Geekbang 极客邦科技 InfoQ

通联数据 www.datayes.com
让投资更容易

QCon 全球软件开发大会 【北京站】2016

基于大数据的机器学习在金融投资行业的应用

蔡弘， 博士
CTO， 通联数据股份公司

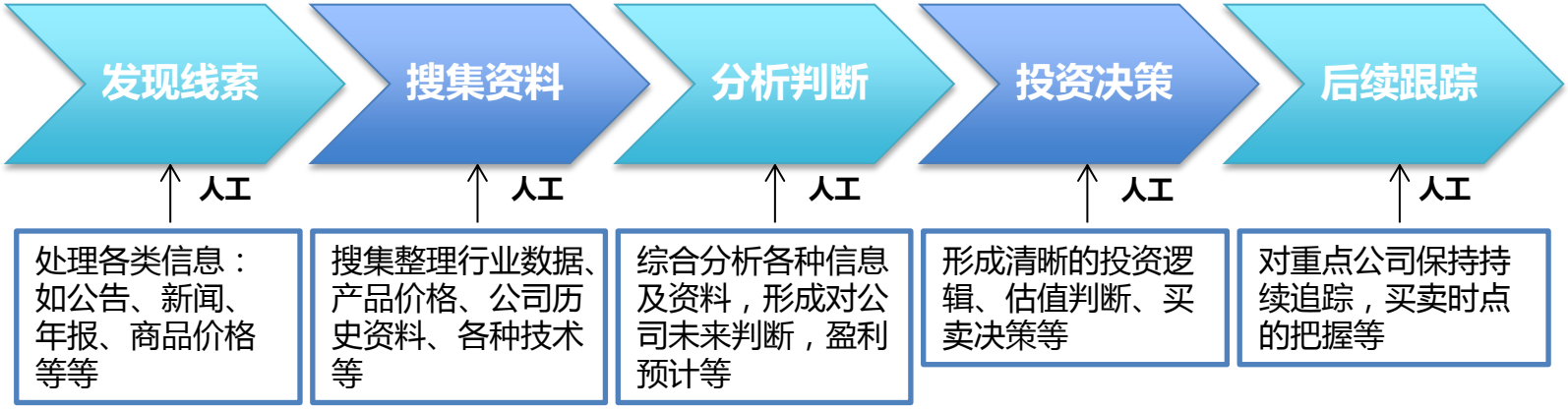
Intemational Software Development Conference

1 投研领域的挑战

投研管理业务场景及痛点

通联数据 www.datayes.com 让投资更容易

1、基本面投资流程很长、需处理的信息量极大、种类繁多



2、仅依靠人工很难跟踪来自新闻、公告、研报、社交网络等各种来源的海量信息

- A股上市公司已经超过2600家，注册制推出后上市公司数量还将大幅增加
- 新闻网站、微博/雪球大V、微信公众号、论坛、股吧等社交媒体信息
- 电商数据、招聘信息、司法、诉讼，行业上下游关联等非结构化数据

挑战也是机会：

大数据如何变成投研团队可以迅速吸收并用于投资活动的小数据？

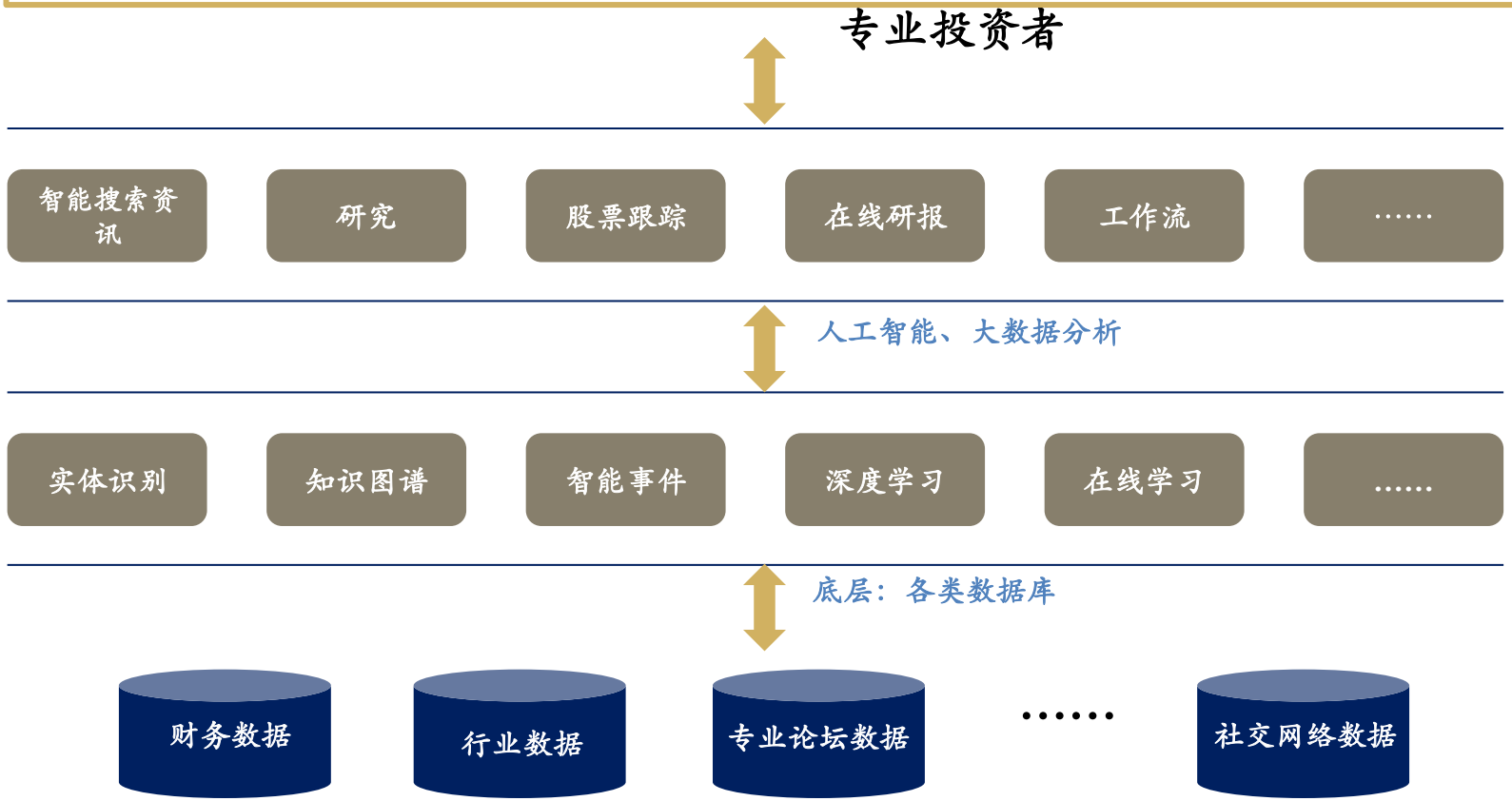
重点是：如何把与证券有关信息中的“关联关系”展现出来

2 架构

通联智能投研平台架构

通联数据 www.datayes.com
让投资更容易

通联智能投研平台定位于构建一个开放、分享、高效的基本面投研平台，通过**自然语言处理和机器学习等技**术，高效地从海量的信息中提炼对研究员有价值的信息；同时，该平台实现投研流程中的过程数据和结果数据管理，满足客户在证券研究过程中对信息响应、研究协作的迫切需求，使得碎片化的研究成果得以沉淀积累，为投资决策提供重要支持。



通联投研平台主要功能

通联数据 www.datayes.com
让投资更容易

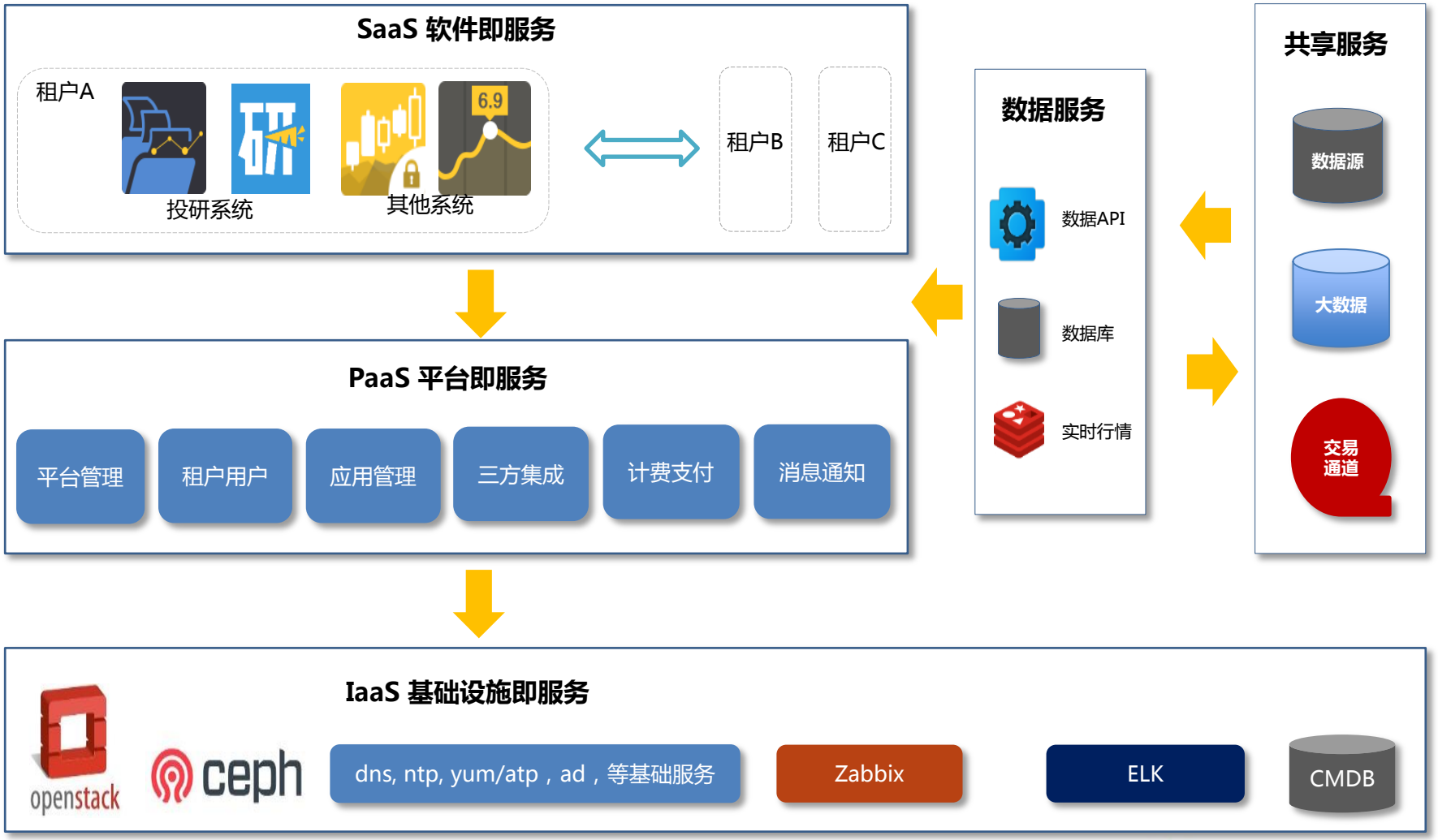
主要功能

资讯中心	股票管理	研究中心	监控中心	个人空间
• 新闻资讯 • 公司公告 • 微信订阅 • 分享中心	• 私有股票池 • 公共股票池	• 内部研究 • 晨会研究 • 外部研究	• 宏观监控 • 行业监控 • 我的监控 • 内部监控	• 我的研究 • 我的收藏 • 我的笔记 • 自定义数据

大数据分析 + 研究过程与结果沉淀 + 智能投研分析工具 = 深入洞察行业、创造卓越价值

通联投研平台SaaS服务架构

通联数据 www.datayes.com
让投资更容易



3 投研分析中的机器智能揭秘

通联数据投研平台机器学习技术框架

通联数据 www.datayes.com
让投资更容易

投研产品策略层
- 智能研报服务
- 智能提醒服务

机器学习技术应用层
- 深度学习技术
- 增强学习技术

交易策略算法层
- 索引行情、状态识别
- 自动化学习策略

搜索推荐服务层
- 搜索相关技术
- 推荐相关技术

机器学习技术基础层
- 特征抽取、聚类技术
- 关联分析、排序技术

大数据分析方法论
- 回测技术框架
- 分析实验设计

数据采集层
- 爬虫技术
- 数据清洗与质控

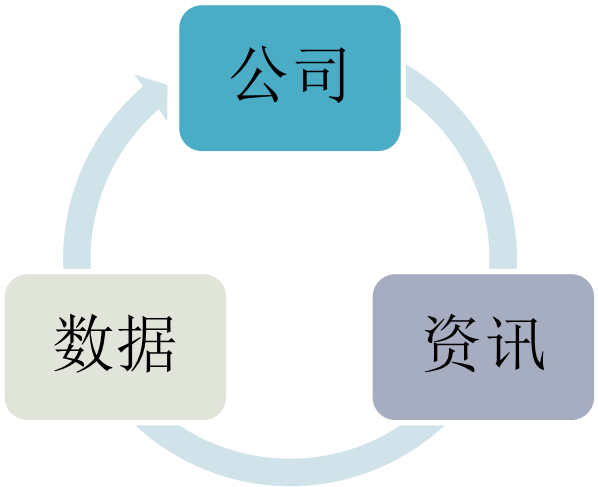
数据整理层
- 分类、标记
- 实体识别、事件抽取

逻辑整理层
- 知识图谱
- 事件序列

通联数据机器学习

通联数据 www.datayes.com
让投资更容易

投资研究的目的是选择合适的投资标的



公司⇔数据
公司数据推荐
公司财务模型

公司⇔资讯
公司资讯推荐
公司情感分析
产品数据挖掘

数据客观描述宏观/行业/公司景气程度
数据可视化
数据搜索
数据清洗

数据⇔资讯
资讯推荐数据
行业数据提取

从资讯中寻找投资线索
资讯搜索
资讯聚类
资讯分类
资讯去重

资讯：公司资讯推荐

通联数据 www.datayes.com
让投资更容易

背景：浏览某个公司相关的资讯

挑战：实体语义消歧

苹果与苹果公司？

> 据工信部网站消息，12月18日-19日，由中国机器人（300024）产业联盟、中国电子信息产业发展研究院、广州工业机器人制造和应用产业联盟……

提到还是相关？

> 东吴证券(13.550, -0.16, -1.17%)分析师徐力认为，中国联通与电信合作的红利将会逐渐体现，成本优势将愈加明显

资讯：公司资讯推荐

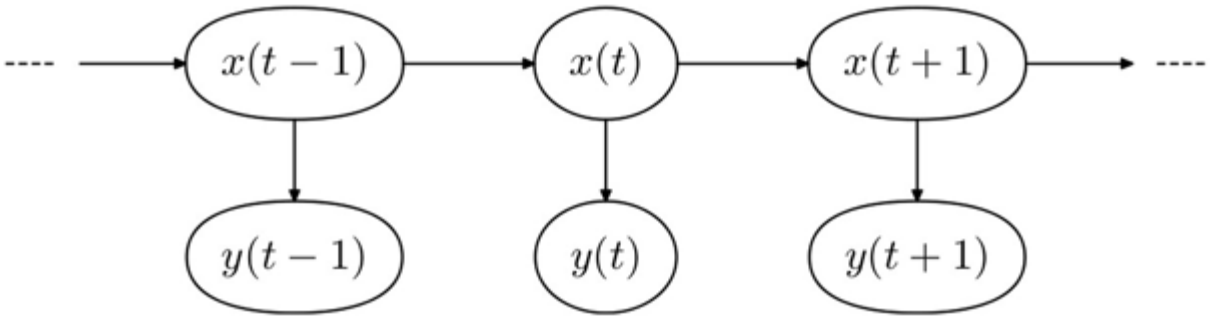
苹果与苹果公司？

据工信部网站消息，12月18日-19日，由中国机器人（300024）产业联盟、中国电子信息产业发展研究院、广州工业机器人制造和应用产业联盟……

方法：NER（命名实体识别）；提到的是一个公司还是一个普通词组

效果：解决bad case中60%的例子，包括常见上市公司名称，如机器人、农产品、新能源

常见的NER方法：隐马尔科夫模型



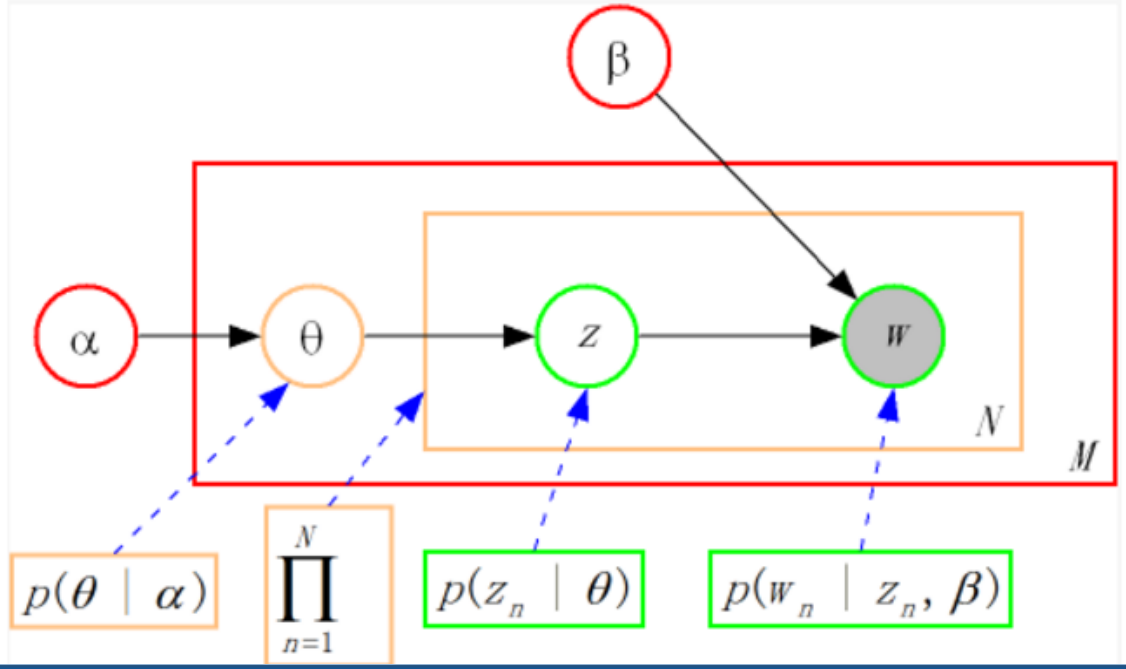
资讯：公司资讯推荐

通联数据 www.datayes.com 让投资更容易

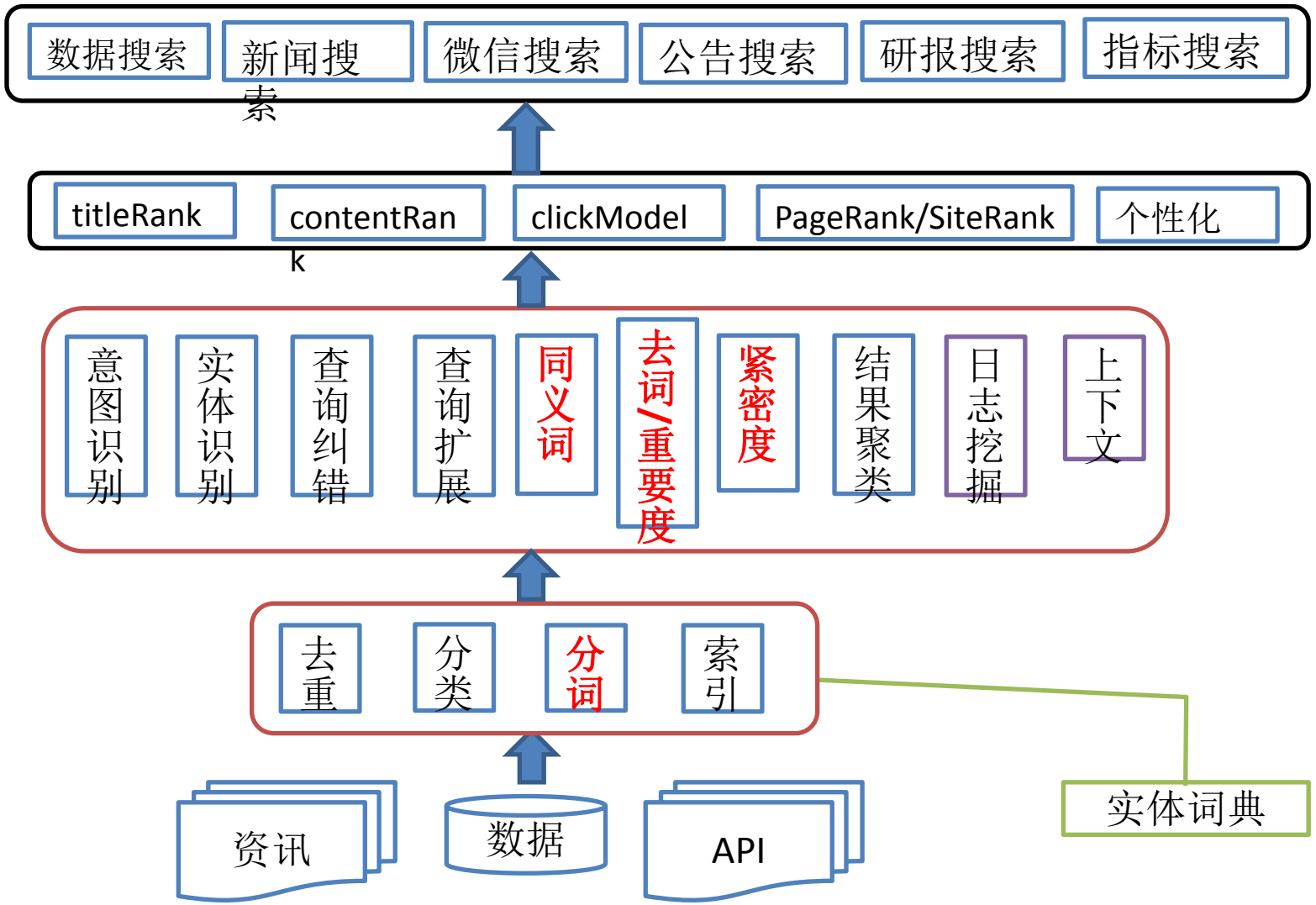
提到还是相关？

> 东吴证券(13.550, -0.16, -1.17%)分析师徐力认为，中国联通与电信合作的红利将会逐渐体现，成本优势将愈加明显

方法：LDA（主题模型）；提到的新闻和公司是不是相同主题
效果：基本上可以去掉证券类公司、网站类公司的问题，占bad case 30%



智能搜索



智能搜索：分词

查询词/query：a股市场的上市公司

正确的分词/terms：a股、市场、的、上市、公司

错误的分词/terms：a、股、市场、的、上市、公司

智能搜索：紧密度

通联数据 www.datayes.com 让投资更容易

- 查询词：a股市场的上市公司
- 分词结果：a、股、市场、的、上市、公司
- 紧密度：（a、股）、市场、的、（上市、公司）
- a和股是紧密的，所以a和股在搜索结果中必须连续出现

- 紧密度是分词的延伸
- 高级别紧密的term已由分词解决。例如：中国、苹果、手机
- 紧密度解决：（中国、银行）（通联、数据）（荷兰、 猪）

应收	账款	5.358284
粗	钢产量	4.017519
沪	深	3.931121
液化	石油气	3.13324
股份	有限公司	3.117022
三井	住友	3.09504
萝	岗区	3.06203
大气污染	防治	3.005964
消费品	零售总额	2.966966
保税	港区	2.458586
动力	煤	2.38944
耐用	消费品	2.096894
融资	融券	2.001254
占	比	1.988273
减值	准备	1.906077
线型	低密度	1.882662
同业	拆借	1.876567
街道	办事处	1.855189
桑	蚕茧	1.792127
意外	伤害	1.757629

智能搜索：紧密度

通联数据 www.datayes.com 让投资更容易

方法：词语连接测度(Symmetric Conditional Probability and Context Dependency，SCPCD)

- 一个词组在文中出现的前缀/后缀数量越多样，它和其它词形成固定搭配的可能性越小，SCPCD越大
- 一个词组被拆分后，拆分的两个部分在文章中出现的次数与词组本身出现的次数一致，则SCPCD越大

智能搜索：重要度

通联数据 www.datayes.com 让投资更容易

重要度：区分query中不同term的重要程度，降低冗余词、停用词等在匹配候选搜索结果的权重

查询词：全国猪肉的平均价格

重要词：全国猪肉的平均价格

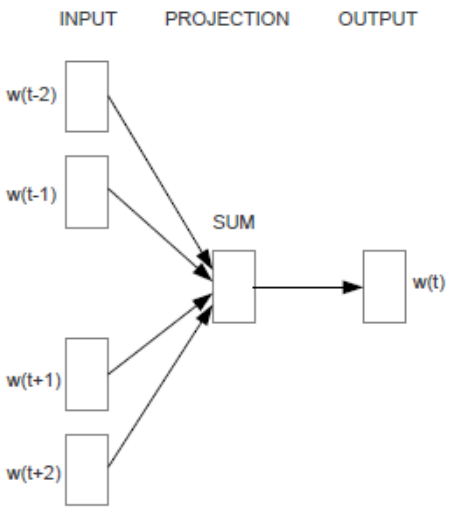
方法：query结构、term自身以及与query的信息（位置、词性、长度）、全局统计信息（language model）

智能搜索：同义词

通联数据 www.datayes.com
让投资更容易

生猪	猪
汽车	小汽车
轿车	小轿车
物价	价格
铜矿	铜矿石
进口额	进口
出口额	出口
进口额	进口量
出口额	出口量
北京市	北京
上海市	上海

同义词发现
word2vec



我们在招聘，欢迎志同道合的伙伴！

通联数据 www.datayes.com
让投资更容易

- 前端开发工程师
- 前端架构师
- 移动端开发工程师
- 交易系统开发工程师
- Java前端开发工程师
- Java后端开发工程师
- ETL开发工程师
- 云平台架构师(Openstack, Docker, Spark)
- 测试开发工程师
- 性能测试和调优工程师

- 大数据分析工程师
- 搜索算法专家和工程师
- 机器学习算法专家和工程师

（精通机器学习（SVM、LR、AdaBoost），数据挖掘（Apriori、决策树、随机森林，了解深度学习（CNN、LSTM等）或者知识图谱相关理论）

- 自然语言处理专家和工程师

（熟悉常用的自然语言处理方法，包括但不限于HMM、CRF、word2vec）

- 爬虫开发工程师

- 投研产品经理
- 大数据可视化用户体验设计师

- DEVOPS

(支持IaaS（Openstack/Ceph）和Container（Docker/Ceph）应用部署环境)

- 开发运维工程师

(负责应用系统相关高可用设计、监控、升级部署、应用配置修改、日志收集与分析等工作，并尽可能实现运维自动化)

- 数据运维工程师

(参与相关应用项目的ETL设计、开发、维护工作)

- 数据库工程师

(MySQL, SQL Server)

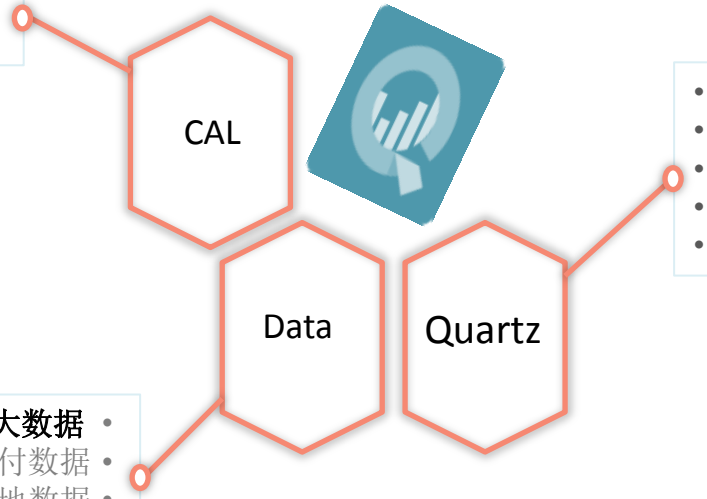
- 信息安全工程师
- 信息系统工程师

优矿（uqer)： 用python快速验证投资想法

通联数据 www.datayes.com
让投资更容易

欢迎大家到38号展台体验通联产品

金融计算分析库
- 权益/固定收益及衍生品建模
- 中国市场定制
- 强大的定价工具



- 支持各类型的股票量化策略
- 策略回测
- 策略表现评估
- 无需关心底层实现
- 更多策略框架添加中

海量金融大数据
- 覆盖市场行情、财报、宏观、电商支付数据
- 自定义的本地数据

如何在优矿上一个人干掉一家公募量化团队？Alpha！Go！

扫描查看源码

Geekbang 极客邦科技 InfoQ

THANKS!

通联数据
www.datayes.com

让投资更容易

Intemational Software Development Conference