百度基础技术研发

赵世奇 百度

2012-08-16

定位

前瞻

通用

基础

研究内容

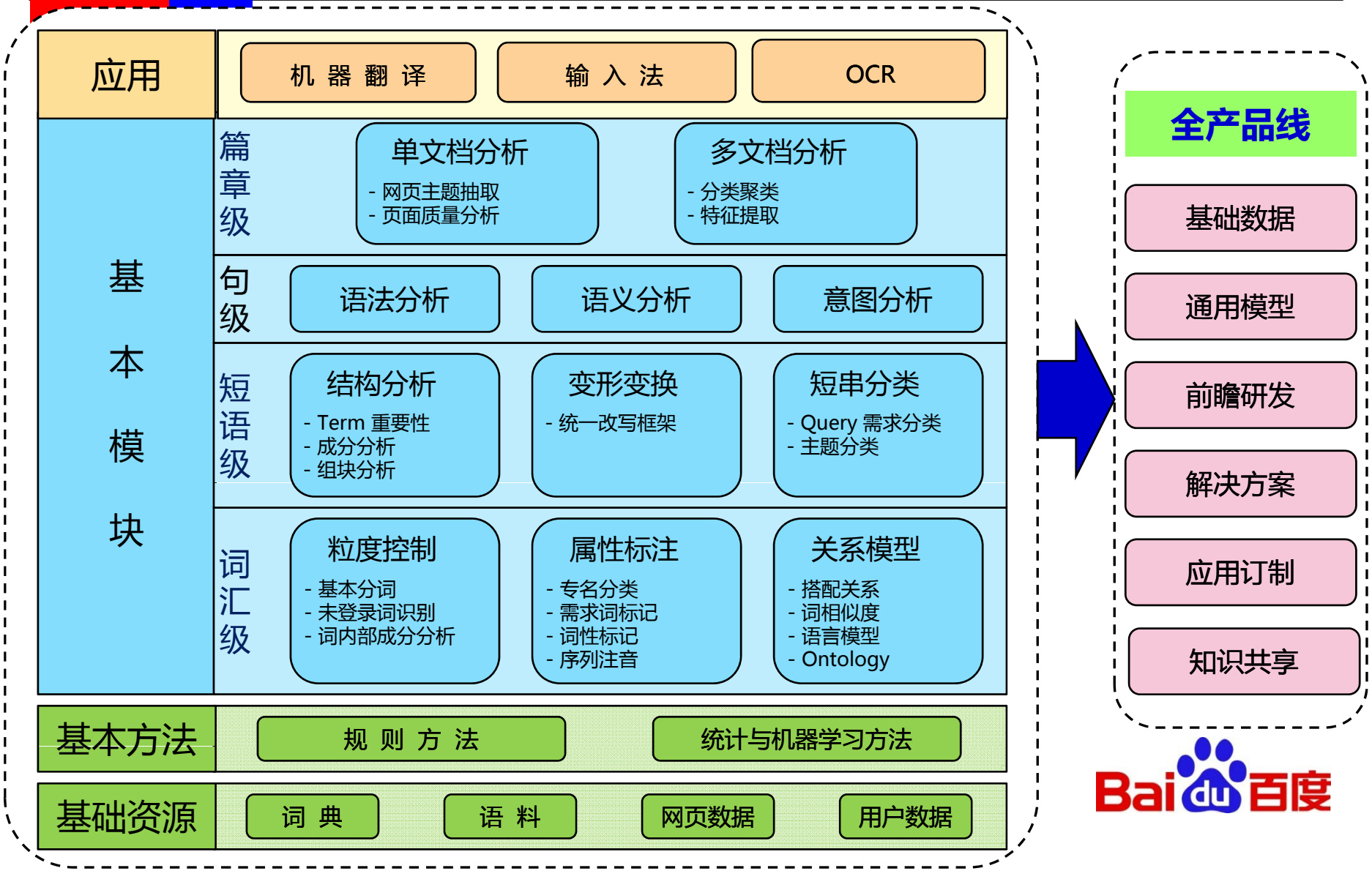
声　图　文

研发内容

- 自然语言处理
- 互联网数据
  - Spider
  - 互联网数据挖掘
- 多媒体
  - 语音
  - 视觉

Baidu 百度

百度自然语言处理全景图



NLP业务

- 基础模型与算法
  - 统计机器学习模型
- 基础数据与技术
  - 基础语言资源
  - 用户数据
  - 词法、句法、语义及篇章分析
- 应用技术
  - Query处理技术
  - 用户理解和通用表示空间构建
- 独立产品
  - 机器翻译
  - 输入法
  - OCR

Baidu 百度

Ontology建设与应用

提纲

Ontology应用价值

Ontology构建 — 新词挖掘

Ontology构建 — 关系挖掘

Ontology应用实例

Ontology应用意义

- 互联网应用
  - 基础应用：分词与专名识别
    - 新词/专名的挖掘、分类与属性计算
  - 互联网搜索：从传统搜索到语义搜索
    - 丰富的用户引导与知识展现
  - 推荐与个性化：用户建模的底层框架
    - 将用户模型映射到Ontology的知识结构
  - 自动问答：问答系统的知识基础
    - 提供事实性问题的精准答案
  - ……

Ontology应用意义

- 应用中体现的新问题
  - 新词层出不穷
    - 各类专名（电影、网络红人等）、新词等不断涌现
  - 知识动态变化
    - 如“价格”等类别的属性、“相关实体”等类别的关系
  - 涉及领域繁多
    - 应用的方方面面：母婴、医疗、旅游、美食、娱乐……
  - 数据规模巨大
    - 亿级别数量级
  - 挖掘方法多样
    - 往往是将多种方法的挖掘结果融合在一起

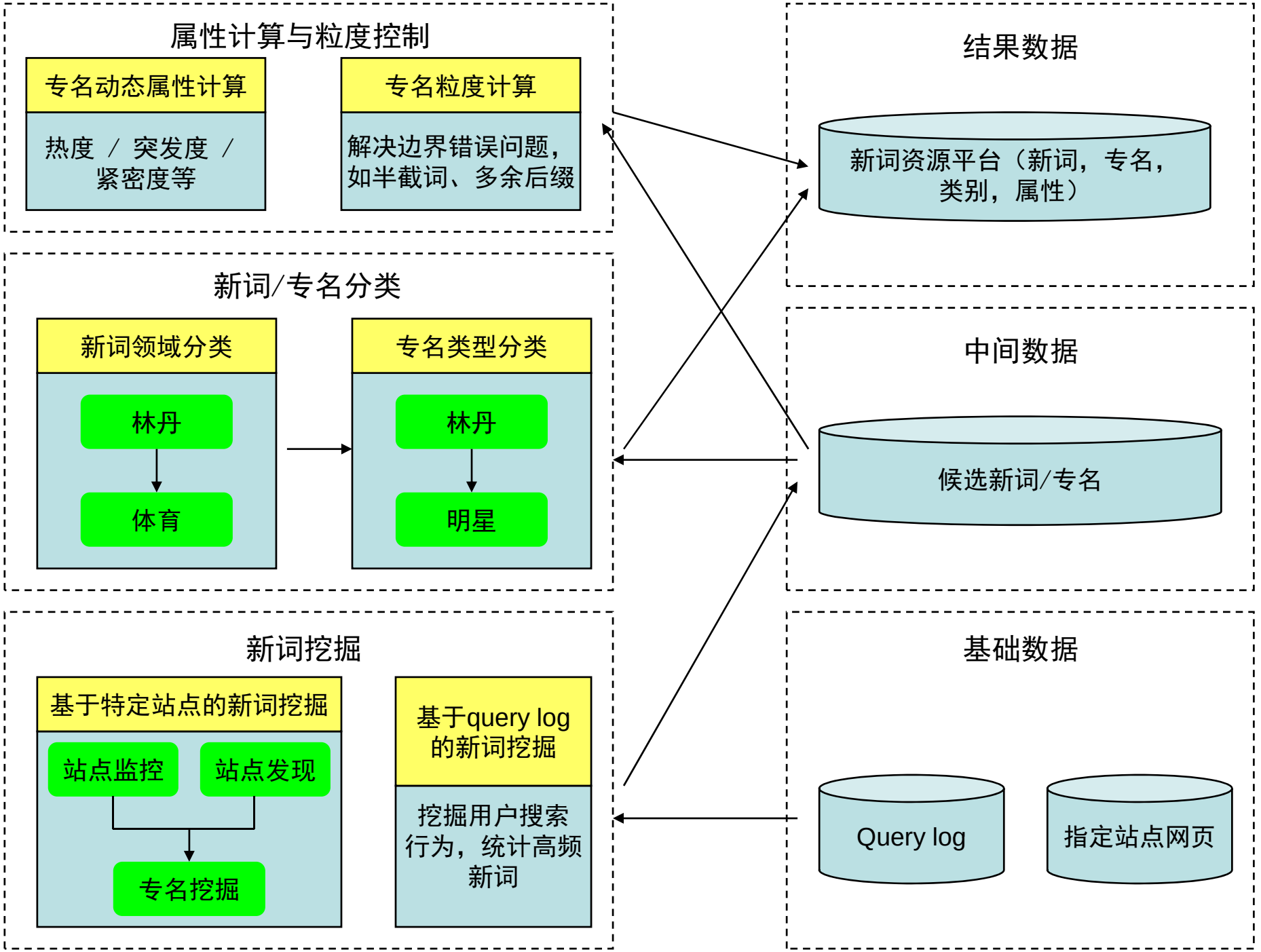
提纲

Ontology应用价值

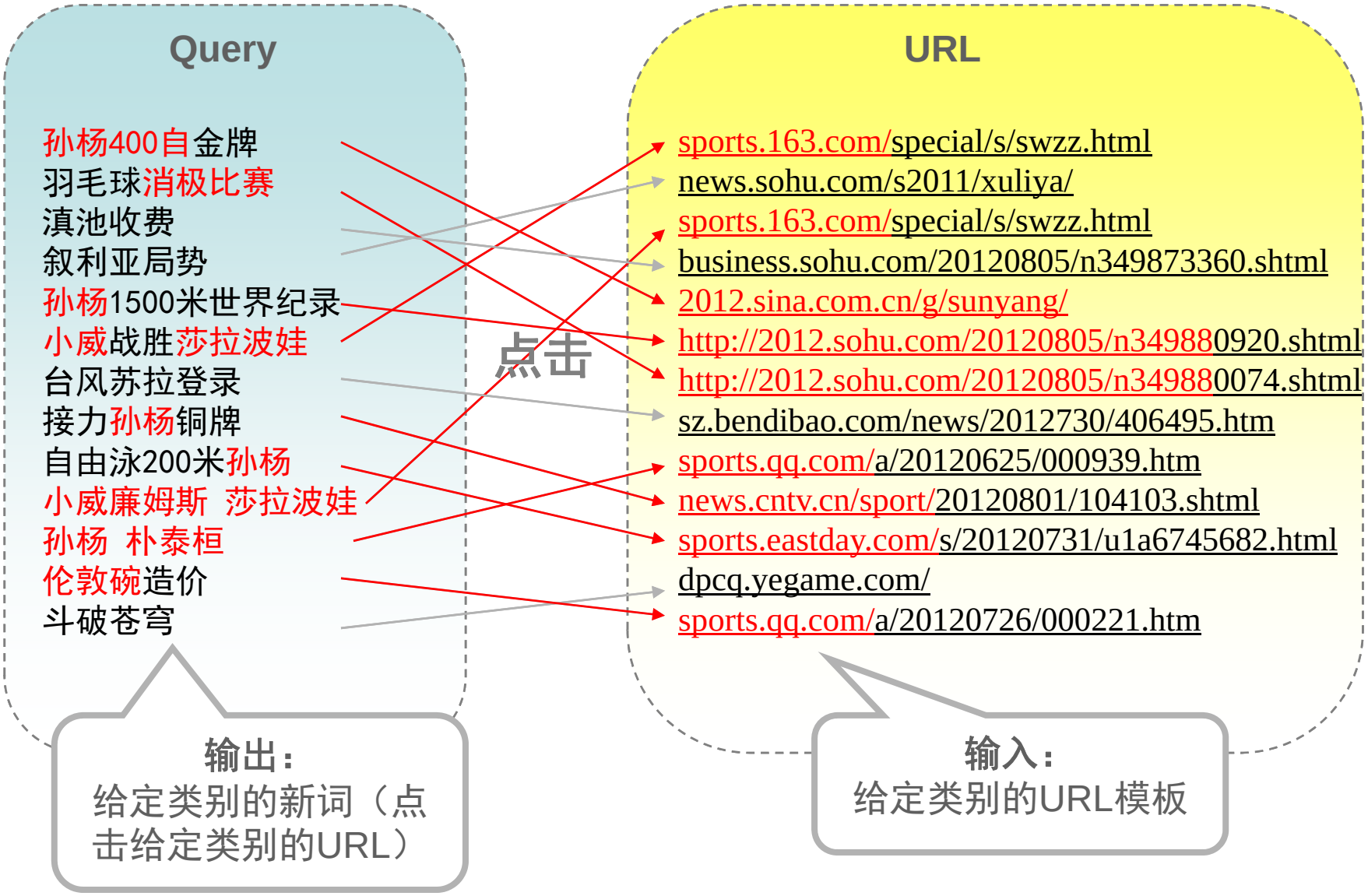
Ontology构建 — 新词挖掘

Ontology构建 — 关系挖掘

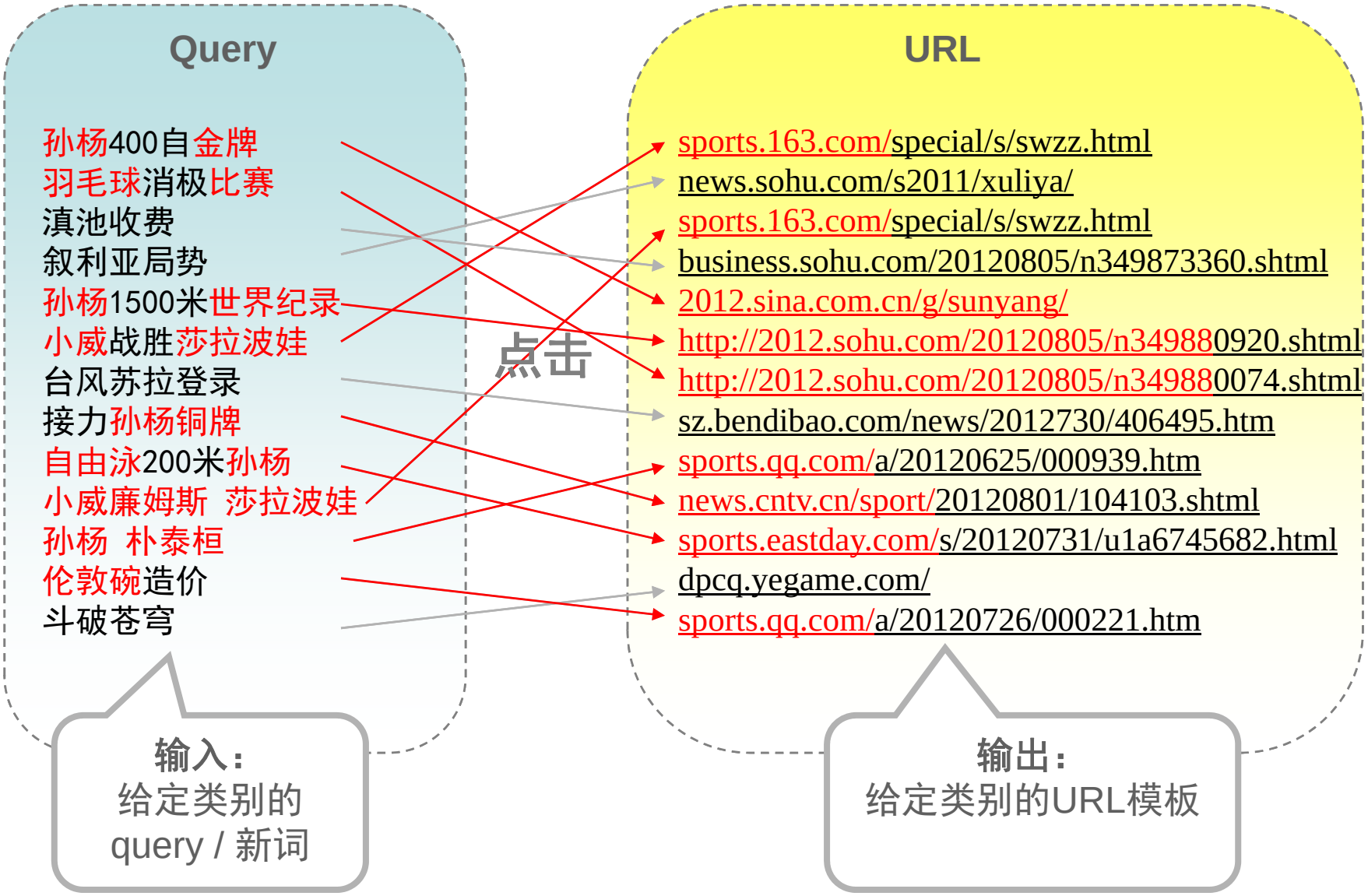
Ontology应用实例


属性计算与粒度控制
专名动态属性计算
热度 / 突发度 / 紧密度等
专名粒度计算
解决边界错误问题，如半截词、多余后缀
新词/专名分类
新词领域分类
林丹
体育
专名类型分类
林丹
明星
新词挖掘
基于特定站点的新词挖掘
站点监控
站点发现
专名挖掘
基于query log的新词挖掘
挖掘用户搜索行为，统计高频新词
结果数据
新词资源平台（新词，专名，类别，属性）
中间数据
候选新词/专名
基础数据
Query log
指定站点网页

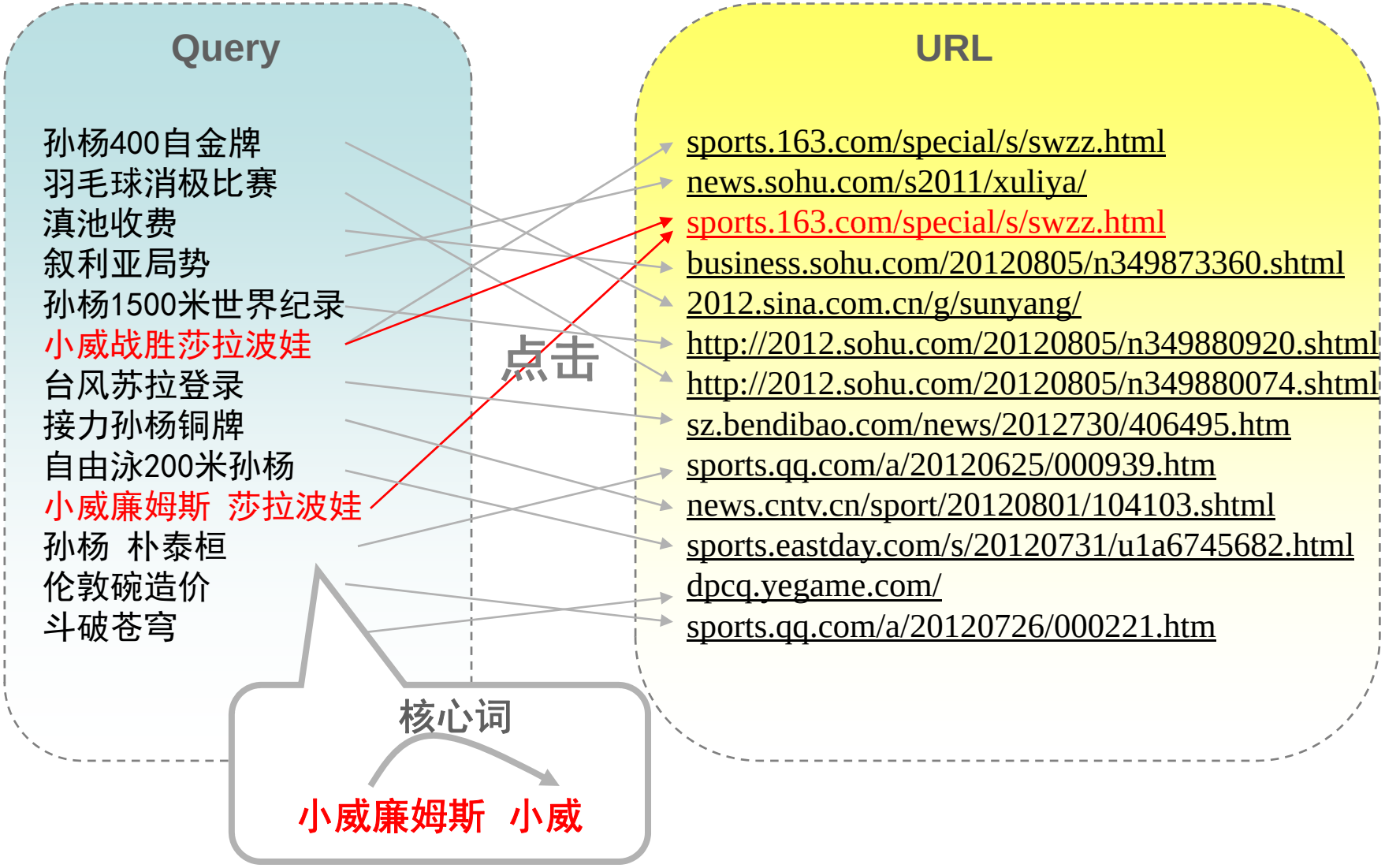
基于用户日志的中文新词挖掘－新词分类



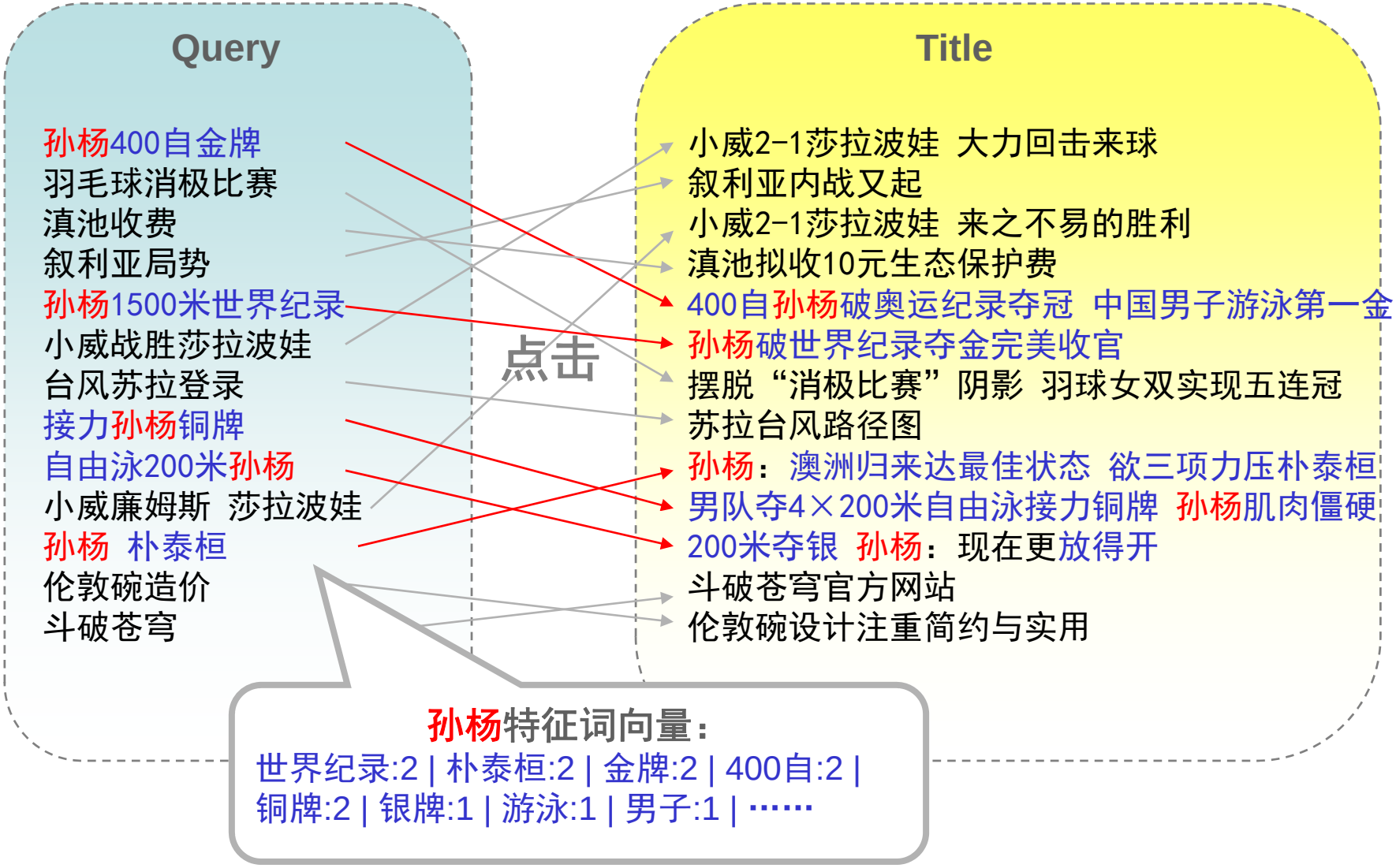
基于用户日志的中文新词挖掘 – URL模板挖掘



基于用户日志的中文新词挖掘 – 核心词挖掘



基于用户日志的中文新词挖掘 – 特征词挖掘



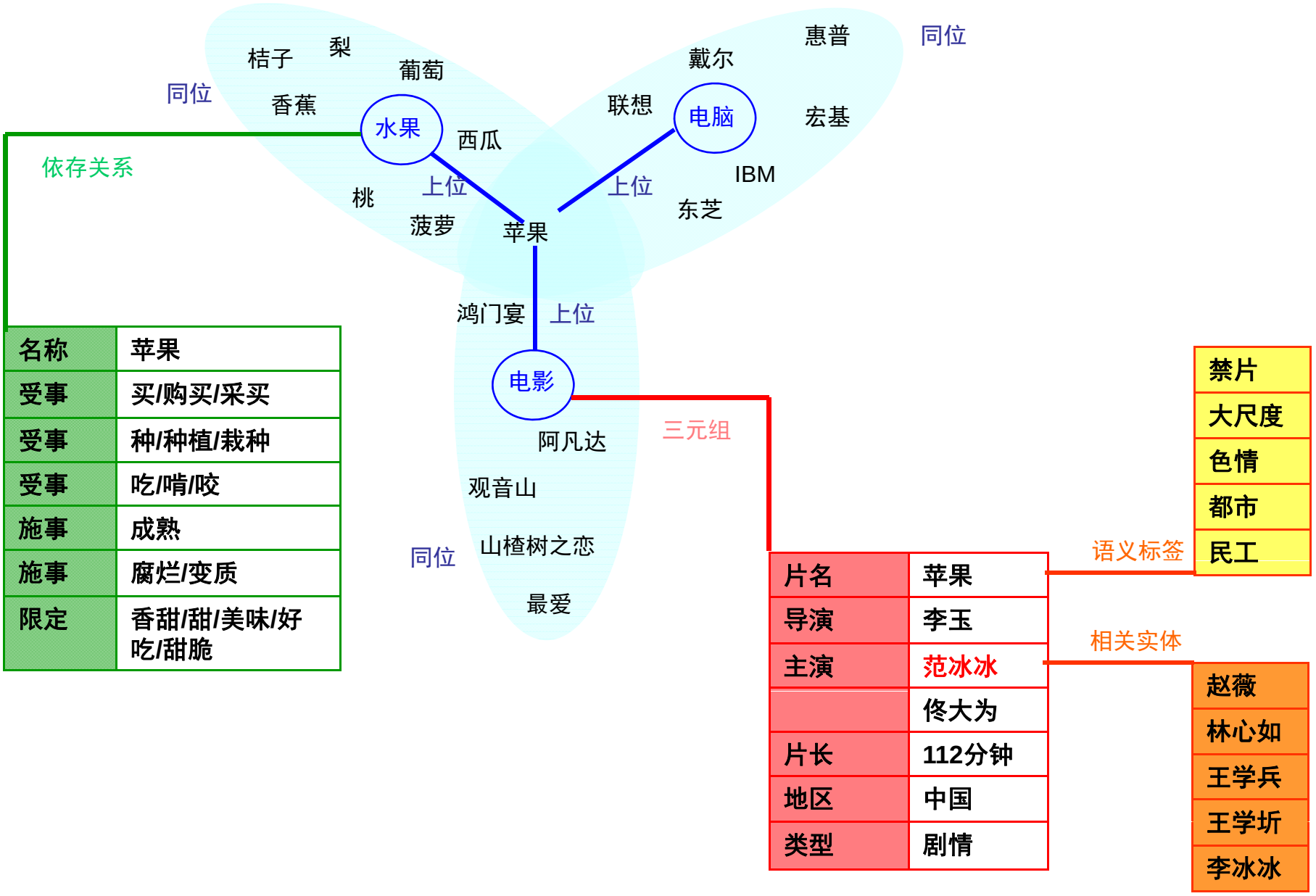
提纲

Ontology应用价值

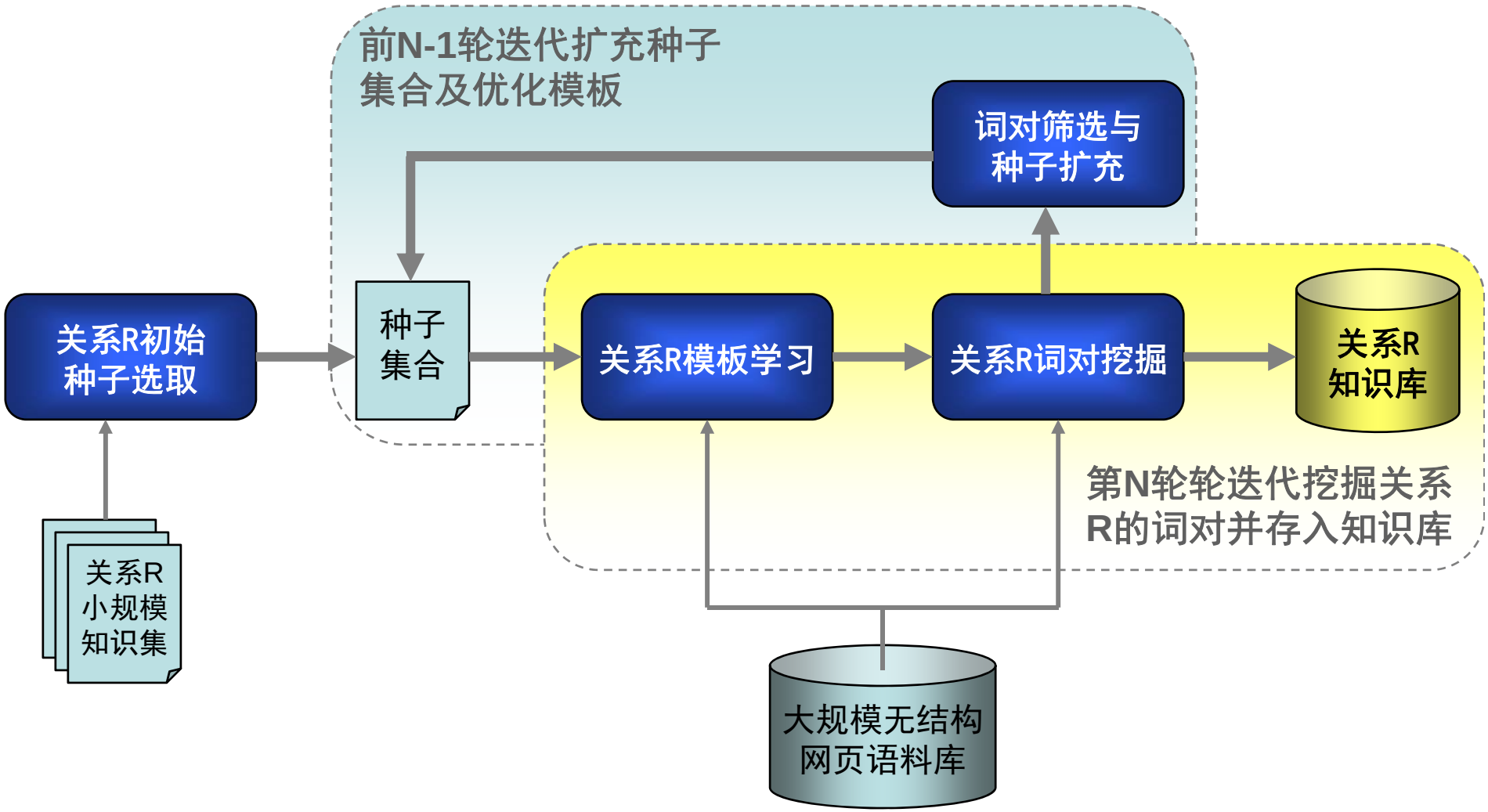
Ontology构建 — 新词挖掘

Ontology构建 — 关系挖掘

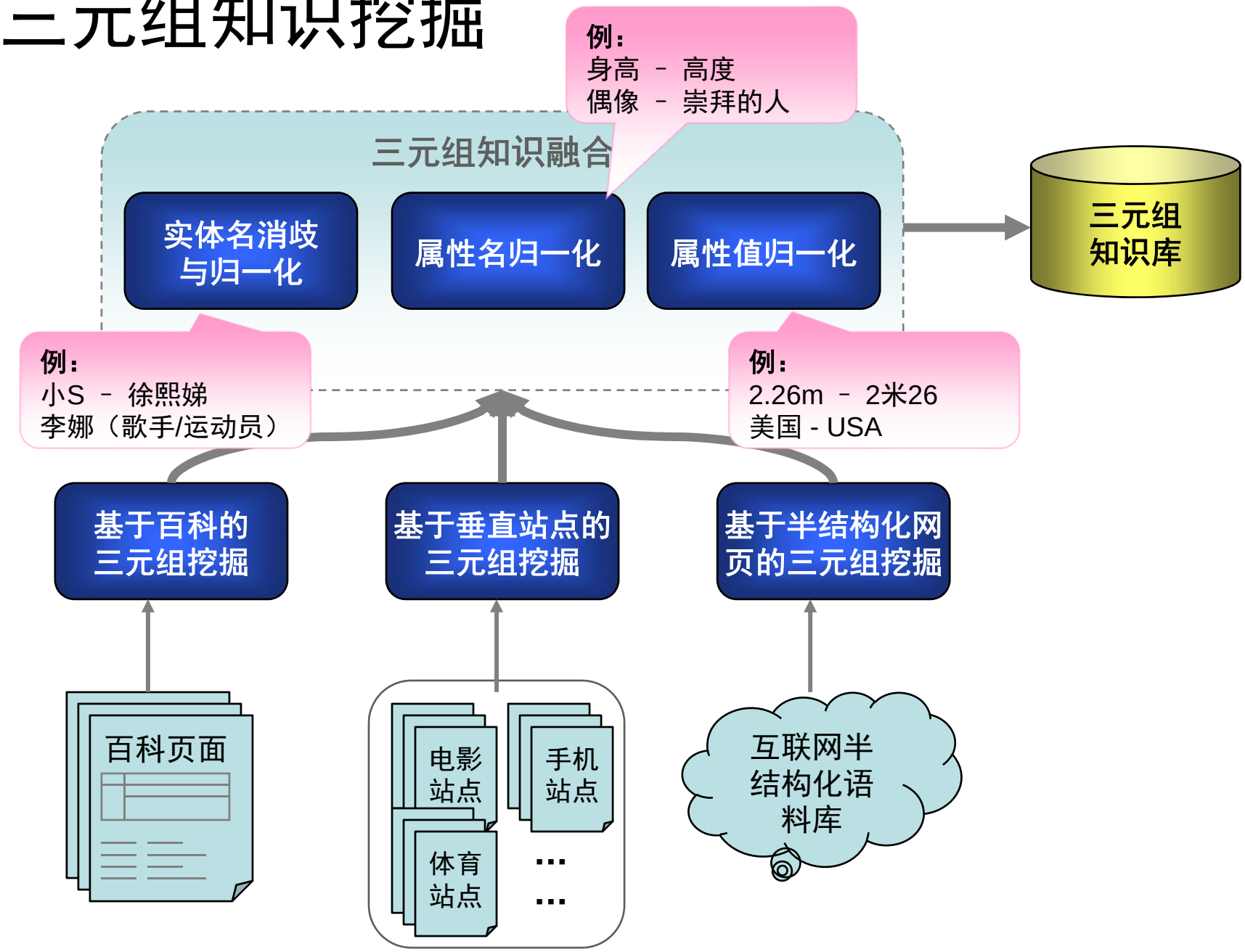
Ontology应用实例


桔子
梨
葡萄
同位
香蕉
水果
西瓜
依存关系
桃
上位
菠萝
苹果
戴尔
惠普
同位
联想
电脑
宏基
IBM
上位
东芝
鸿门宴
上位
电影
三元组
阿凡达
观音山
同位
山楂树之恋
最爱
名称 苹果
受事 买/购买/采买
受事 种/种植/栽种
受事 吃/啃/咬
施事 成熟
施事 腐烂/变质
限定 香甜/甜/美味/好吃/甜脆
片名 苹果
导演 李玉
主演 范冰冰
佟大为
片长 112分钟
地区 中国
类型 剧情
语义标签
禁片
大尺度
色情
都市
民工
相关实体
赵薇
林心如
王学兵
王学圻
李冰冰

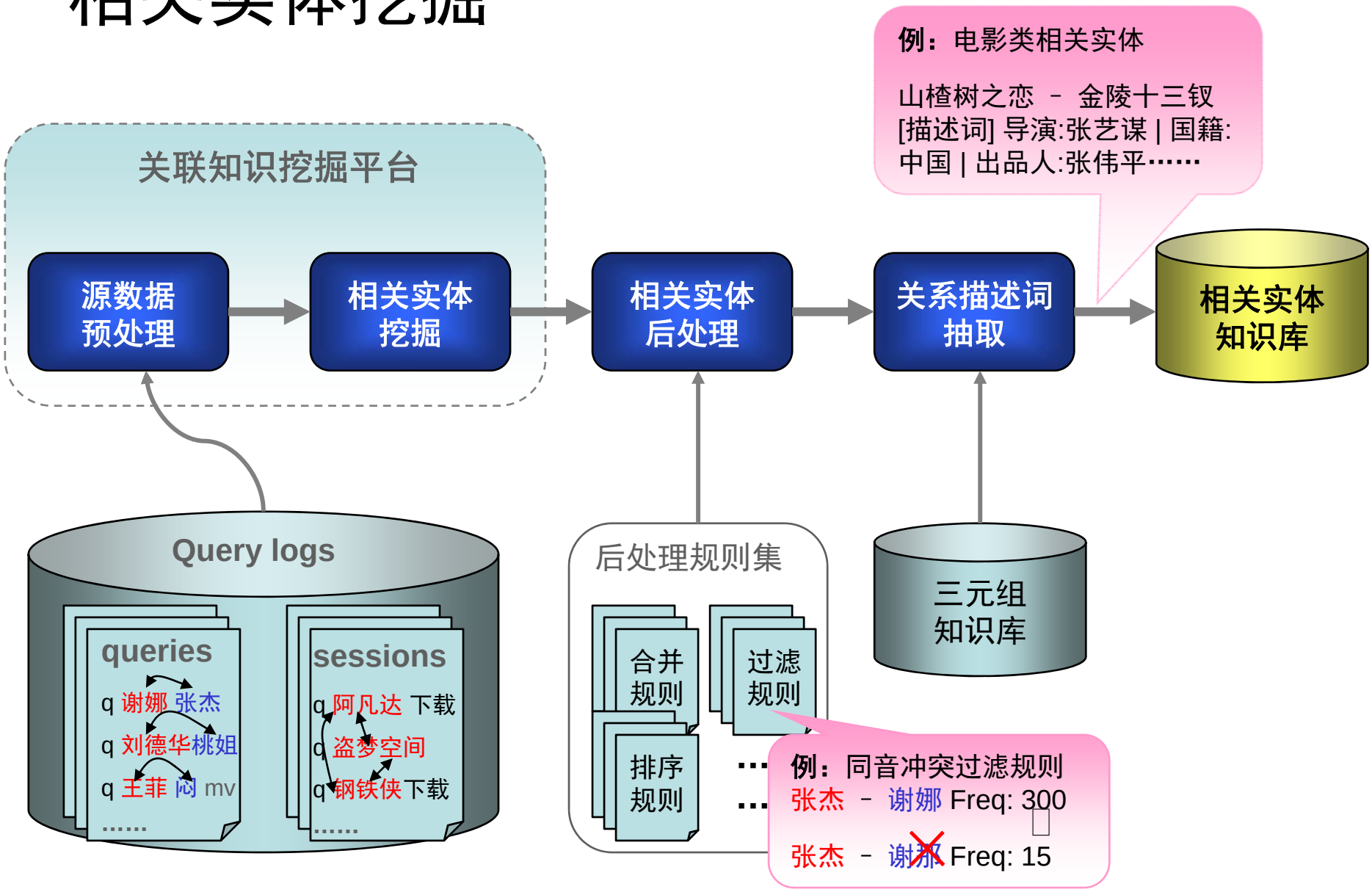
上下位/同位知识挖掘


前N-1轮迭代扩充种子
集合及优化模板
词对筛选与
种子扩充
关系R初始
种子选取
种子
集合
关系R模板学习
关系R词对挖掘
关系R
知识库
第N轮轮迭代挖掘关系
R的词对并存入知识库
关系R
小规模
知识集
大规模无结构
网页语料库

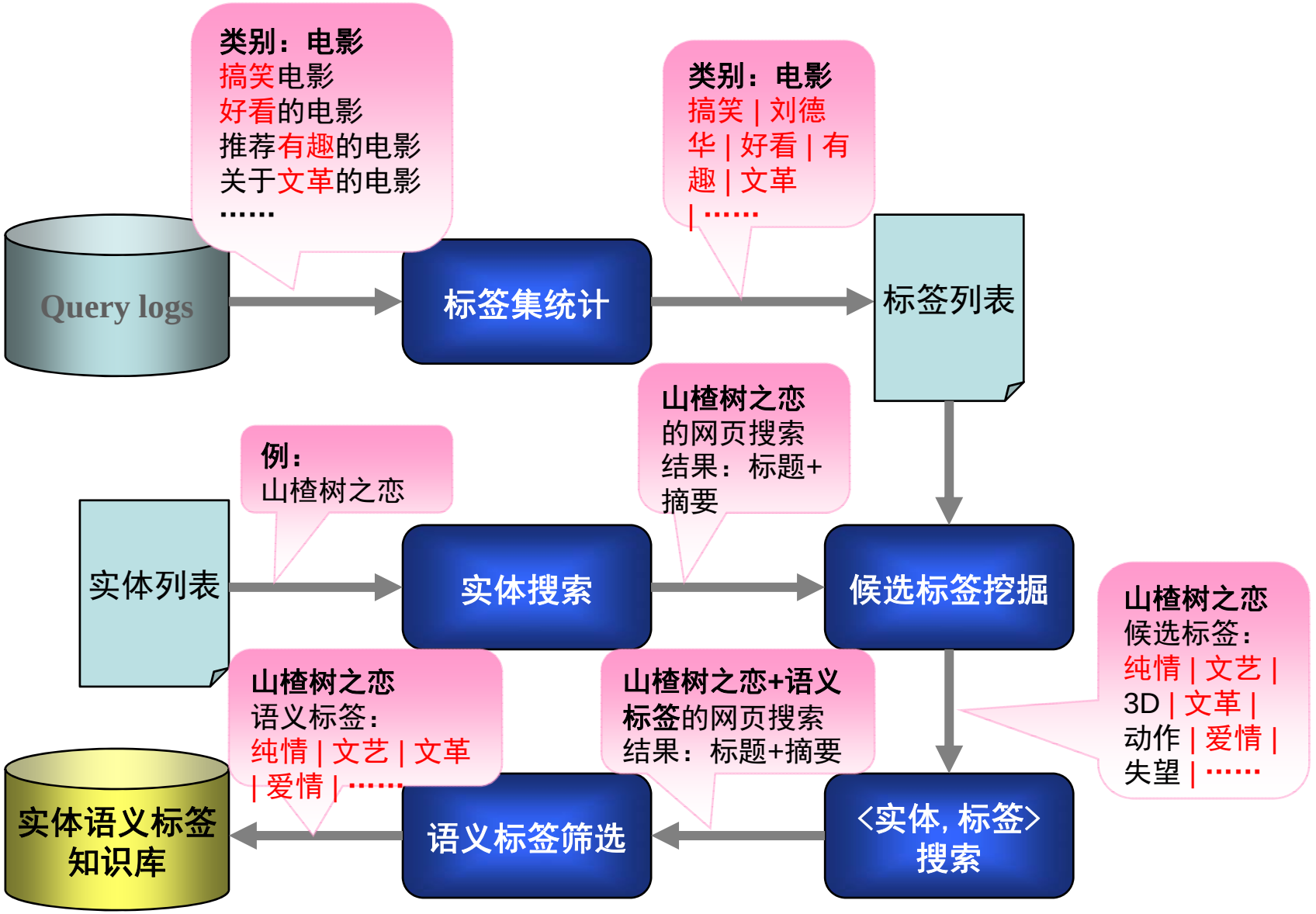
三元组知识挖掘


例：
身高 – 高度
偶像 – 崇拜的人
三元组知识融合
实体名消歧与归一化
属性名归一化
属性值归一化
三元组知识库
例：
小S – 徐熙娣
李娜（歌手/运动员）
例：
2.26m – 2米26
美国 - USA
基于百科的三元组挖掘
基于垂直站点的三元组挖掘
基于半结构化网页的三元组挖掘
百科页面
电影站点
手机站点
体育站点
...
...
互联网半结构化语料库

相关实体挖掘


例：电影类相关实体
山楂树之恋 - 金陵十三钗
[描述词] 导演:张艺谋 | 国籍:
中国 | 出品人:张伟平……
关联知识挖掘平台
源数据
预处理
相关实体
挖掘
相关实体
后处理
关系描述词
抽取
相关实体
知识库
Query logs
queries
q 谢娜 张杰
q 刘德华桃姐
q 王菲 闷 mv
.....
sessions
q 阿凡达 下载
q 盗梦空间
q 钢铁侠 下载
.....
后处理规则集
合并
规则
过滤
规则
排序
规则
三元组
知识库
例：同音冲突过滤规则
张杰 - 谢娜 Freq: 300
张杰 - 谢那 Freq: 15

实体语义标签挖掘


类别：电影
搞笑电影
好看的电影
推荐有趣的电影
关于文革的电影
……
Query logs
标签集统计
类别：电影
搞笑 | 刘德华 | 好看 | 有趣 | 文革 | ……
标签列表
例：
山楂树之恋
实体列表
实体搜索
山楂树之恋
的网页搜索结果：标题+摘要
候选标签挖掘
山楂树之恋
候选标签：
纯情 | 文艺 | 3D | 文革 | 动作 | 爱情 | 失望 | ……
山楂树之恋
语义标签：
纯情 | 文艺 | 文革 | 爱情 | ……
实体语义标签知识库
语义标签筛选
山楂树之恋+语义标签的网页搜索结果：标题+摘要
〈实体，标签〉搜索

提纲

Ontology应用价值

Ontology构建 — 新词挖掘

Ontology构建 — 关系挖掘

Ontology应用实例

Ontology应用实例

- 网页搜索
  - 基于新词挖掘资源改善查询词紧密度

Ontology应用实例

- 输入法
  - 基于新词资源更新输入法词库

北京卓思天成国际市场研究咨询有限公司
北京卓思天成国际市场研究咨询有限公司卓思是一家定位于为主流汽车厂商提供信息咨询服务的企业，目前主要为各大汽车厂家提供新产品上市策略、品牌研究、竞争分析、用户...
www.sunjob.com.cn/guanggao/adView.asp?Adl ... 2012-7-29 - 百度快照

北京卓思天成国际市场研究咨询有限公司,北京卓思天成国际市场研究...
北京卓思天成国际市场研究咨询有限公司简介 卓思是一家新近成立的中小型信息咨询企业，以汽车行业市场研究咨询业务为主。 卓思定位于为少数高端客户提供高品质产品和...
www.01hr.com/company/c-342334696357.html 2012-8-3 - 百度快照

Ontology应用实例

- 自动问答
  - 基于三元组知识提供确定性问题的答案
    - http://zhidao.baidu.com

Ontology应用实例

- APP推荐
  - 基于相关实体知识推荐相关的app

谢谢!

Q&A