面向中文知识图谱构建的知识融合与验证

孙乐　　韩先培

中国科学院软件研究所
基础软件国家工程研究中心

中文知识图谱

- NLP和AI的终极目标之一是构建比肩人类的文本阅读和理解系统
- 缺乏支撑计算机智能推理和决策的知识库一直是构建上述系统的瓶颈之一
- **目标：逐步构建可支撑上述目标的中文知识图谱**

相关工作—传统知识库

- 基于人工编写方式，构建了一系列的中小规模中文知识库
  - 知网（HowNet）[董振东 和 董强，1999]
  - 《同义词词林》[梅家驹等，1996]
  - 概念层次网络（HNC）[黄曾阳，1997]
- 特点
  - 规模相对较小
  - 建模的知识范围特定
  - 不同知识库构建的目的不一样，因此使用不同的语义描述元数据，覆盖不同类别的知识

相关工作—协同知识库

- 基于Web 2.0的方式，各个领域都有丰富的Web 2.0知识站点创立
  - 通用知识：百度百科，维基百科，互动百科
  - 书籍、音乐、电影：豆瓣
  - 商品：淘宝，中关村在线，太平洋
  - 餐馆：大众点评
  - 医学：丁香园

- 由于Web的去中心化结构，这些知识以分散、异构、自治的形式存在，而不是一个统一、一致的知识整体

特点总结

1. 分散：知识独立自治的存在于多个源中
2. 异构：不同知识资源使用不同的结构和元数据
3. 冗余：各个知识源中的知识具有一定的重叠（同构或异构的方式）
4. 噪音：Web 2.0方式会引入大量错误和噪音
5. 不确定：通常需要集成不确定的信息抽取系统结果
6. 非完备：知识的长尾性 → 仅仅覆盖特定领域的高频知识，大部分是常识知识库
7. 中文知识的缺乏：现在已经有大规模的英文知识图谱，但是大规模中文知识图谱的工作相对缺乏

出发点

- 如何从当前的这些知识出发，构建准确、高覆盖、一致的大规模中文知识图谱？

- 策略一：融合
  - 充分利用现有知识库，融合这些分散、冗余和异构的知识，作为构建中文知识图谱的出发点

- 策略二：验证
  - 对新加入知识图谱的知识（如信息抽取系统的结果，众包标注）进行验证，确保新知识与知识图谱的一致性，持续更新中文知识图谱

知识融合

知识融合

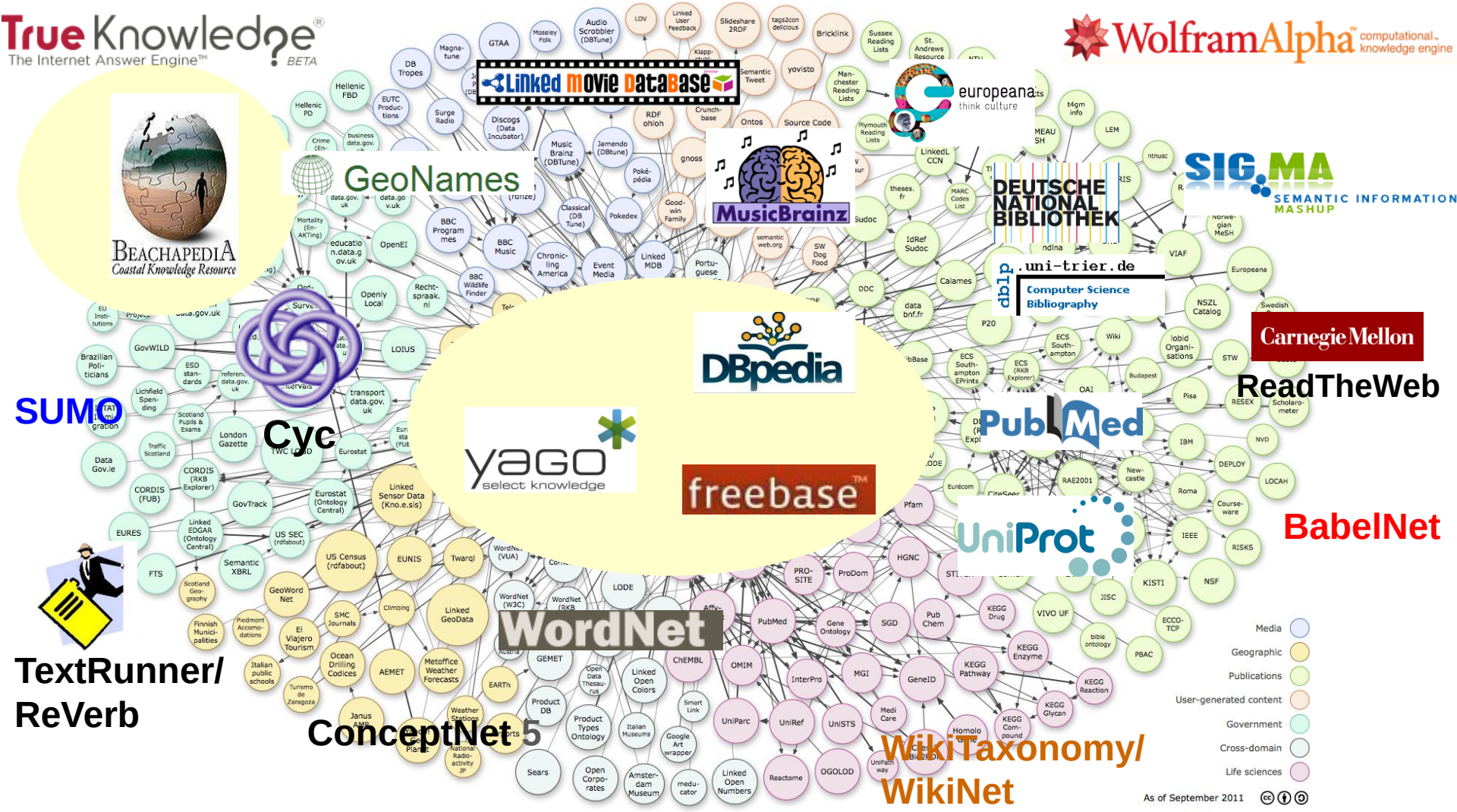
- 定义（Wikipedia）: The merging of information from heterogeneous sources with differing conceptual, contextual and typographical representations
- 融合的层面
  - 数据层融合
    - Record Linkage/Entity Linking/Entity Resolution
    - 百度百科：中国<--> Wikipedia: China<-->互动百科：中国
  - 语义描述层融合
    - Schema Mapping
    - 百度百科：科学家类别<--> Wikipedia:Scientist Category
    - 百度百科：人物.出生信息<-->Wikipedia：人物.出生日期 和 人物.出生地点

数据层融合关键技术--实体链接

- 等同性（Equality）判断
  - 给定不同数据源中的实体，判断其是否指向同一个真实世界实体
  - 大陆：贝克汉姆 == 香港：碧咸==北美：Beckham？

- 基于等同性判断，我们可以连接不同知识源中的等同知识，从而将多个分散的知识源连接成为一个整体
  - Linked Data

Linked Data全景：300亿事实（还在不断增长中）



http://richard.cyganiak.de/2007/10/lod/lod-datasets_2011-09-19_colored.png

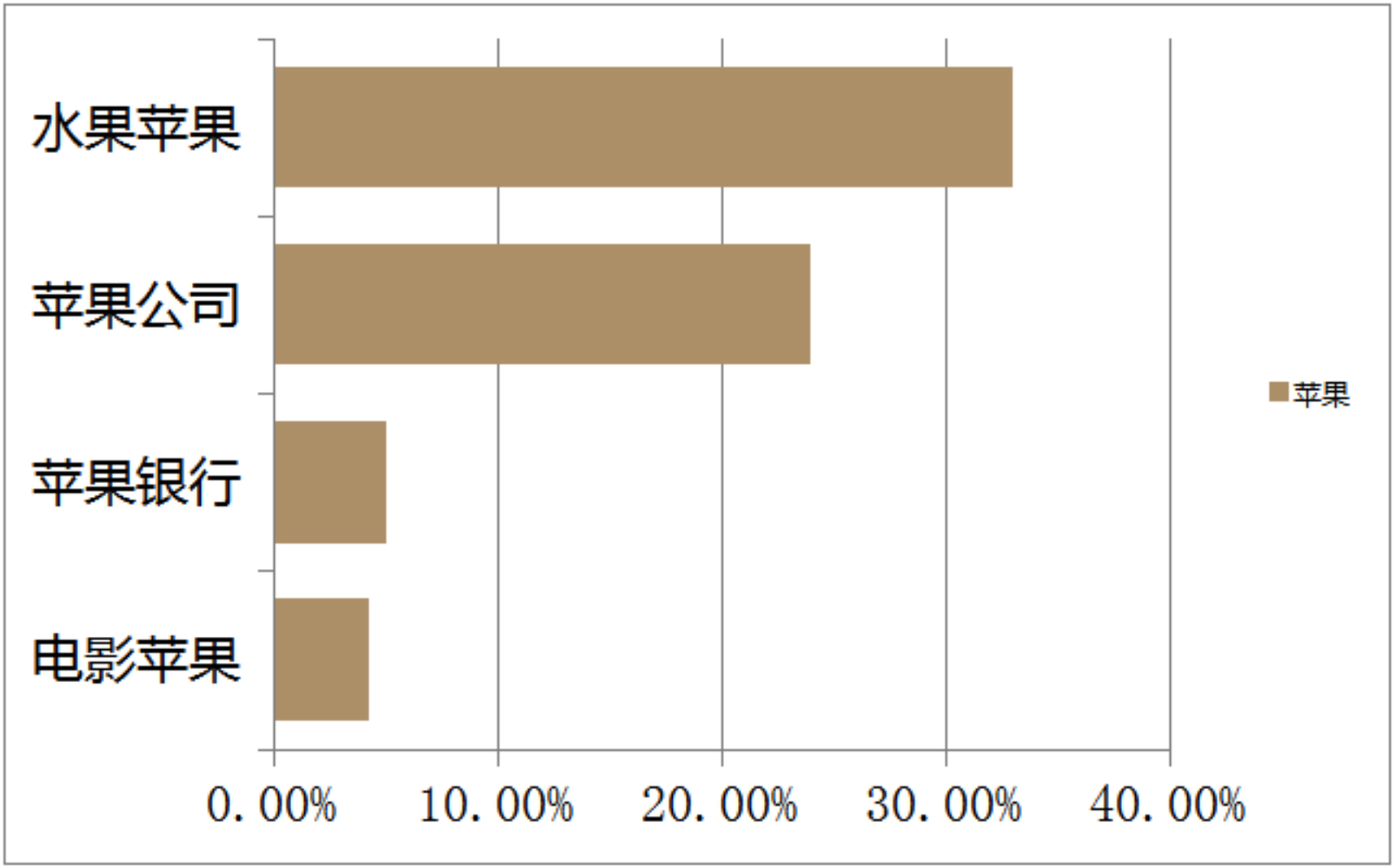
实体链接方法(1)：基于实体知识的链接

基于实体-提及模型的实体链接

- 人们在进行链接工作时，使用了大量关于实体的知识
  - 实体的知名度
  - 实体的名字分布
  - 实体的上下文分布

- 提出了实体-提及模型来融合上述异构知识

实体知名度



实体的名字分布

- 一个实体的名字通常是固定的，且以一定的概率出现

- IBM和国际商用机器公司都可以作为IBM公司的名字，但是BMI，Oracle不会作为它的名字

- IBM比全称国际商用机器公司更常作为IBM公司的名字出现

实体名字模型

- 建模了许多不同的名字构建方式
  - 保持原始形式： 迈克尔→ 迈克尔
  - 缩写： 亲爱的顾客 → 亲（淘宝体）
  - 省略： 李克强 总理 → … 总理
  - 翻译：乔丹→ 佐顿
  - 其它方式：科比→大神，薄熙来→不厚

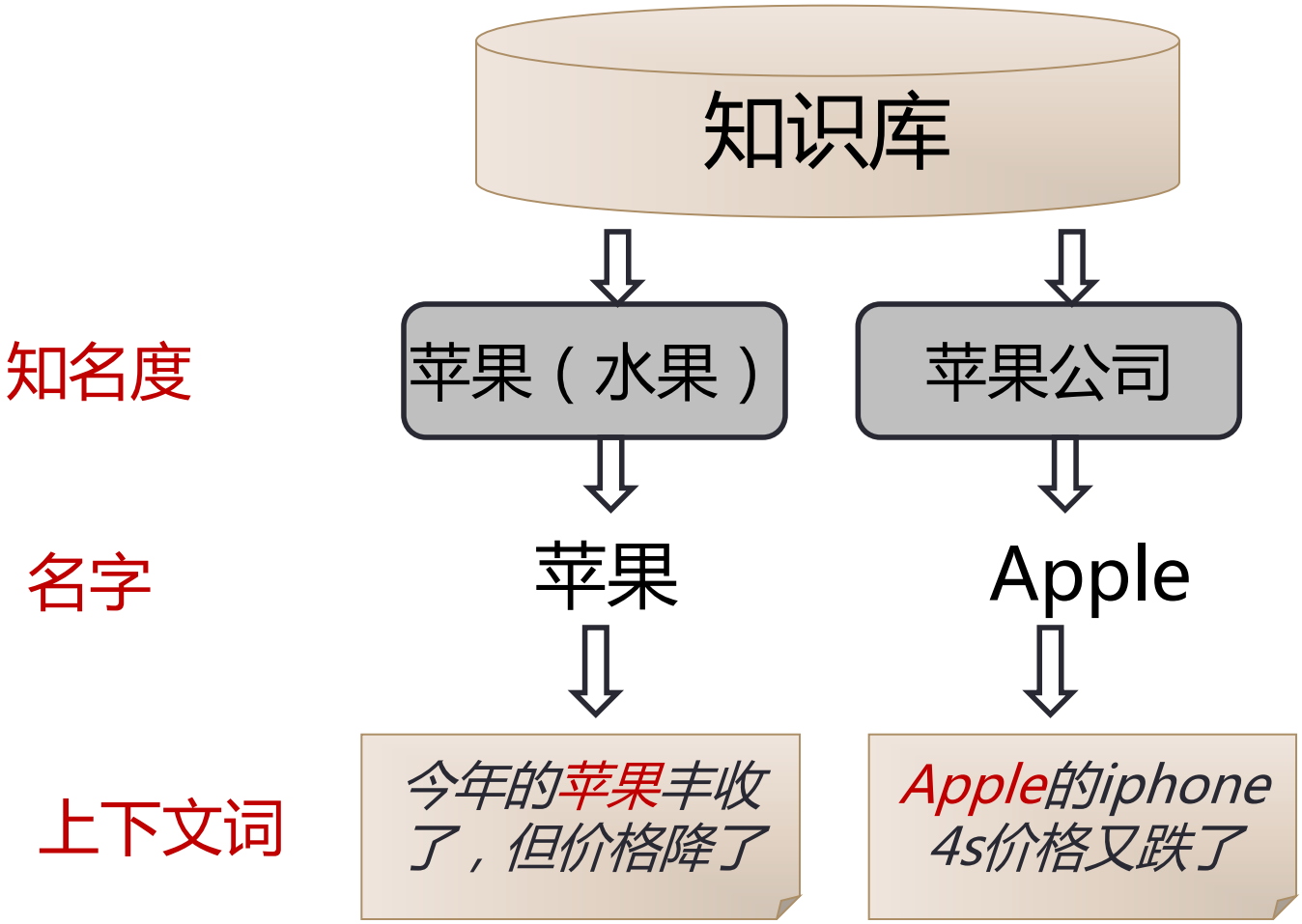
实体的上下文词分布

- 不同实体的上下文词分布通常有极大的差异

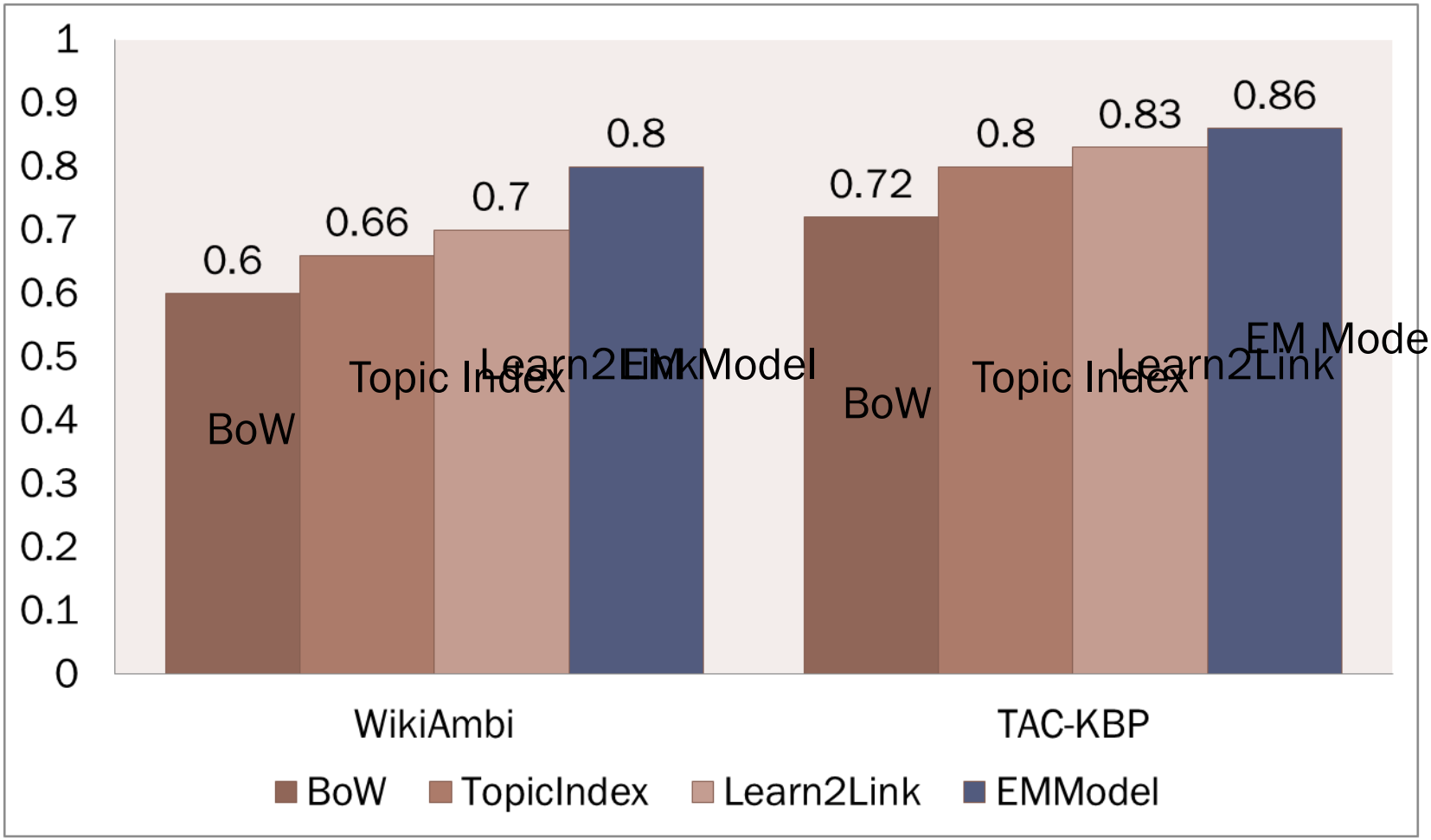
苹果公司

苹果银行

基于实体-提及模型融合上述知识



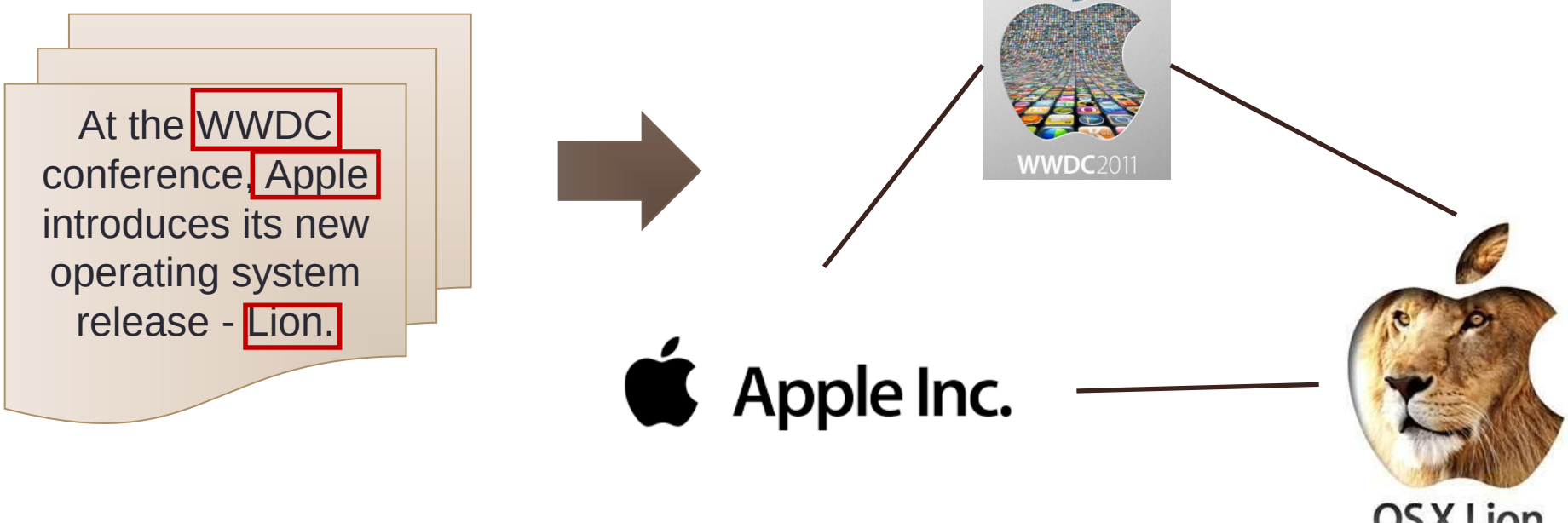
实验性能


准确率
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.6
0.66
0.7
0.8
0.72
0.8
0.83
0.86
BoW
Topic Index
Learn2Link
EM Model
WikiAmbi
TAC-KBP
BoW
TopicIndex
Learn2Link
EMModel

实体链接方法(2)：基于篇章主题的链接

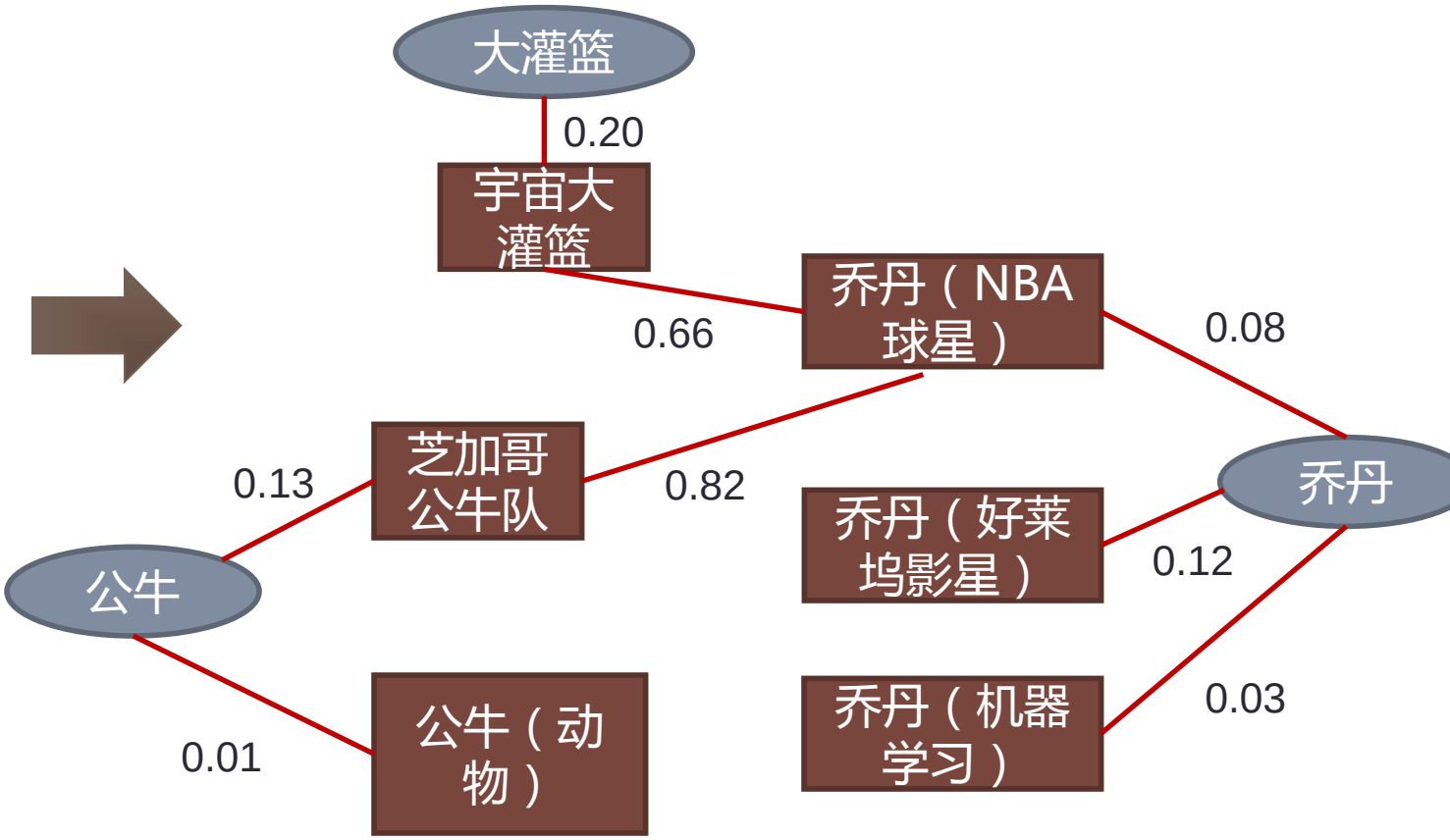
主题一致性假设

- 文章中的实体通常与文本主题相关，因此这些实体相互之间语义相关
  - 出现实体ipad和iphone的文章也更有可能出现苹果公司，而不是水果苹果或苹果银行



基于图的协同推断

在公牛期间，乔丹出演了电影大灌篮



协同推导

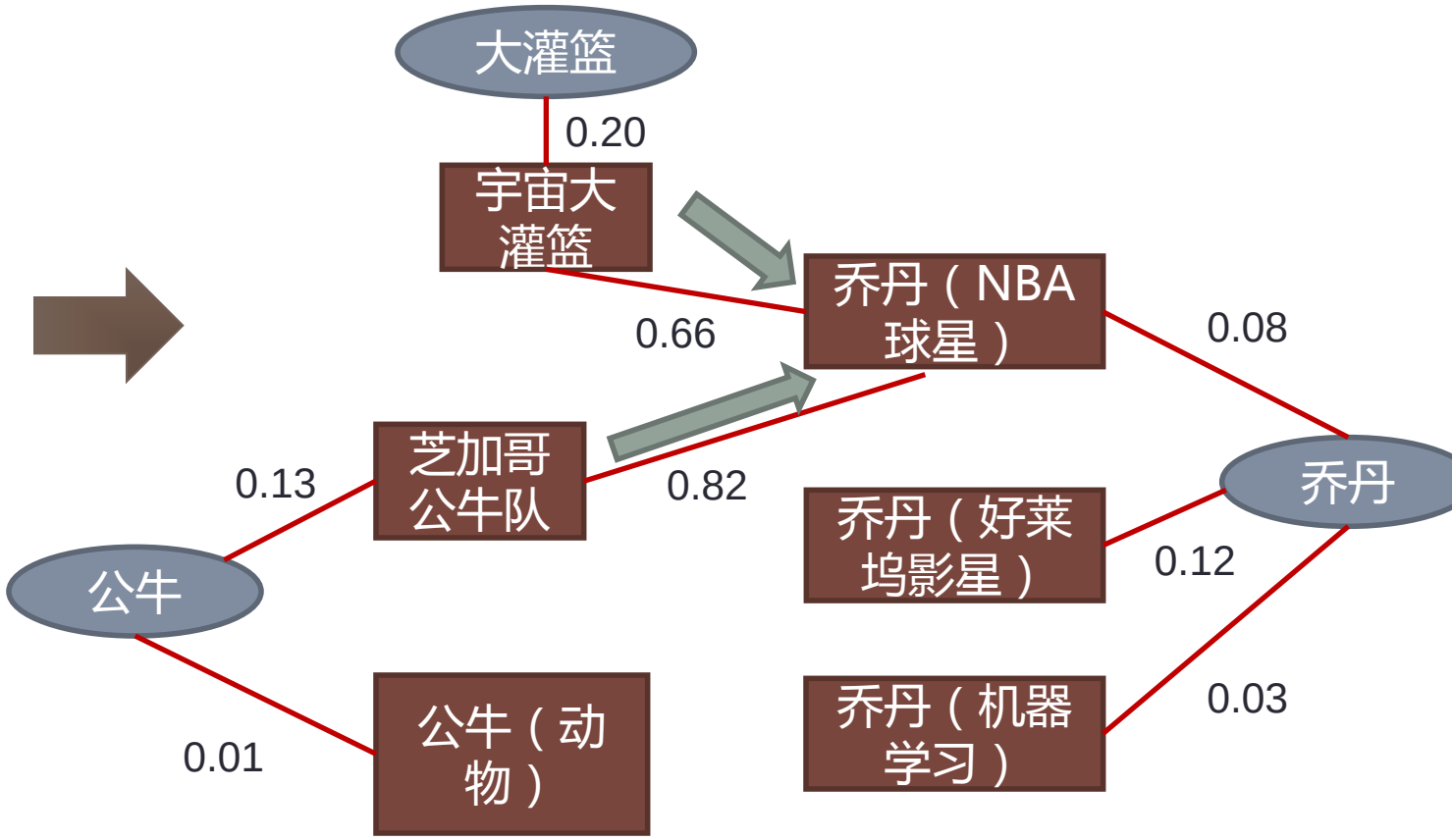
- 通过将证据在图上的依存结构上传递来协同增强证据直至收敛

$$r^{t+1} = (1-\lambda) \times T \times r^{t} + \lambda \times s$$

- r^{t+1}：在时间 t+1的证据
- T：证据传递率矩阵(Referent Graph的归一化相邻矩阵)
- λ：证据重分配率
- s：初始证据

基于图的协同推断

在公牛期间，乔丹出演了电影大灌篮



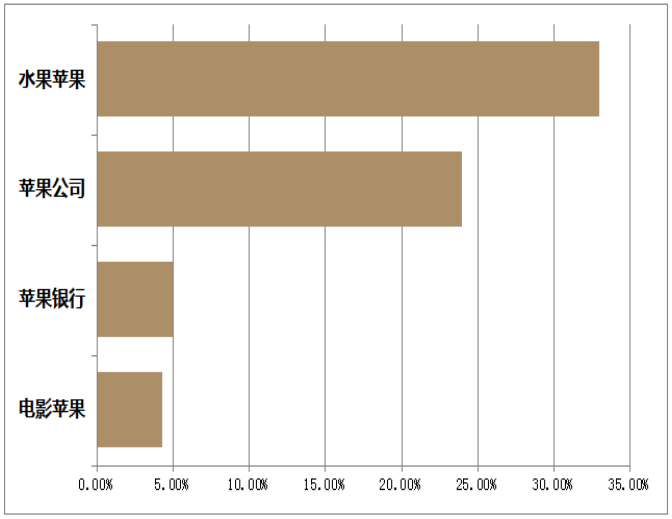
基于图的协同推断

实体	宇宙大灌篮	芝加哥公牛队	乔丹（NBA球星）
链接概率	35%	23%	5%
链接概率（增强后）	21%	30%	46%

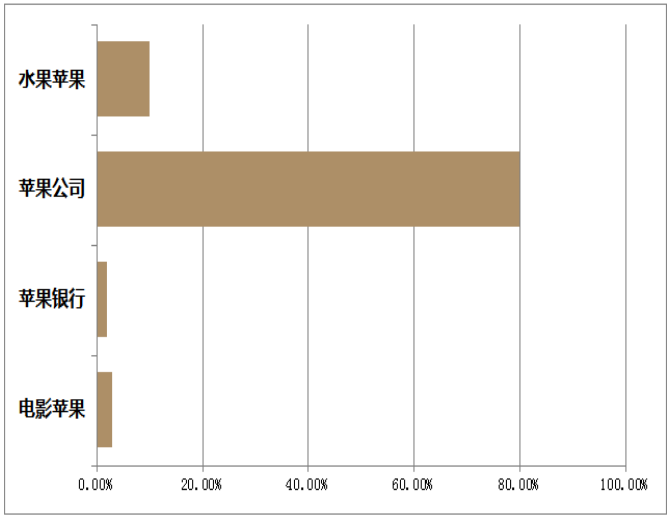
实体	公牛（动物）	乔丹（机器学习）	乔丹（好莱坞演员）
链接概率	2%	5%	21%
链接概率（增强后）	0.2%	0.7%	3%

实体链接方法(3)：融合实体知识与篇章主题的链接

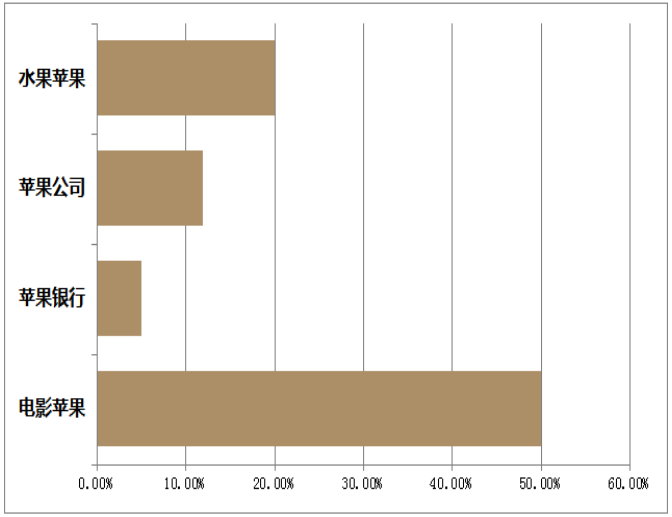
仅有实体知识是不够的



普通新闻



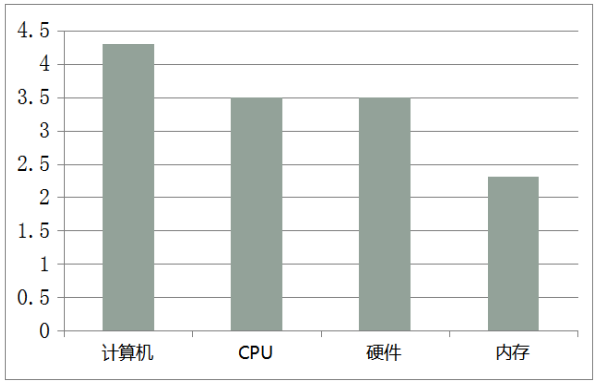
IT新闻



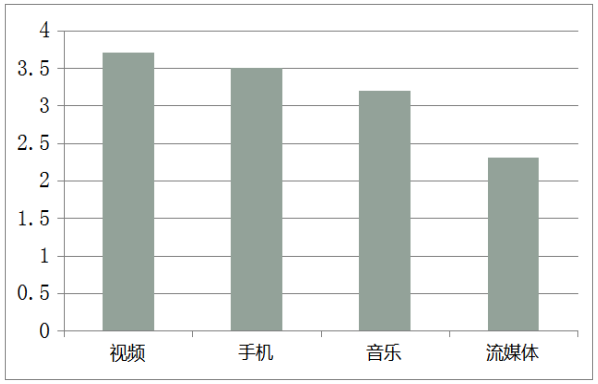
娱乐新闻

建模文本主题

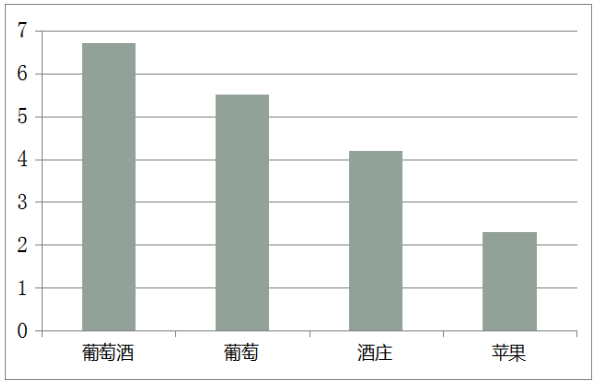
- 假设每一篇文本都有N个内在主题，每一个主题是实体的多项式分布
  - 苹果发布iPhone → {IT$^{0.41}$，手机$^{0.23}$， 苹果公司$^{0.33}$}
  - 苹果丰收→ {植物$^{0.45}$，水果$^{0.33}$，贸易$^{0.21}$}



计算机



娱乐

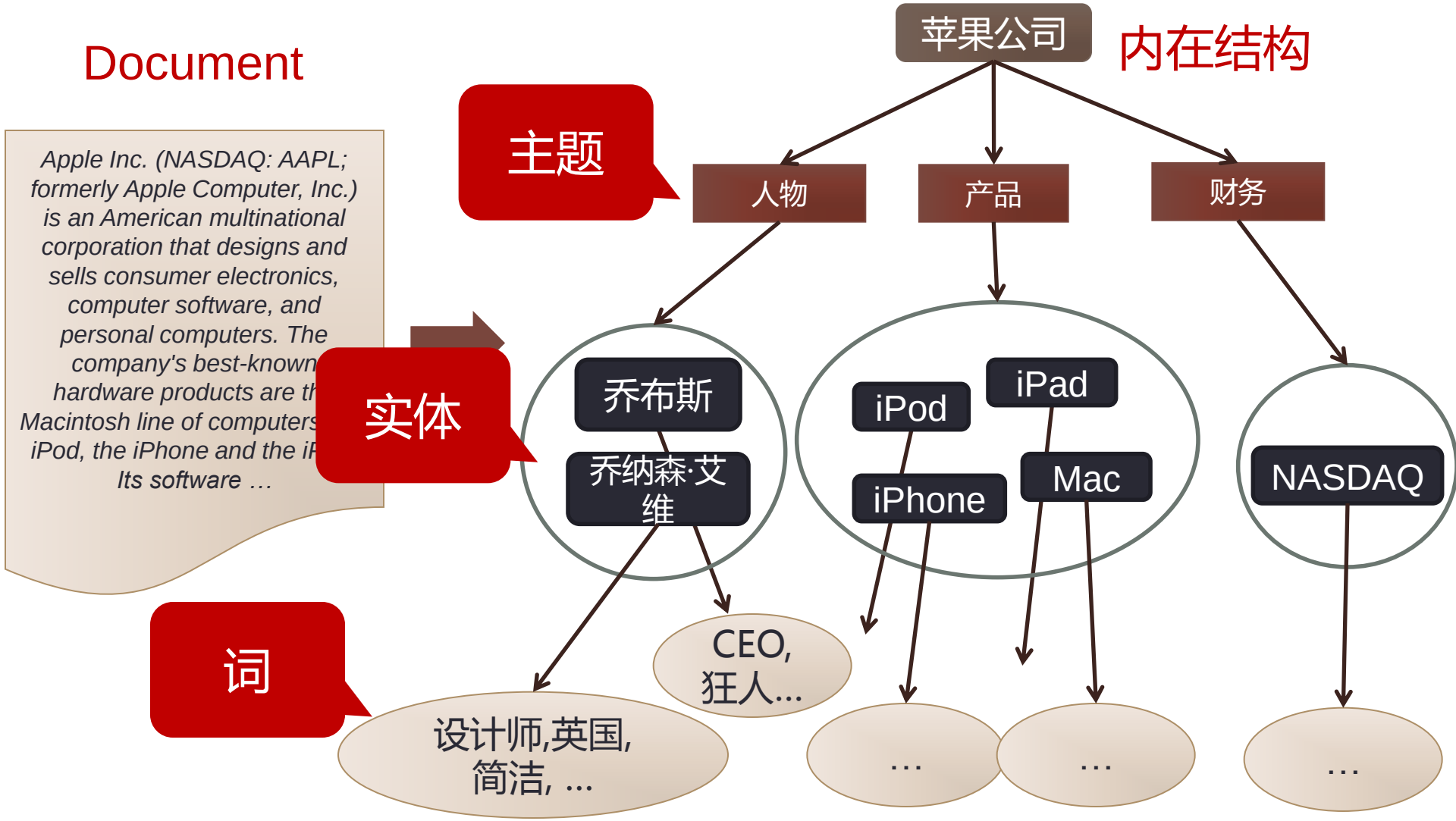


酒

基于实体-主题模型融合实体知识

Document

Apple Inc. (NASDAQ: AAPL; formerly Apple Computer, Inc.) is an American multinational corporation that designs and sells consumer electronics, computer software, and personal computers. The company's best-known hardware products are th Macintosh line of computers iPod, the iPhone and the iP Its software …



实验结果

	Precision	**Recall**	**F1**
Wikify!	*0.55*	*0.28*	*0.37*
EM-Model	*0.82*	*0.48*	*0.61*
M&W	*0.80*	*0.38*	*0.52*
CSAW	*0.65*	*0.73*	*0.69*
EL-Graph	*0.69*	*0.76*	*0.73*
Our Method	***0.81***	***0.80***	***0.80***

Table 1. The overall results on IITB data set

描述层知识融合

描述层知识融合-Schema Mapping

- 我们有一个集合的知识源，每一个知识源使用不同的分类体系和属性体系
- 需要将这些Schema（分类体系和属性体系）统一为一个全局的schema

基本资料

公司名称：国际商业机器公司
外文名称：IBM(International Business Machines Corporation)
总部地点：美国纽约州阿蒙克市
成立时间：1911年
经营范围：信息技术和业务解决方案
公司性质：外资
公司口号：停止空谈，开始行动！

现状

年营业额：1069亿美元（2011）
员工数：399,409人(2009年)

其他信息

现任CEO：Virginia Rometty（罗睿兰）

国际商业机器股份有限公司
International Business Machines Corporation

公司类型	上市公司（NYSE：IBM）
成立	1911年
董事长	弗吉尼亚·罗曼提（Ginni Rometty）
总裁	弗吉尼亚·罗曼提
首席执行官	弗吉尼亚·罗曼提
总部地点	美国纽约州阿蒙克市
标语口号	随需应变的业务（On Demand Business）
产业	电脑硬件、电脑软件
产品	IT服务、行业解决方案、服务器、其他计算机硬件等
营业额	▼$1045亿美元（2012年）
息税前利润	▲219亿美元(2012年)
净利润	▲176亿美元(2012年)
资产净值	▼188.6亿美元（2012年）
员工人数	434,246（2012年）
网址	www.ibm.com

百度百科	Wikipedia
公司类型	公司性质
成立时间	成立
公司口号	标语口号
年营业额	营业额

Schema Mapping难点

- 属性体系并非简单的一对一关系
  - 公司.成立 = {公司.成立时间，公司.成立地点，公司.成立方式}
  - 出生={出生日期，出生地点}
  - 出生日期={出生年份，出生月，出生日}
- 需要综合利用多种类别的信息
  - 属性的语义信息
    - 成立={创立，建立}
    - 出生={诞辰，诞生}
  - 属性的值分布信息
    - 出生日期的主要值为时间
    - 总部的主要值为机构
  - 属性的联合分布
    - 出生日期+出生地点+职业+单位 =>人

Schema Mapping解决方案

- 建立一个全局的Schema
  - 例如，以Freebase的Schema作为基准
- 利用一个集合的Base learners，将不同知识源中的schema与全局Schema进行映射
- 使用Meta-Learner来综合利用Base learner的分类结果并利用属性的联合分布信息，从而得到最终的Schema mapping全局结果

Schema Mapping样例

- Base Learner
  - 训练数据：
    - 人物(出生，北京)
    - 人物(生日，1999-1-2)
    - 公司(总部所在地，北京)
  - mapping
    - (出生地点，北京) => (出生，0.8)， (总部所在地，0.2)
- Meta Learner
    - 人物(出生，北京)
    - 人物(生日，1999-1-2)
    - 公司(总部所在地，北京)
    - => P（出生|生日）= 0.5,P(总部所在地|出生) = 0.001
  - (出生地点，北京)+(生日，1991) => (出生，0.9)， (生日，1.0)

Schema Mapping的挑战

- 建立全局Schema的标准是什么？如何建立统一的全局Schema？
- Scalable Schema Mapping算法
- Schema和自然语言表述之间的关系？描述一个特定Schema的表达方式有哪些？
  - 人物.出生日期
    - PER 出生于 Date
    - Date 是 PER 的诞辰
    - Date 哪一天，PER 的母亲生下来他。
    - ....
- Schema之间的蕴含关系
  - 公司.创始人 => 公司.员工
  - 收购（公司，公司）=>合并（公司，公司）

知识验证

知识验证

- 知识图谱构建不是一个静态的过程，需要
  - 及时更新动态知识
  - 加入新知识
- 需要判断新知识是否是否正确？与已有知识是否一致？
- 黄河长度是多少？
  - 黄河全长5494公里（知道）
  - 黄河全长5464公里（百科全书）
  - 黄河全长5464公里（问问）

知识验证证据（1）

- 权威度
  - 权威度高的信息源更有可能出现正确的答案
  - 百科全书 > 知道 ~= 问问
- 冗余度
  - 正确的答案更有可能出现
  - 黄河 + 5494 出现 39,600次
  - 黄河 + 5464 出现 338,000次
- 多样性
  - 正确的答案会以不同的方式表达

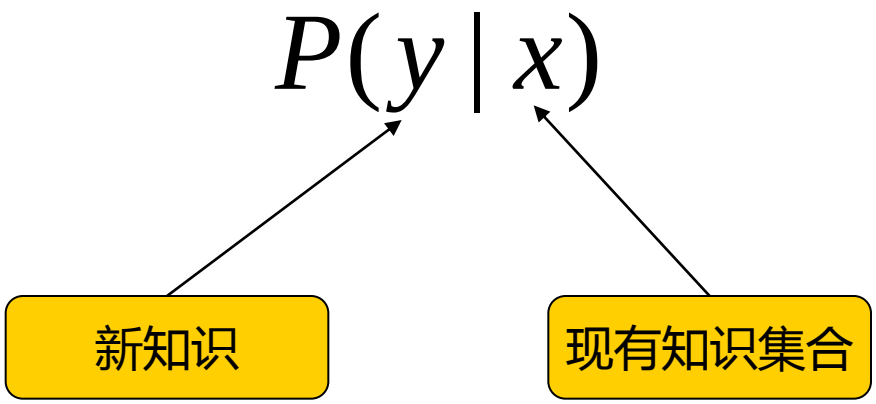
知识验证证据（2）

- 一致性
  - 正确的答案应当与其它知识相容无冲突

- 例子
  - 黄河是世界第5大河
  - 密西西比河是第4大河，长6262公里
  - 澜沧江全长4880公里，是第5大河
  - 4880公里 < 长度（黄河） <6262公里

知识验证的统计模型

- 计算新知识与现有知识相容的可能性概率

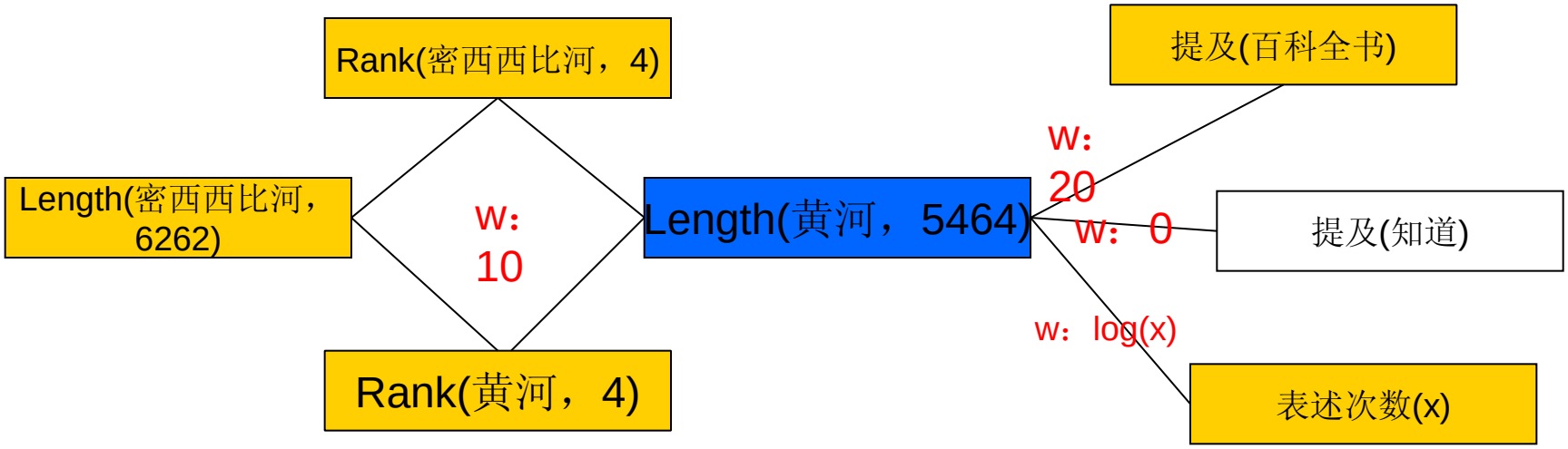


一种解决方案-马尔科夫逻辑网络

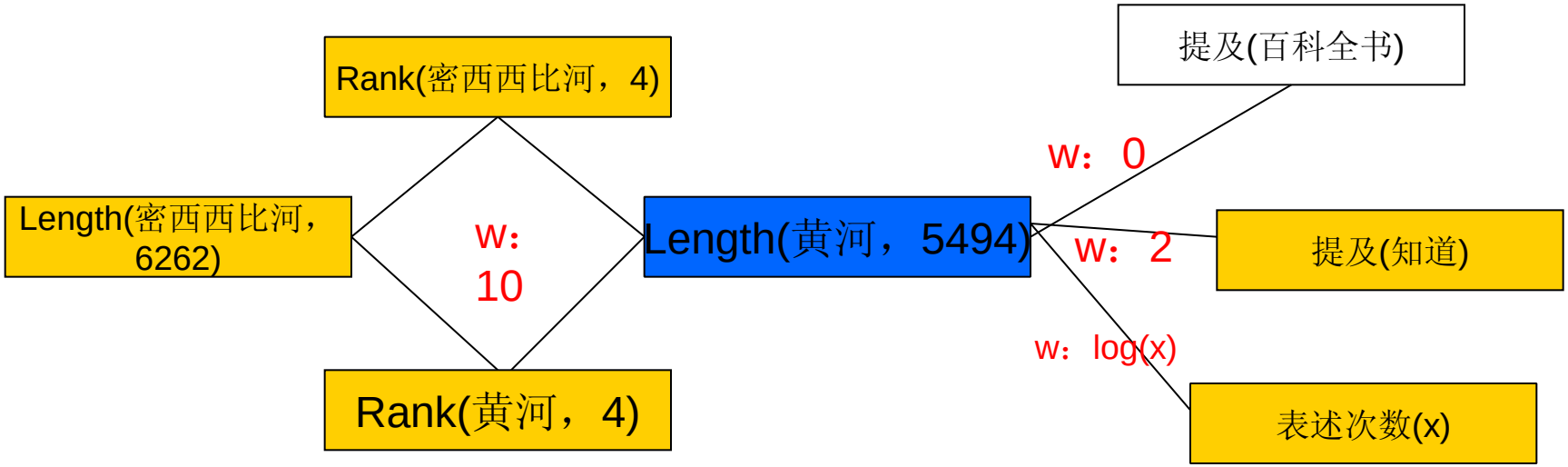
- 将知识和知识之间的约束建模为逻辑规则
  - 河流(r1) && 河流(r2) && 长度排名低于(r1，r2) => Length(r1)<Length(r2)
  - 提及(x, kb)
- 对这些规则赋予权重表示违反该条规则的代价
  - Reference(Length(黄河，5464公里)，百科全书)：10
  - Reference(Length(黄河，5494公里)，知道）：2
  - 10>2表示百科全书出现错误的可能性小于知道，因此P(Length(黄河，5464公里)) > P(Length(黄河，5494公里))

基于MLN的知识验证

- 所有陈述按逻辑规则相互链接
- 一条知识与当前知识图谱的相容性取决于其违反逻辑规则的多少和重要性



可能性正比于e^(10+20+0 + log(338,000)) ~= e^(35.5)



可能性正比于e^(10+0+2 + log(39,600)) ~= e^(16.6)

P(Length(黄河,5464)) > P(Length(黄河，5494))

知识验证挑战

- 如何发现知识之间的有效约束规则/联系
- 学习规则之间的约束强度
- 多源异构证据的统一建模
- 面向海量数据的可扩展Inference方法

总结

中文知识图谱构建

- 中文知识图谱的构建
  - 需要充分利用现有的知识资源
  - 处理Web去中心化结构带来的知识分散、异构、冗余、噪音、非确定性和非完备问题
- 解决方案
  - 多源知识融合：充分利用现有知识库，融合这些分散、冗余和异构的知识，作为构建中文知识图谱的出发点
  - 新知识验证：对新加入的知识进行验证，保证新知识与知识图谱的一致性和准确性，保证知识的持续更新

挑战

- **低成本**：Unsupervised/Distant Supervised的融合和验证算法
- **大规模**：处理Web规模的知识问题需要Scalable Inference Algorithm
- **不确定性**：信息抽取结果置信度的估计，基于噪音知识的推理系统
- **结构**：不仅仅抽取知识，而且能够分析知识的结构，如Ontology，Taxonomy…
- **部署**：多源、异构、演化文本上信息抽取模块的自动更新、管理和学习

分析

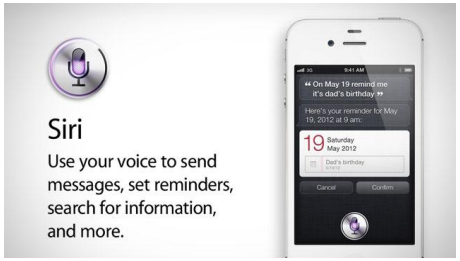
知识获取

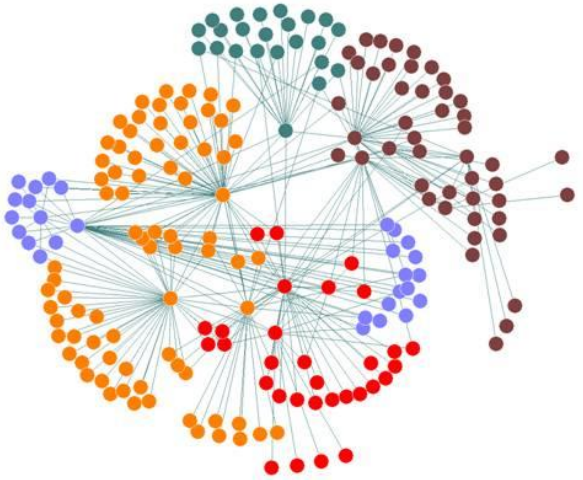
Web & Text

知识图谱

综合

解析





敬请大家批评和指导！

lesunle@163.com
xianpei@nfs.iscas.ac.cn