

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

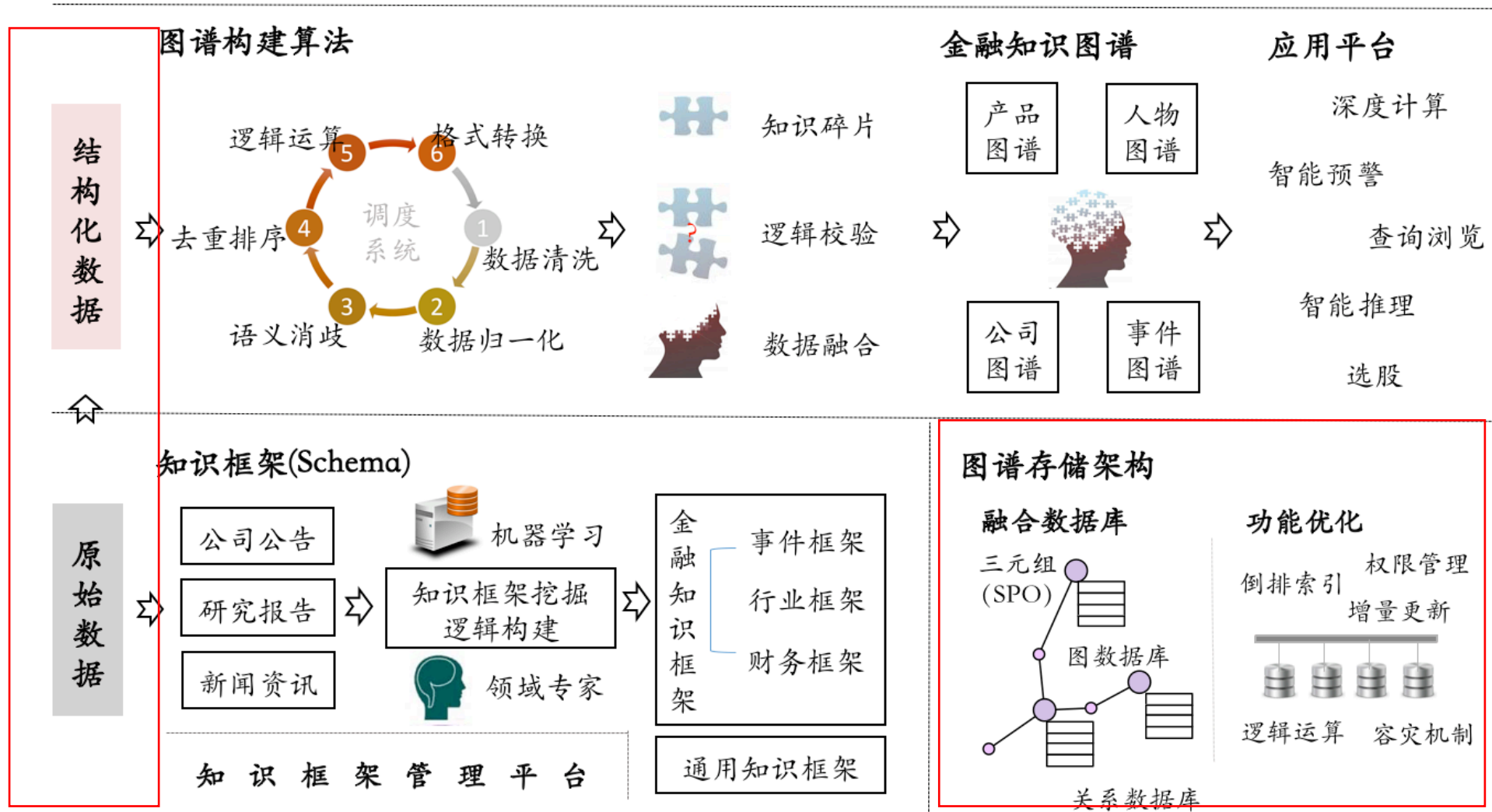
提纲

- 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



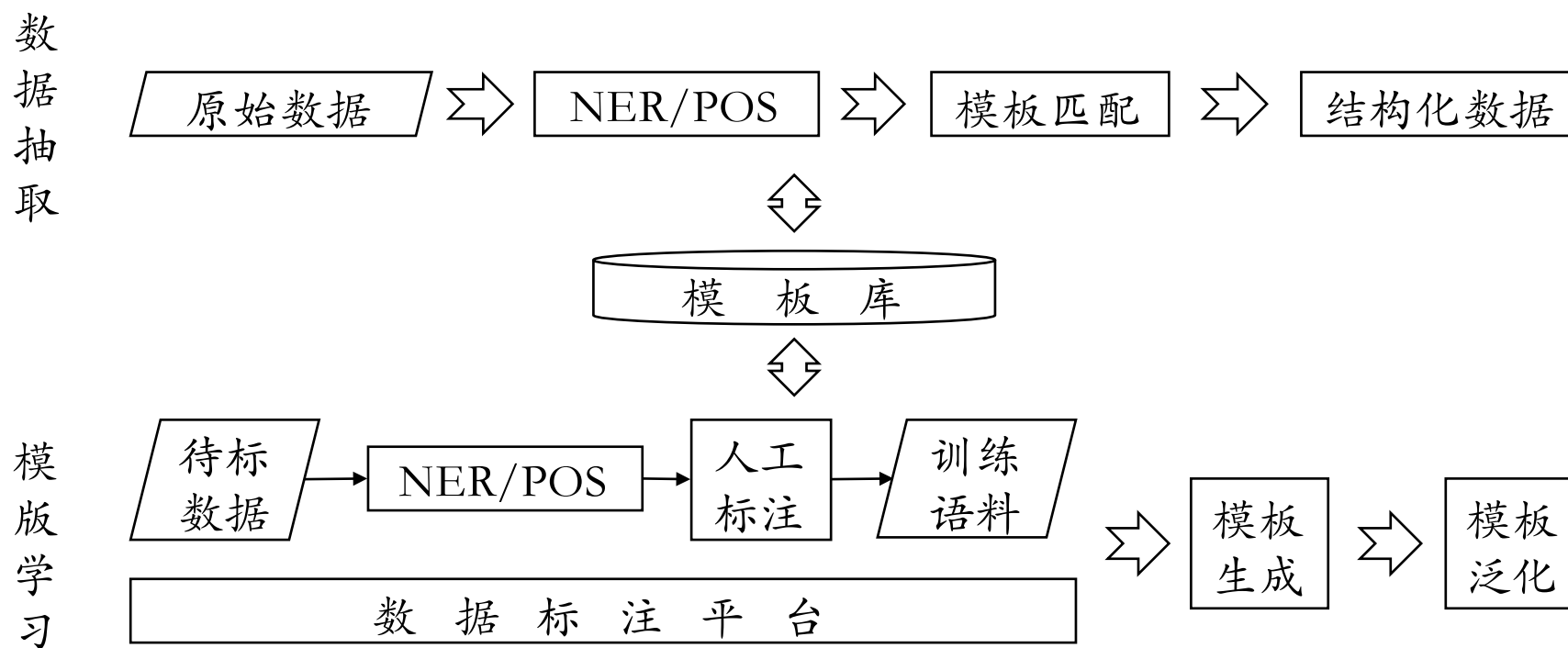
金融行业数据抽取

- 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>

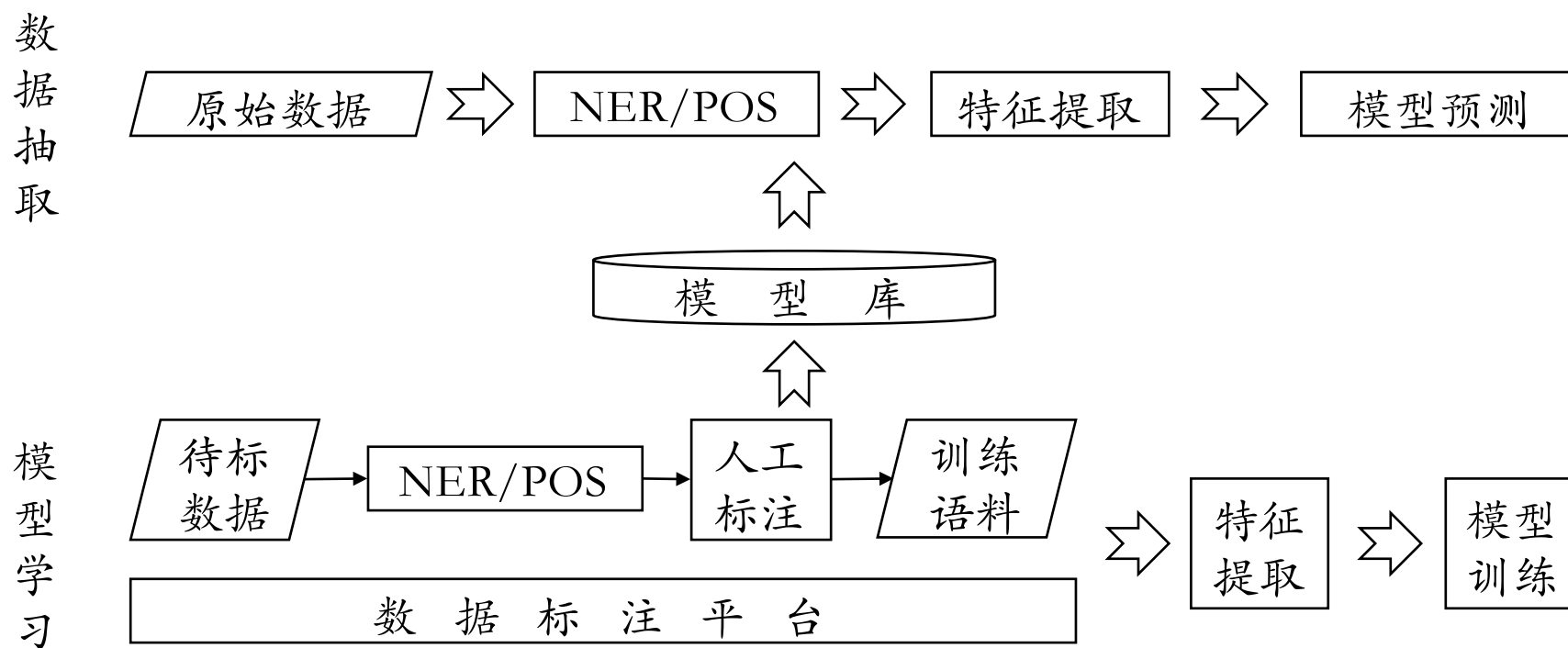
甲某某 先生 因工作原因 申请 辞去 公司 监事会主席 职务



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%）
- $C_{\text{属性-值}}$

- 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 $C_{\text{属性-值}}$ R (研发投入总额, 30,004.5万元) :1 R (研发投入总额, 1.22%) :0
R (费用化支出, 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 (费用化支出, 22,427.28万元) R (费用化支出, 27.60%)
- C_{属性一值}
 - R (研发投入总额, 30,004.5万元) : 1
 - R (费用化支出, 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	10000000000	20000000000
万科	2016-12-31	10000000000	20000000000
民生银行	2015-12-31	10000000000	20000000000

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

宏观结果			
农业:牲畜饲养:生猪存栏头数	农业:牲畜饲养:生猪存栏头数:能繁母猪	农业:牲畜饲养:生猪定点屠宰量	出场价格:生猪
生产资料市场价格:农产品:生猪:外三元			

产品上下游

农业
大豆
玉米
疫苗
豆粕
饲料
种猪

生猪						
公司	◆ 主营业务产品	◆ 产品划分方式	◆ 主营业务收入	◆ 收入占比	◆ 毛利率	◆ 报表日期
牧原股份	生猪	按产品(项目)分类	3002995668.79	99.98	24.61	2015-12-31
罗牛山	生猪	按产品(项目)分类	237011917.38	32.47	19.91	2015-12-31
大康农业	生猪	按产品(项目)分类	149855729.99	3.88	38.52	2015-12-31
正邦科技	生猪	按产品(项目)分类	476814895.62	6.61	10.03	2013-06-30
新五丰	生猪	按产品(项目)分类	89585600.00	100.00	10.85	2006-09-30
顺鑫农业	生猪屠宰业	按行业分类	2724888567.42	28.35	7.44	2015-12-31
雏鹰农牧	生猪产品	按产品(项目)分类	1497736079.56	41.39	19.24	2015-12-31
雏鹰农牧	生猪屠宰业	按行业分类	1008937025.69	27.88	9.54	2015-12-31
天邦股份	生猪养殖	按产品(项目)分类	651899363.08	30.44	0.00	2015-12-31
新五丰	生猪内销	按产品(项目)分类	411563783.41	31.05	11.94	2015-12-31

养殖
屠宰加工
生猪养殖
硫酸软骨素
肉制品
肝素粗品
肝素钠原料药

产品相关公司

人物图谱

相关公司

职位履历

相关人物

综合

股票

网页

公告

宏观

人物

df 鼎复数据

王石

Q

...

王石

出生日期

1951年

性 别

男

学 历

本科学历

毕业学校 兰州铁道学院

工作经历

地点	职务	开始时间	结束时间
深圳现代科教仪器展销中心	总经理	1984	—
万科	总经理	1988	1999
万科	董事、董事长	2002/06/12	2017/06/29
万科	董事会主席	2011/03/31	2017/06/29
万科	董事会名誉主席	2017/06/30	2020/05/29
美邦服饰	独立董事	2007/09/20	2014/01/09

获得荣誉

1994年王石荣获“深圳市第一届优秀企业家金牛奖”。
1998年1月王石受到国家总理朱镕基接见，朱总理对王石对房地产的市场走势和看法给予充分肯定。
1998年12月王石入选《中央电视台》为纪念改革开放二十年所拍摄的大型电视人物传记片——《20年、20人》节目。
1999年9月应邀出席“‘99《财富》论坛”，并作专题演讲，在会上呼吁21世纪的中国房地产企业走产业化、规模化的发展道路，适应新世纪、新市场的挑战。
2001年5月应邀出席在香港举行的“2001《财富》论坛”。
2001年11月，荣获“深圳市第二届优秀企业家金牛奖”。
2000年、2001年，万科连续两年被福布斯评为“世界最佳小企业”。
2000—2002年连续三年当选“中国最具发展潜力上市公司”，被誉为“中国房地产业领跑者”。
2003年5月，被中国企业家协会授予“中国创业企业家”称号。

相关人物

查看更多

郁亮

万科企业股份有限公司

魏斌

华润置地有限公司
华润电力控股有限公司
万科企业股份有限公司

宋林

深圳证券交易所
华润置地有限公司
华润电力控股有限公司
万科企业股份有限公司

陈鹰

华润置地有限公司
华润电力控股有限公司
万科企业股份有限公司

顾云昌

深证证券交易所
中国房地产协会

©鼎复数据 京ICP备16001128号

用户协议

功能介绍

选股+智能预警

人物图谱

逻辑分析 搜索技术

df 鼎复数据

高市值



市值 ▾

大于100亿

在股票中为您找到了1171条数据

更多操作, 请到鼎复智能投研云平台

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

高市值, 股权集中

市值大于100亿

关联股东合计持股比例大于60%

股东签署
一致行动
协议

各股东实际控
制人重合或存
在关联关系

股东关系
为直系亲
属

各股东受统一
实际控制人影
响

数据 实时 更新
订阅 搜索
选股结果更新 预警

df 鼎复数据

高级C++开发工程师

自动化测试工程师

高级NLP工程师

金融分析师

前端开发工程师

前端开发工程师

谢谢

销售副总裁

数据专员

2017.8

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

