知识图谱，金融大数据治理的基石

证券行业知识图谱应用分享

夏磊

围绕“知识”的提取和创造
以此辅助企业决策

2009年
成立于上海

2011年
穆迪入股

2015年
京东金融
入股

2016年
KPMG Top
Fintech 50

目录

PART 1

为什么要构建知识图谱

金融机构、数据驱动型企业面临的难题

分析师

- 信息收集所需时间长，尤其在文本信息方面，投入精力大
- 个体知识有局限
- 分析对象之间的隐含关系不易察觉

机构

- 人力投入大，需要不同行业的分析师以覆盖主流行业
- 分析师之间的个体差异，增加了信息流转的壁垒
- 分析师的个体经验难以留存和管理

通过“知识图谱”来治理证券大数据

通过“知识图谱”来治理证券大数据

结构化

从非结构化数据到结构化

非结构化数据
- 公告挖掘
- 资讯挖掘

半结构化数据
- 公告挖掘
- 财务科目
- 产品容错表
- 命名实体对齐

标准化

元数据定义、字典、标签

元数据定义、字典定义
- 数据接入
- 清洗
- 标准化

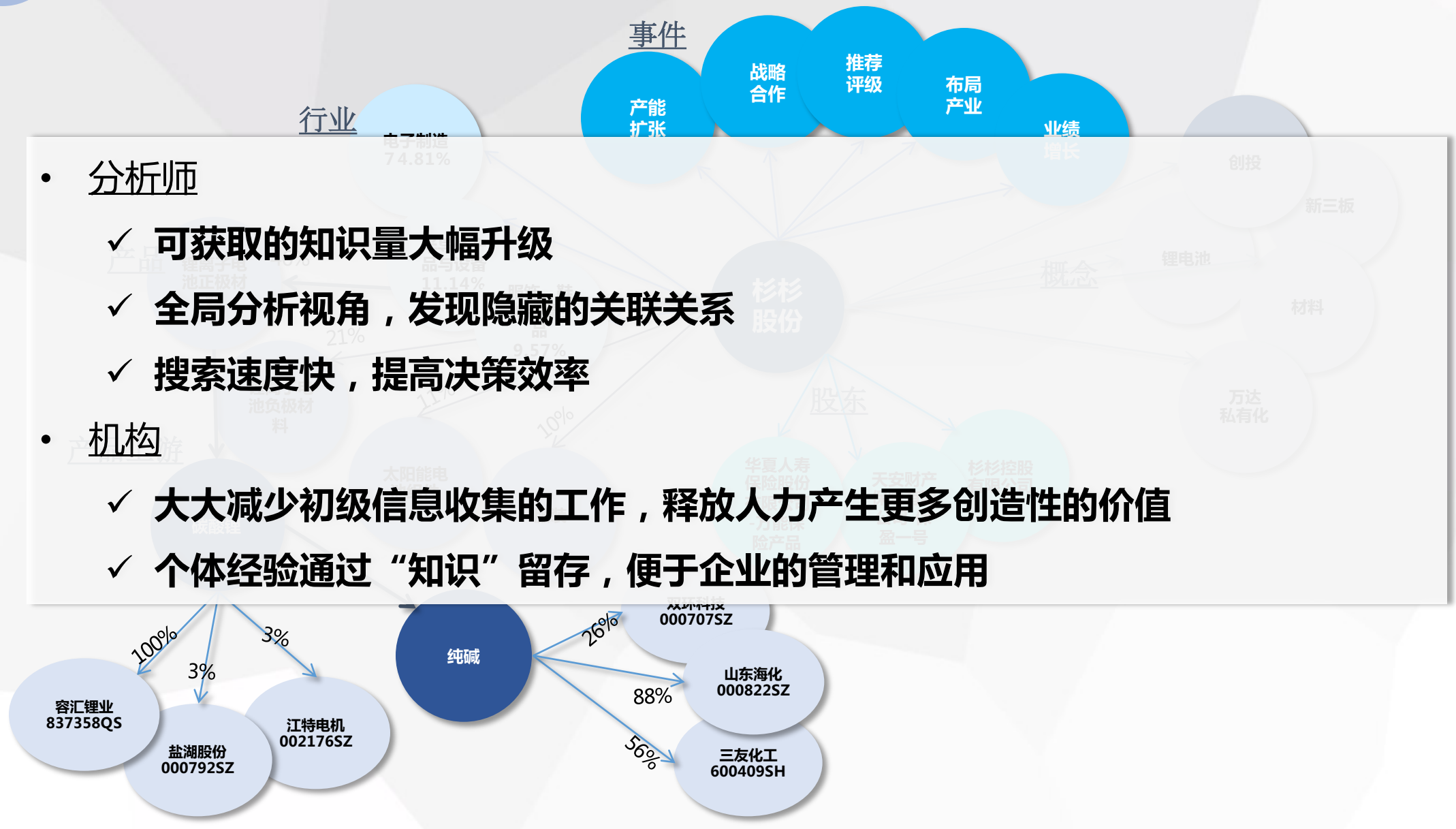
标签体系
- 产品标签
- 用户标签
- 交易标签

关联性

数据关联产生价值

产品上下游关系
事件传导关系

知识图谱给金融企业带来的价值



- 分析师
  - ✓ **可获取的知识量大幅升级**
  - ✓ **全局分析视角，发现隐藏的关联关系**
  - ✓ **搜索速度快，提高决策效率**
- 机构
  - ✓ **大大减少初级信息收集的工作，释放人力产生更多创造性的价值**
  - ✓ **个体经验通过“知识”留存，便于企业的管理和应用**

知识图谱在企业技术架构中发挥的重要作用

PART 2

数库如何构建知识图谱

——以证券行业为例

构建知识图谱——挖掘目标

结构化

- 数据挖掘
  - 数据库
  - 时间序列

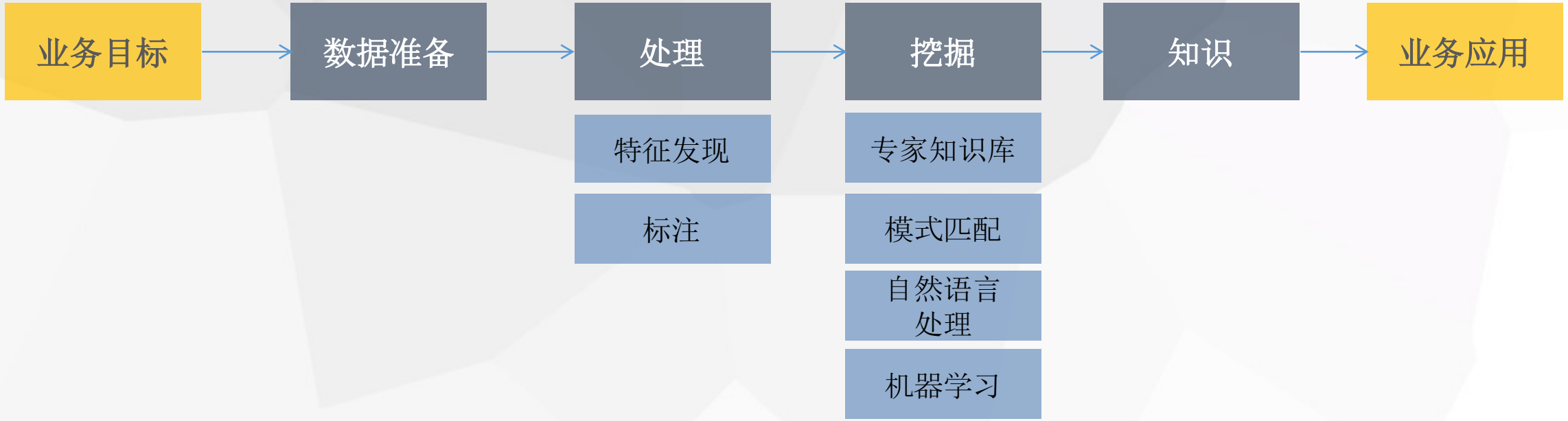
半结构化

- 过程挖掘
  - 日志文件
- 标准挖掘
  - 不同公司披露的财务和产品数据

非结构化

- 文本挖掘
  - 公告
  - 研报
  - 网站新闻
  - 社交媒体

构建知识图谱——挖掘流程



构建知识图谱——关键技术

实体抽取

- 命名实体识别：
人名、机构名、行业、地名、时间

事件提取

- 主体、地点、时间、原因、后果

多义词消歧

- Ngram、word2vec、潜语义分析 LSA/LSI
- 分类算法

关系挖掘

- 实体与事件关系
- 产业链上下游：规则、句法分析、深度学习

聚类/分类算法

- 层次聚类、密度峰值聚类
- SVM、CNN

搜索引擎及UGC标签

构建知识图谱

存储

- **关系型数据库-结构化**：mysql
- **非关系型数据库-非结构化**：mongodb、hbase、ElasticSearch
- **图数据库-查询和推理**：neo4j

工具

- **数库挖掘工具 CSTG**
  - ✓ 词库、规则库
  - ✓ 专家知识库
  - ✓ 标注
  - ✓ 图谱
- **数库机器学习平台 CSMLP**
  - ✓ 预处理
  - ✓ 机器学习+自然语言处理算法
  - ✓ 模型训练
  - ✓ 测试

PART 3

从数据中提取知识
——实例：上市公司产业链挖掘

实例：产业链挖掘

目标

- ***从公告、研究报告中挖掘出产业链上下游相关信息***

挖掘流程

- 样本选择：如从某个行业的研究报告开始
- 数据准备：选择含有产业链特征的段落，标注集，词库
- 算法：自然语言处理算法（分词、句法分析、特征分析）
- 验证测试

方法

- 基于规则和正则表达式
- 基于LTP分析
- 基于深度学习

产业链挖掘 - 基于正则表达式的流程

字典替换

正则匹配

句式分析

探寻

结果输出

Step1:
根据规则，找到上下游/原材料 对应的内容,并用上下游/原材料进行替换

找到上下游及对应词

INDUST0下游多为PRODU0、PRODU1等小企业

INDUST0下游等小企业 + {下游:[PRODU0,PRODU1]}

Step2:
根据简单表达式，找到主语,并生成关系

小范围(简单句式)寻找受体

((?:PRODU|INDUST)\d)(?:的|涉及|以及|和)?([上下中]游|原材料|材料)

INDUST0下游等小企业 ➔ {‘ INDUST0’ :{下游：[PRODU0,PRODU1]}}

Step3:
在第二步没有找到合适的主体情况下，找到主句，靠前的作为主体

在主句中寻找

受INDUST1、PROD1等下游行业景气度下降影响，11年度国内PROD0市场增长缓慢

受下游行业景气度下降影响，11年度国内PROD0市场增长缓慢 + {下游:[INDUST1,PROD1]}

受下游行业景气度下降影响，11年度国内PROD0市场增长缓慢 ➔ {下游:[INDUST1,PROD1]}

Step4:
第三部还未找到，则回溯到前面几句，找到主句/标题中的行业词/产品词，示例略

回溯寻找

产业链挖掘 - 基于LTP的处理流程

字典替换

正则粗筛

获取ltp树

基于ltp规则的提取和分析

探寻

结果输出

用行业词、产品词替换

行业词|产品词替换

INDUST1下游主要涉及INDUST2、INDUST3、INDUST4、INDUST5等多个行业，三季度在保增长政策和重点行业振兴计划的支撑下，PROD1的需求也值得看好。

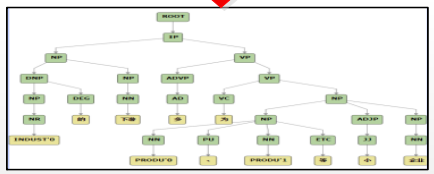
{‘ indust’ :[煤炭,电力,钢铁,化工,建材], product’ :[无烟煤]}

Step2:
使用正则表达式对目标进行初筛

找到上下游及对应词

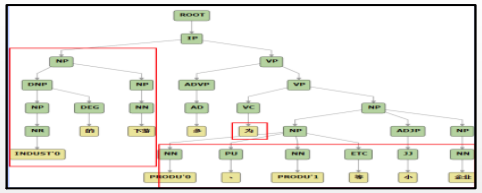
INDUST0下游多为PRODU0、PRODU1等小企业

Step3:
获取LTP树



Step3:
在第二步没有找到合适的主体情况下，找到主句，靠前的作为主体

在主句中寻找



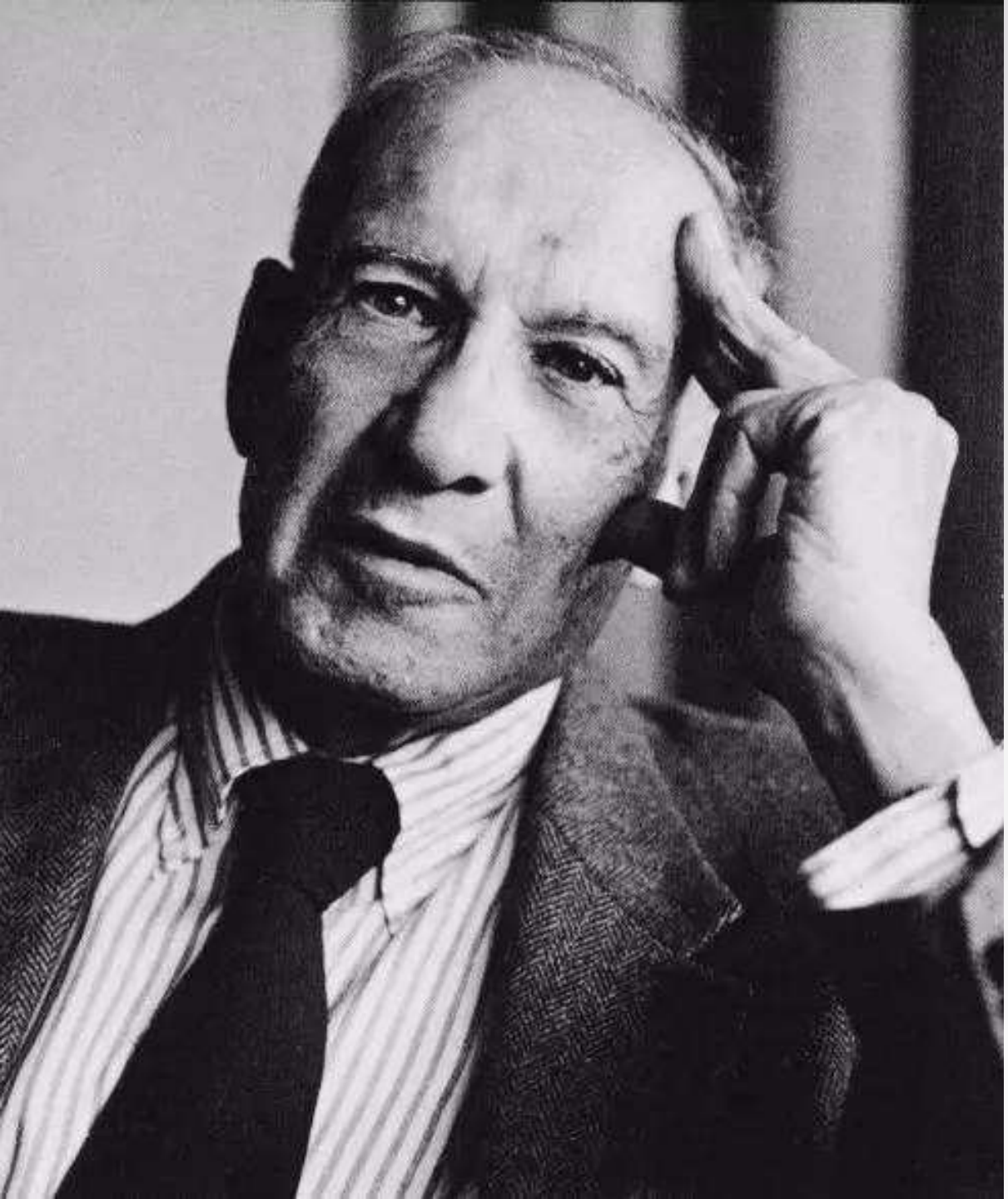
([INDUST’ 0],下游,[PRODU’ 0 PRODU’ 1等])

Step4:
根据关键词出现和位置和固定搭配，在分句的前后找句子进行补充

回溯寻找

两种方式的比较

	优势	劣势
基于正则	✓ 适合预处理，用于发现目标关系 ✓ 简单，方便，有较高的召回率	✓ 对内容提取的精确度不够 ✓ 人工处理量较大
基于LTP	✓ 对于简单结构的句子处理结果优势明显 ✓ 可用于关键词/目标词之间的关联分析和关系发现	✓ 对文档的写作质量要求较高 ✓ 复杂句式处理能力有限



今天，知识就是力量。

它决定了机会的获得和事业的前程。

——彼得·德鲁克

THANKS!

http://www.chinascope.com