介绍

- 鼎复数据
  - 面向二级市场金融机构（券商、公募、私募等）
  - 数据服务、提升工作效率、提升投资收益

- 基于金融知识图谱的技术服务
  - 数据结构化、图谱化
  - 图谱数据之上的智能应用

知识图谱自动构建

结构化数据

原始数据

图谱构建算法

知识碎片

逻辑校验

数据融合

金融知识图谱

产品图谱　人物图谱

公司图谱　事件图谱

应用平台

深度计算

智能预警

查询浏览

智能推理

选股

知识框架(Schema)

公司公告

研究报告

新闻资讯

机器学习

知识框架挖掘
逻辑构建

领域专家

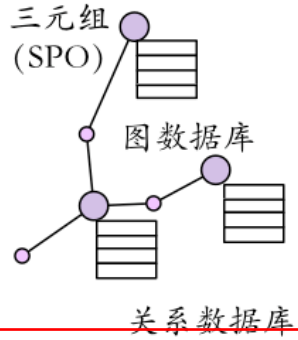
金融知识框架：事件框架、行业框架、财务框架

通用知识框架

知 识 框 架 管 理 平 台

图谱存储架构

融合数据库



功能优化

倒排索引　权限管理

增量更新

逻辑运算　容灾机制

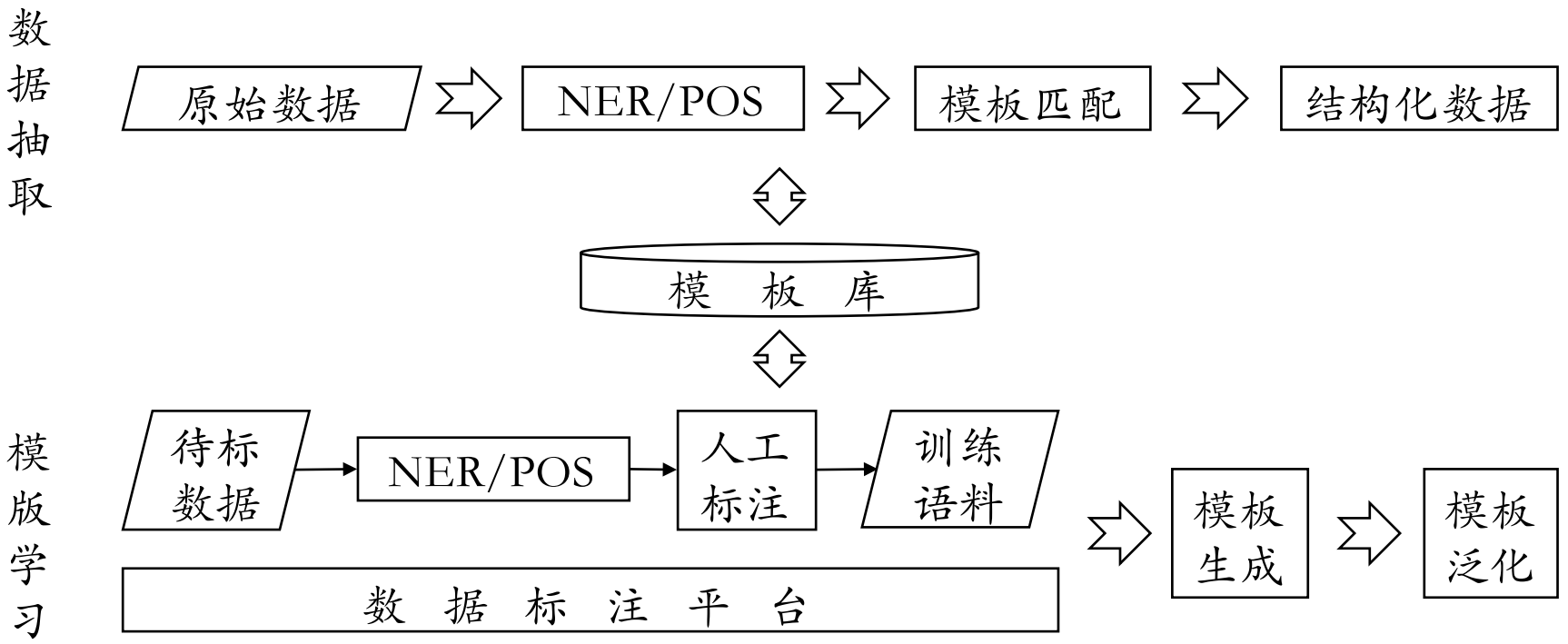
金融行业数据抽取

- 目标
  - 文本抽取结构化数据
- 数据抽取算法
  - 基于模板
  - 基于机器学习模型

算法分类	介绍	优点	缺点
基于模板	当数据匹配模板时进行知识抽取	准率高； 效果相对可控；	扩展性不好； 召回低；
基于机器学习模型	通过机器学习算法学习知识各因素关系	扩展性好； 召回较高；	准确率略低； 可控性略低；

金融行业数据抽取—基于模板

- 示意图



金融行业数据抽取—基于模板

- 模板
  - 实体标签组成
  - 类正则表达式
- 样例
  - <name> <gender> <resign_reason> <resign_indicator> .* <job>
- NER/POS
  - 粒度
    - 鼎复数据科技（北京）有限公司（以下简称“本公司”）
  - 模型
    - CRF＋词典

金融行业数据抽取—基于模板

- 模板学习

甲某某 先生 因工作原因 申请 辞去 公司 监事会主席 职务
乙某某 女士 因个人原因 申请 辞去 公司 独立董事
丙某某 先生 因家庭原因 请求 辞去 证券事务代表 一职

模板自动挖掘结果

<name> <gender> * <job> *

词表挖掘+校正

<name> <gender> <resign_reason> <resign_indicator> .* <job>

金融行业数据抽取—基于模板

- 模板匹配

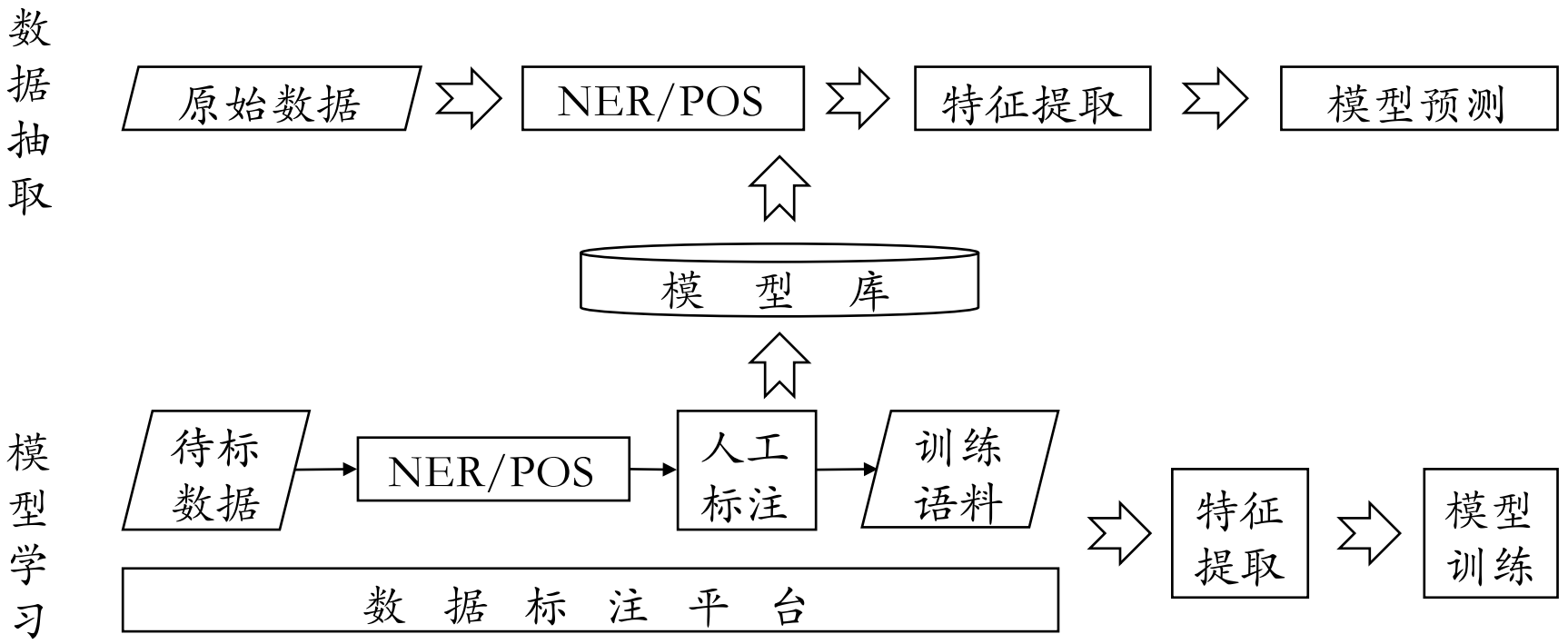
<name> <gender> <resign_reason> <resign_indicator> .* <job>

甲某某 先生 因工作原因 申请 辞去 公司 监事会主席 职务

属性名称	值
姓名	甲某某
性别	男
辞职原因	因工作原因
辞职岗位	监事会主席

金融行业数据抽取—基于机器学习模型

- 示意图



金融行业数据抽取—基于机器学习模型

- 模型
  - 表格拆解为关系
  - 分类模型

时间	主体	属性	值
2016年	公司	研发投入总额	30,004.5万元

- 样例
  - 2016年公司研发投入总额30,004.5万元,占营业收入的比重较去年同期减少1.22%
  - R（研发投入总额，30,004.5万元）、R（研发投入总额，1.22%）
  - $C_{属性-值}$
- 特征设计
  - 文本特征
  - 位置特征
  - 句法特征

金融行业数据抽取—基于机器学习模型

样本　2016年公司研发投入总额30,004.5万元,占营业收入的比重较去年同期减少1.22%,其中费用化支出也有很大的变化，达到22,427.28万元,占研发投入的比例同比增长27.60%

标注　$C_{属性-值}$
R（研发投入总额，30,004.5万元）:1　R（研发投入总额，1.22%）:0
R（费用花支出，22,427.28万元）:1 R（费用化支出，27.60%）:0

特征
文本特征：文本长度，分词term数等
位置特征：av_dis, av_v_conflict, av_a_conflict等
句法特征：句法距离，句子成分距离等

模型
混合
rf模型
lr模型
gbdt模型

金融行业数据抽取—基于机器学习模型

- 原文
  - 2016年公司研发投入总额30,004.5万元,占营业收入的比重较去年同期减少1.22%,其中费用化支出22,427.28万元,占研发投入的比例同比增长27.60%
- 关系预测
  - R（研发投入总额，30,004.5万元） R （研发投入总额，1.22%）
  - R（费用花支出，22,427.28万元） R （费用化支出，27.60%）
  - $C_{属性-值}$
    - R（研发投入总额，30,004.5万元）：1
    - R（费用花支出，22,427.28万元）：1
- 结构化数据

时间	主体	属性	值
2016年	公司	研发投入总额	30,004.5万元
2016年	公司	费用化支出	22,427.28万元

知识图谱架构探索

- 基本存储格式
  - 三元组（SPO）
- 数据分片
- 倒排索引
  - 主语、宾语
- 融合数据库
  - 时序数据
- 实时更新

知识图谱架构探索

- 图谱架构主要包含存储和查询两个部分
  - 以三元组（SPO）为基本存储格式

subject	predicate	object
df:000c2b8928a	df:bad_prep_total	df:3758408b6cba

  - 支持 SPARQL 1.0 查询

```
select ?friend where {
  ?var <person.name> "lisi" .
  ?var <person.friend> ?friend .
  ?friend <person.last_name> "zhang" .
}
```

知识图谱架构探索

- 三元组数据分片存储
  - 根据 Predicate 分片

知识图谱应用架构探索

- 分片内建倒排索引
  - Subject, Object
  - String 到分片行的映射

- 支持表结构存储
  - 某些数据以表结构存储性能更好，比如时序数据

公司	时间	净利润	营业收入
招商银行	2016-12-31	1000000000	2000000000
万科	2016-12-31	1000000000	2000000000
民生银行	2015-12-31	1000000000	2000000000

知识图谱金融应用探索

- 金融知识图谱特点
  - 准确率高
  - 覆盖广
  - 实时性强
  - 逻辑性强

- 产品
  - 公司图谱
  - 产品图谱
  - 人物图谱
  - 选股＋智能预警

公司图谱（持股）

工商数据

公司财报十大股东

基金持仓

增发配股

高管增减持

处置资产

举牌信息

产品图谱

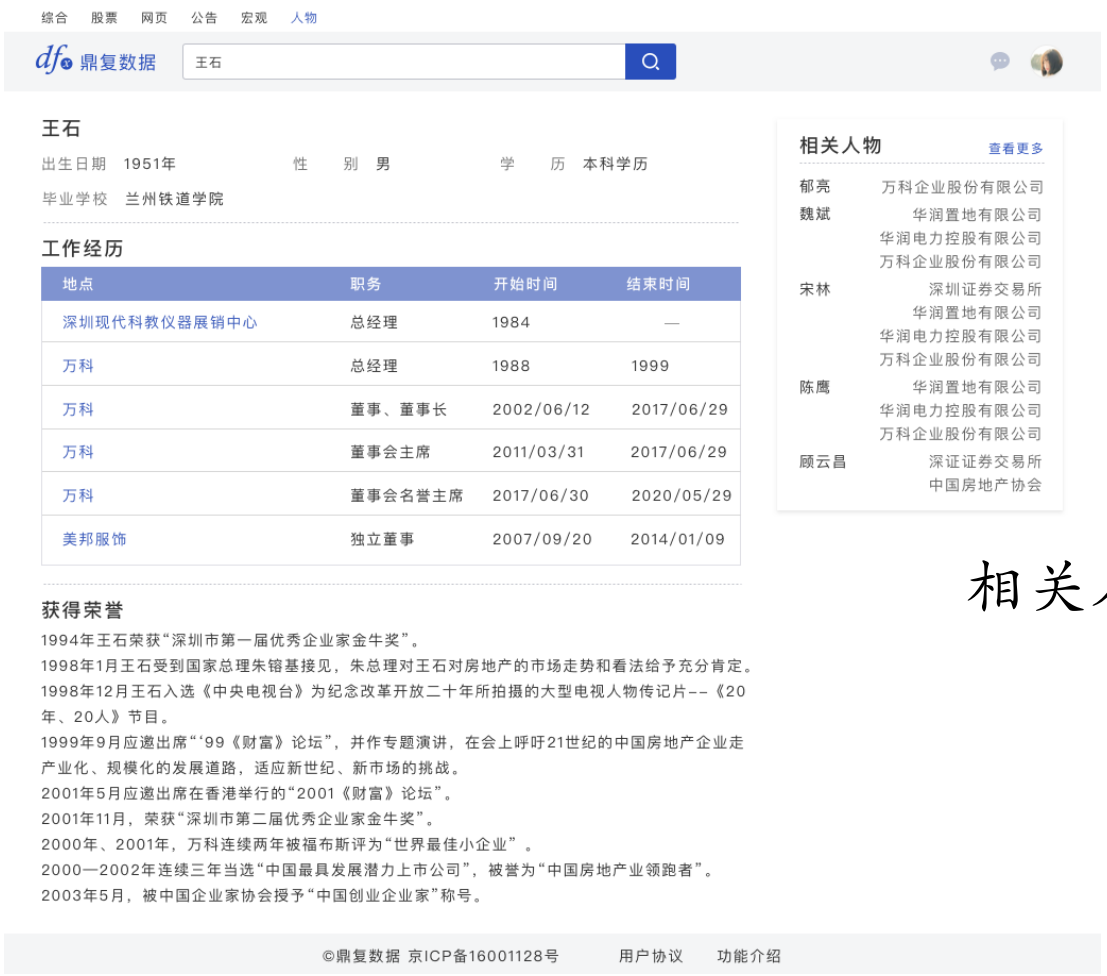
产品相关数据

产品相关公司

人物图谱

相关公司

职位履历



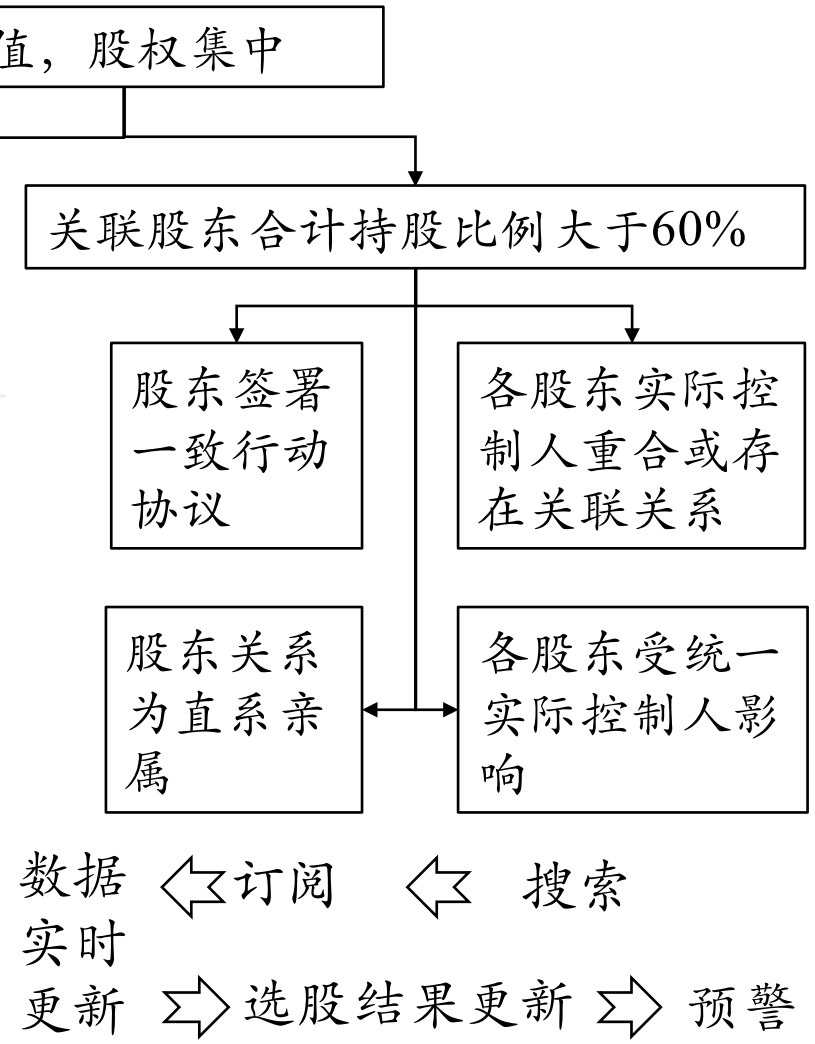
相关人物

选股＋智能预警

人物图谱

逻辑分析　搜索技术

股票代码	股票简称	2017-08-18市值	市盈率(ttm)	市净率	收盘价
601398	工商银行	2,003,003,164,902.00	7.17	0.98	5.62
601939	建设银行	1,682,573,878,575.00	7.21	1.02	6.73
601857	中国石油	1,455,016,773,510.00	53.13	1.05	7.95
601288	农业银行	1,182,250,585,880.00	6.39	0.87	3.64
601988	中国银行	1,177,551,164,800.00	7.15	0.77	4.00
601318	中国平安	971,960,435,238.00	14.94	1.82	53.17
601628	中国人寿	799,608,504,450.00	39.93	2.55	28.29
600028	中国石化	716,741,560,832.00	12.61	0.84	5.92
600036	招商银行	641,088,475,152.00	9.70	1.52	25.42
600519	贵州茅台	615,097,252,770.00	32.09	7.78	489.65
601328	交通银行	463,399,413,984.00	6.87	0.71	6.24
601088	中国神华	384,267,468,060.00	12.67	0.97	19.32
601166	兴业银行	355,030,920,772.00	6.46	0.89	17.09



高级C++开发工程师

自动化测试工程师

金融分析师

高级NLP工程师

前端开发工程师

前端开发工程师

谢谢

销售副总裁

2017.8

数据专员

Java/PHP开发工程师

招贤纳士

资深UI设计师

请将简历发送到

talents@dingfudata.com

C＋＋/PHP/JAVA研发实习生

产品经理