

实体链接：从文本到概念

韩先培

xianpei@nfs.iscas.ac.cn

信息检索研究室, 中国科学院软件研究所

提纲

- ▶ 背景
 - ▶ 应用场景示例
 - ▶ 关键技术
 - ▶ 引用表构建
 - ▶ 实体知识构建
 - ▶ 链接推理算法
 - ▶ 总结
-



信息过载



缺点:

- 在找到有用的信息之前，我们需要处理更多的无用信息
- 找到多少有用的信息取决于我们从沙子里淘金的本领

优点:

- 只要找，总是能够找到

主要原因：

- 自然语言表达的多样性
- 自然语言表达的歧义性

自然语言表达的多样性

- ▶ 同一意义可以以多种不同方式表达
 - ▶ 找全信息很难
- ▶ 自然语言表达的弹性
 - ▶ 科比今天发挥的不错
 - ▶ 科比今天终于吃了顿饺子
- ▶ 人与人之间表达的差异性和选择性

科密	科黑
科比, 小飞侠 黑曼巴 科神 ...	科比, 科铁 蜗壳 (我科) 建军, 毛八一 唠嗑



语言的歧义性

- ▶ 同一个词语、词组、句子在不同的上下文中有多种不同的意义
 - ▶ 中关村的苹果不错 → 苹果电脑
 - ▶ 新发地的苹果不错 → 水果苹果
- ▶ 找到准确的信息很难, 特别是长尾的信息
 - ▶ 我想找我的高中老师李强



► 但是全国共有 230717 个李强...



实体链接提供了一种解决方案

- ▶ 处理自然语言的多样性问题
 - ▶ 同一意义的不同表达
- ▶ 处理自然语言的歧义性问题
 - ▶ 同一表达的不同意义
- ▶ 手段：通过将自然语言中的文本与知识库中的条目进行链接



实体链接 — 例子

文本

乔丹

:

美国
NBA

著名
篮球
运动员，他为

联盟

带来至少100
亿的收入，也

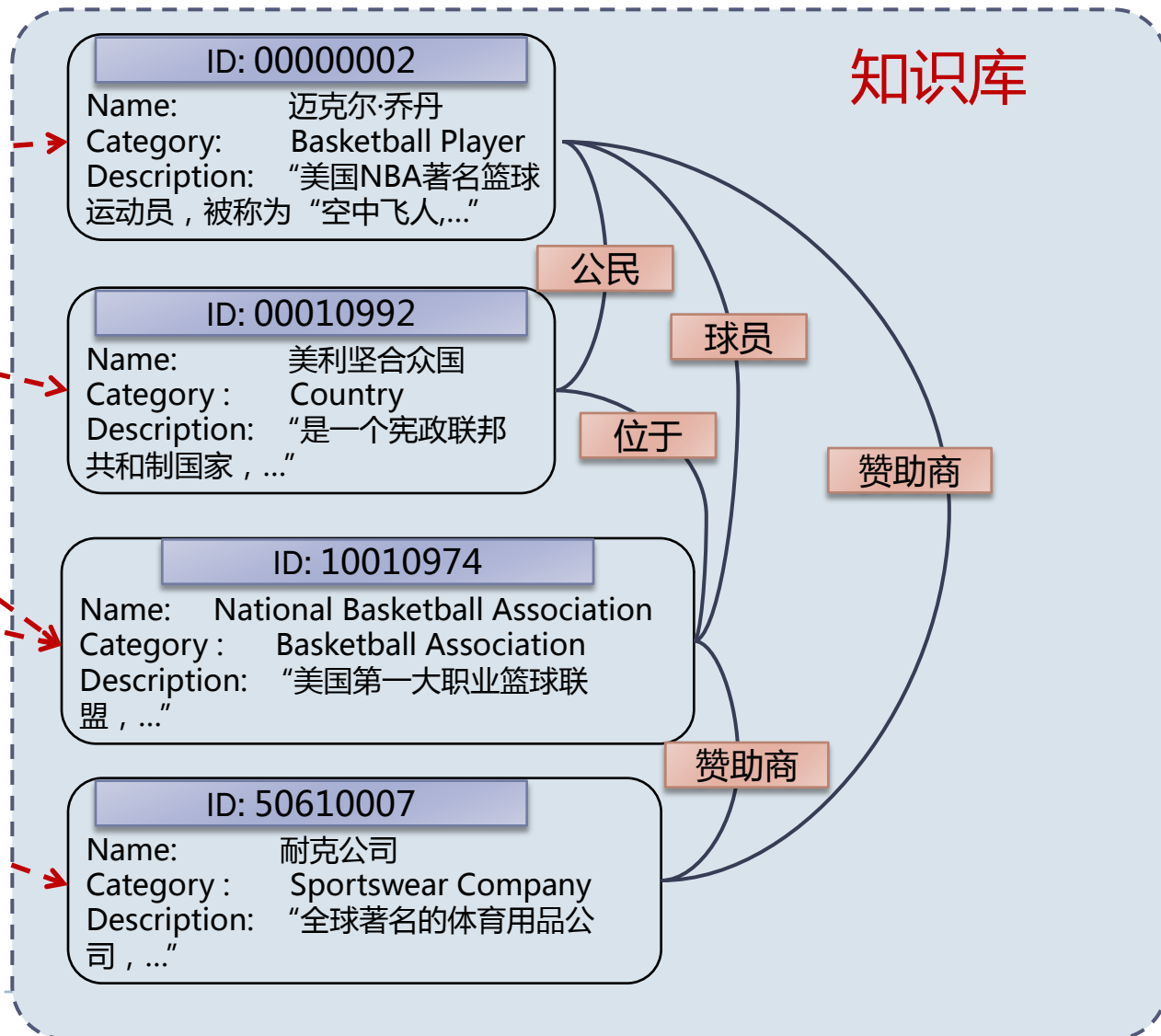
把

耐克公司

从一家小公司

...

知识库



提纲

- ▶ 背景
 - ▶ 应用场景示例
 - ▶ 关键技术
 - ▶ 引用表构建
 - ▶ 实体知识构建
 - ▶ 链接推理算法
 - ▶ 总结
-

阅读增强

乌拉圭总统为苏神辩护：我没看见他咬任何人

北京时间6月26日消息，苏亚雷斯咬人事件发生后，乌拉圭方面不惜与世界为敌，全力为自己的头号球星进行辩护，甚至动用了自己的总统！

[wù lǎ guī]

乌拉圭

乌拉圭（西班牙语：República del Uruguay，东南濒大西洋。居民的亚。1825年8月25日从巴西帝国

乌拉圭属温带气候，以优美白

乌拉圭的人均国内生产总值不



西，

石之

国

中文名称	乌拉圭东岸共和国
英文名称	The Oriental Republic of Uruguay
简称	乌拉圭
所属洲	南美洲
首都	蒙得维的亚
主要城市	蒙得维的亚，萨尔托

人口密度	16.65人/平方公里（2011年）
主要民族	88%白人
主要宗教	罗马天主教
国土面积	176215平方公里

何塞·穆希卡

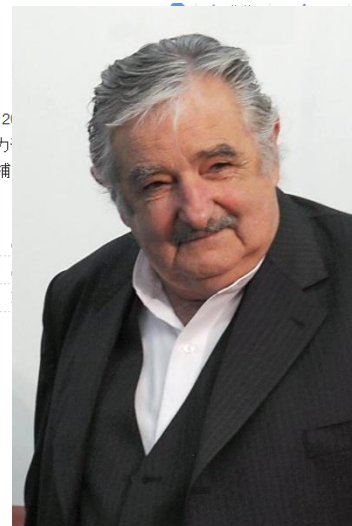
编辑

何塞·穆希卡（José Mujica），乌拉圭现任总统。1935年6月2日在那里度过了童年和少年的生活。他曾是一名自行车运动员，效力于帕马罗斯“游击队”的行动，并同100多名游击队员一起被军政府逮捕，曾身中六弹，险些身亡。

中文名	何塞·穆希卡
外文名	José Mujica
国 籍	乌拉圭

目录

- 1 基本介绍
- 2 竞选总统
- 3 当选总统
- 4 宣誓就职
- 5 世界最穷总统



雷纳，
参与图
，穆希卡

路易斯·阿尔贝托·苏亚雷斯

苏亚雷斯（路易斯·阿尔贝托·苏亚雷斯）即 路易斯·阿尔贝托·苏亚雷斯（Luis Alberto Suarez，

前锋，现效力于英格兰利物浦足球俱乐部。

苏亚雷斯职业生涯开始于乌拉圭民族足球俱乐部，2007年，夺得荷甲金靴奖、荷兰足球先生等荣誉。

2011年1月加盟利物浦队，2013/2014赛季，苏亚雷斯当选年度最佳球员 [2] 和英格兰职业球员协会(PFA)年度最佳球员。

苏亚雷斯2007年入选乌拉圭国家队，2010年参加南非世界杯并当选美洲杯最佳球员 [5]。

2014年6月25日巴西世界杯D组第三轮比赛，乌拉圭队国际足联表示要对苏亚雷斯咬人事件进行调查，禁赛的最



► 跨社交网络的信息集成

- ▶ 将所有与特定实体相关的信息与其进行链接
- ▶ 如新浪微博、Twitter、新闻、水木、等等

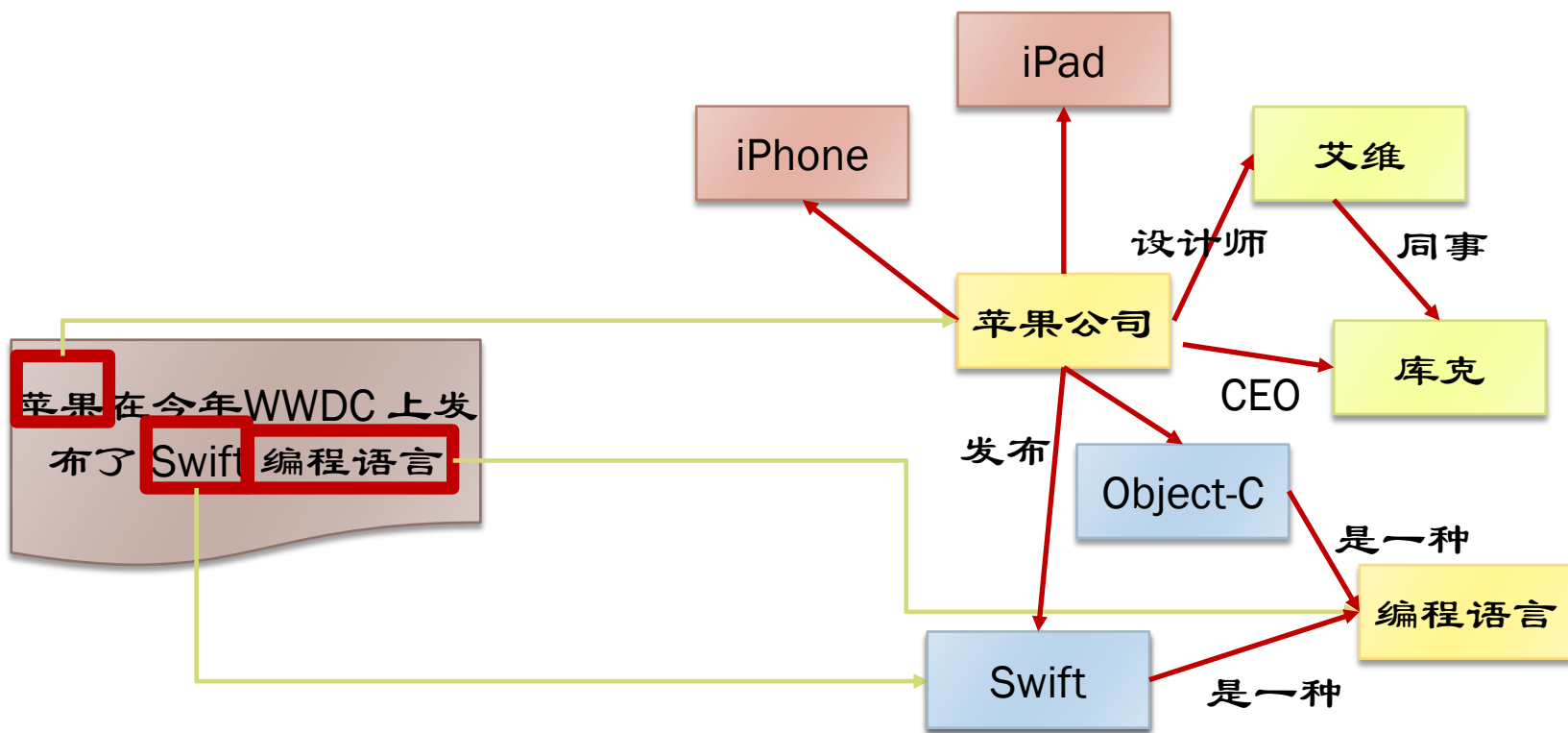


baike.baidu.com/ 2014-06-06 ▼



知识库的构建

- 将文本中的实体与知识库中的实体链接，并将知识库中没有的知识补全到知识库中



提纲

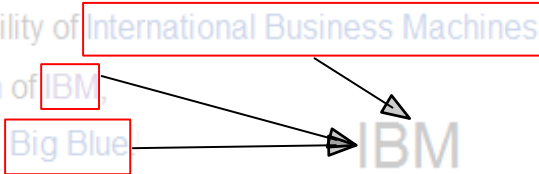
- ▶ 背景
 - ▶ 应用场景示例
 - ▶ 关键技术
 - ▶ 引用表构建
 - ▶ 实体知识构建
 - ▶ 链接推理算法
 - ▶ 总结
-

引用表构建

- ▶ 引用表存储一个名字所有可能指向的实体
 - ▶ 名字 → 实体 的映射关系
- ▶ 构建方法：锚文本数据挖掘

名字	目标实体	次数
苹果	水果苹果	10000
	苹果公司	3000
	苹果电脑	2030
	苹果《电影》	200
	苹果银行	10
AI	Artificial intelligence	581
	Game artificial intelligence	48
	Ai (singer)	10
	Angel Investigations	9
	Strong AI	3
	Characters in the Halo series	2
▶	...	...

IBM Rochester is the facility of International Business Machines
IBM Research, a division of IBM
a play on IBM's nickname, Big Blue



```
graph LR; A[International Business Machines] --> D[IBM]; B[IBM] --> D; C[Big Blue] --> D;
```

提纲

- ▶ 背景
 - ▶ 应用场景示例
 - ▶ 关键技术
 - ▶ 引用表构建
 - ▶ 实体知识构建
 - ▶ 链接推理算法
 - ▶ 总结
-

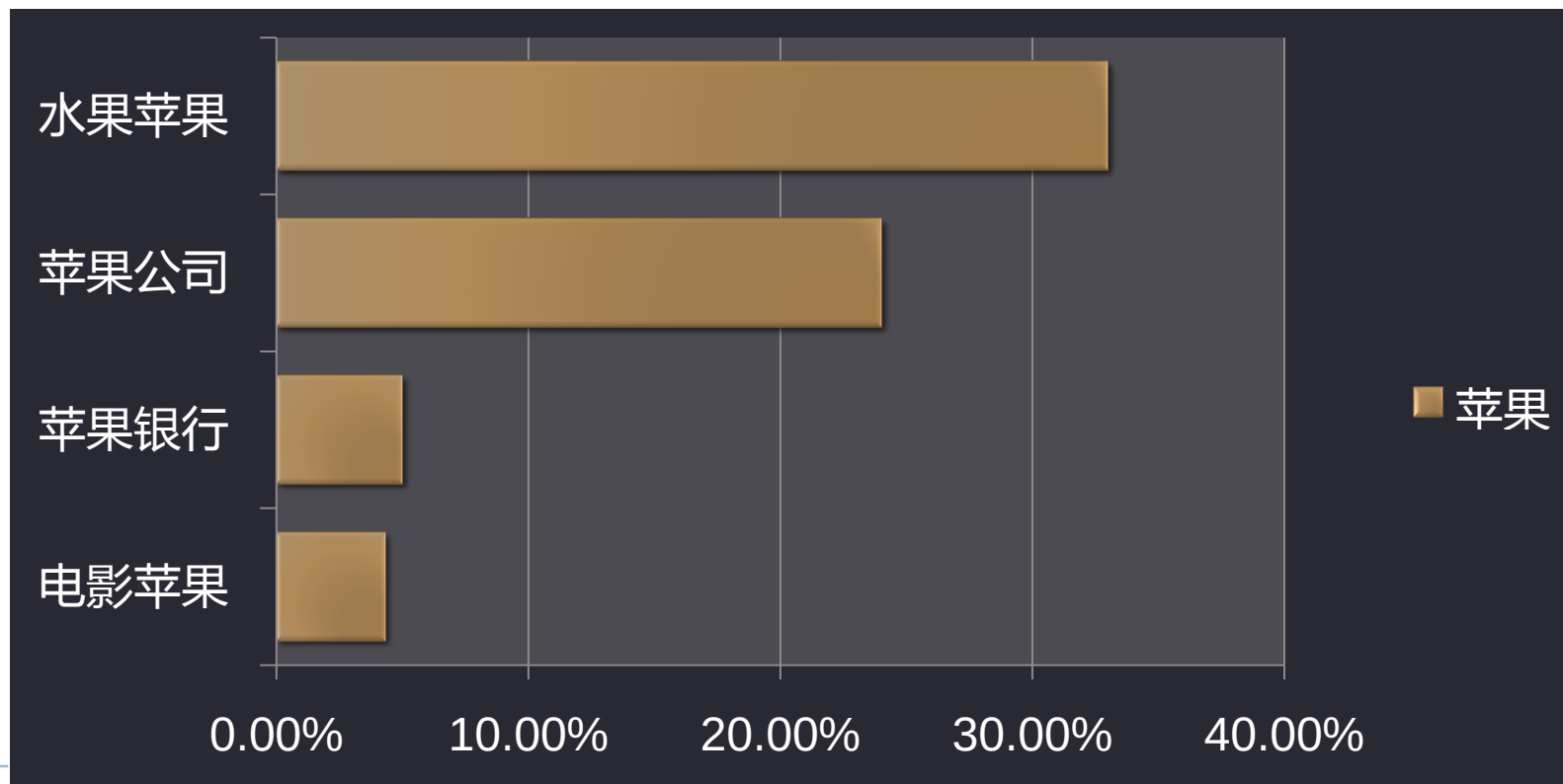
实体知识

- ▶ 实体知名度
- ▶ 实体上下文
- ▶ 实体语义关联度
- ▶ 文章主题



实体知名度

- ▶ 表示一个实体被人们知道了解的程度
- ▶ 高知名度的实体更有可能在文章中被提起



▶ 特定实体的上下文规律性

- [illegible]

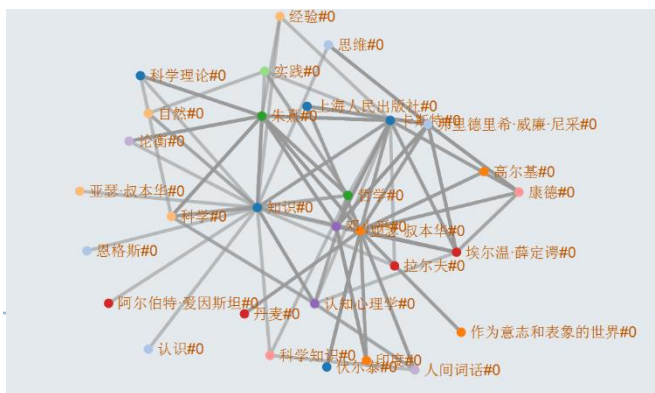
Copyright 2012

tagxedo.com

苹果银行

实体语义关联度

- ▶ 捕捉实体和实体之间的语义关系
- ▶ 相关实体更容易同时出现在一篇文章中
 - ▶ 苹果电脑与乔布斯、iPad、iPhone、库克
 - ▶ 苹果与葡萄、桃子、苹果汁、酒
 - ▶ 苹果电影与范冰冰、华星、华谊
- ▶ 如何衡量两个实体之间的相关度
 - ▶ 在知识网络中的距离
 - ▶ 在文章中共现的次数



	贝叶斯网络	芝加哥公牛
机器学习	0.74	0.00
NBA	0.00	0.71

文章主题

- ▶ 一篇文章中的实体应当与其主题相关
 - ▶ 苹果公司更容易出现在IT相关主题的文档中
 - ▶ 水果苹果更容易出现在吃或农业相关的文档中
 - ▶ 电影《苹果》倾向于出现在娱乐相关的新闻中

苹果公司

计算机

媒体

软件

Topic(Computer)	Topic(Video)	Topic(Software)
Computer	Video	Computer software
CPU	Mobile phone	Microsoft Windows
Hardware	Mass media	Linux
Personal computer	Music	Web browser
Computer memory	Television	Operating system

Table 2. The 3 topics where the *Apple Inc.* has the

水果苹果

酒

食物

植物

Topic(Wine)	Topic(Food)	Topic(Plant)
Wine	Food	Plant
Grape	Restaurant	Flower
Vineyard	Meat	Leaf
Winery	Cheese	Tree
Apple	Vegetable	Fruit

Table 3. The 3 topics where the fruit *Apple* has the

一些有用但是难以构建的知识

- ▶ 作者知识领域
 - ▶ 倾向性偏好
 - ▶ 科黑 vs 科密
 - ▶ 媒体偏好
 - ▶ 新华社 vs 微博
 - ▶ 地点偏好
 - ▶ 北京 vs 广州
 - ▶ 宿舍 vs 图书馆
 - ▶ 精准度与构建成本之间的权衡
-



提纲

- ▶ 背景
 - ▶ 应用场景示例
 - ▶ 关键技术
 - ▶ 引用表构建
 - ▶ 实体知识构建
 - ▶ 链接推理算法
 - ▶ 总结
-

链接推理算法

- ▶ 链接推理算法就是综合实体知识进行决策的过程
 - ▶ 中关村的**苹果**不错 → 水果苹果？苹果电脑？
 - ▶ 水果苹果和苹果电脑的上下文
 - ▶ 相关度(中关村, 水果苹果) = 0.1
 - ▶ 相关度(中关村, 苹果电脑) = 0.7
 - ▶ 局部推理
 - ▶ 全局推理
-

局部推理

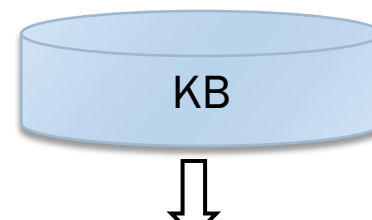
- ▶ 考虑单个实体的上下文，不考虑文章中其它实体对该实体的影响
 - ▶ 文本相似度
 - ▶ 统计生成模型
 - ▶ Learning To Rank



实体-提及模型(EM Model)

在实体-提及模型中,每一个命名性提及 m 都被建模为通过下述生成过程 (generative story)产生的样本:

1. EM Model根据实体的知名度 $P(e)$ 选



实体的知名度知识、名字知识和上下文知识依次被建模为概率分布 $P(e)$, $P(s|e)$, $P(c|e)$

3. EM Model根据实体的上下文知识 $P(c|e)$ 输出提及 m 的上下文 c



基于实体-提及模型的实体链接

- ▶ 基于上述模型, 实体 e 是提及 m 目标实体的概率:

$$P(m, e) = P(s, c, e) = P(e)P(s | e)P(c | e)$$

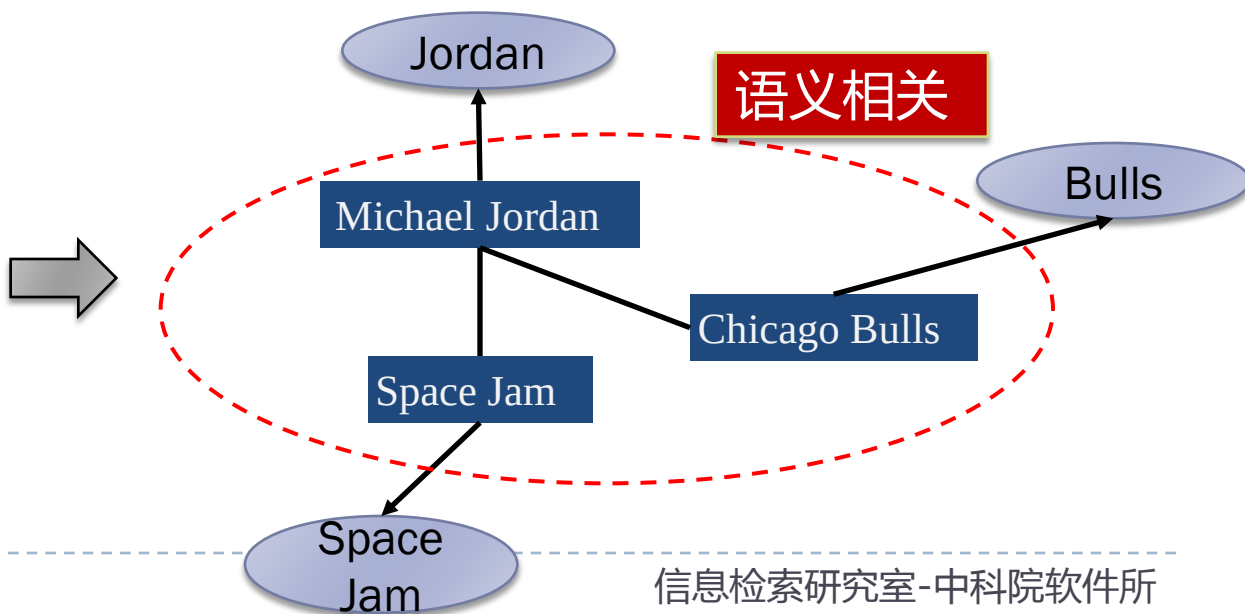
- ▶ 模型选择能最大化条件概率 $P(e|m)$ 的实体 e 作为其提及 m 的目标实体

$$e = \operatorname{argmax}_e \frac{P(m, e)}{P(m)} = \operatorname{argmax}_e P(e)P(s | e)P(c | e)$$

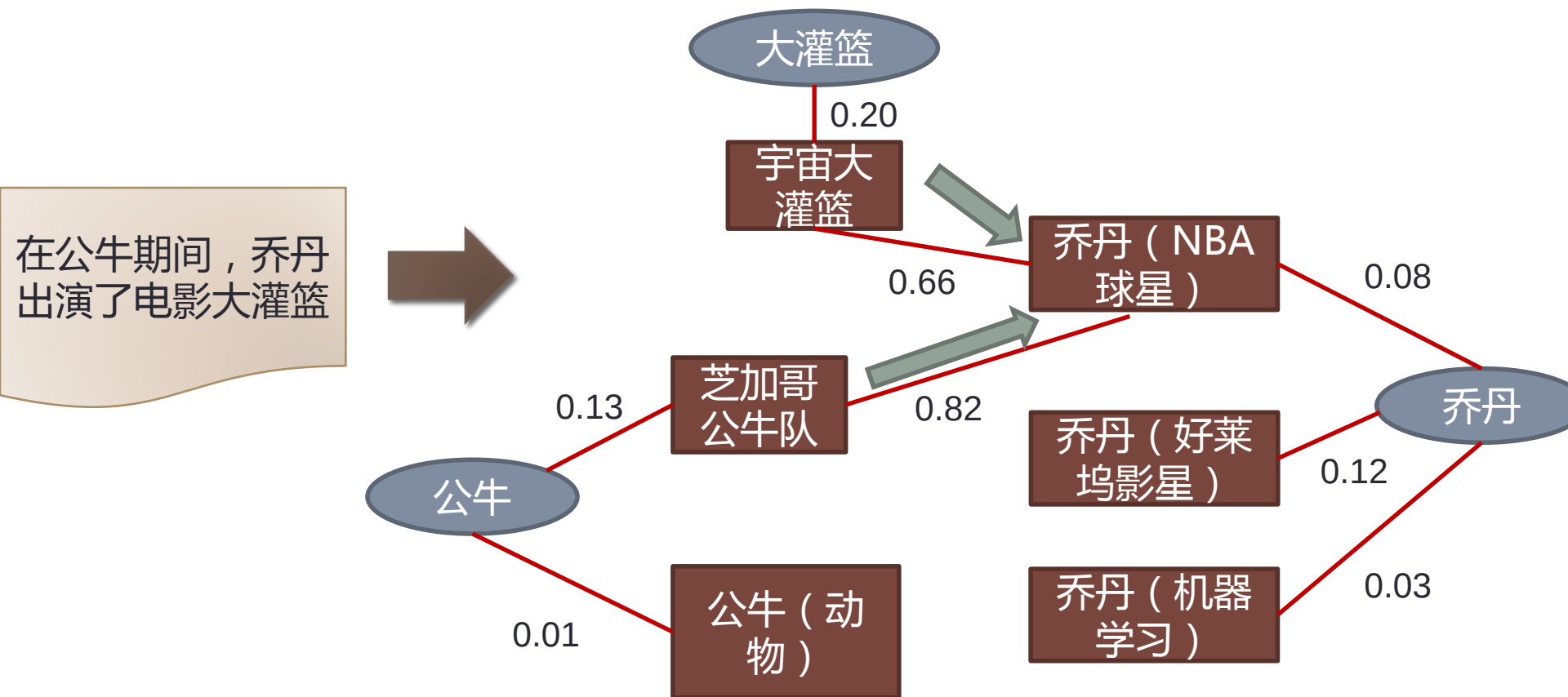
全局链接

利用目标实体之间的语义关联，协同链接单篇文本内的所有提及能有效提升实体链接性能

During his standout career at Bulls, Jordan also acts in the movie Space Jam.



基于图的协同推断



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

在时间
t+1的证据

证据传递率矩阵(Referent
Graph的归一化相邻矩阵)

证据重
分配率

初始
证据

篇章主题-实体-上下文词的协同推断

内在结构

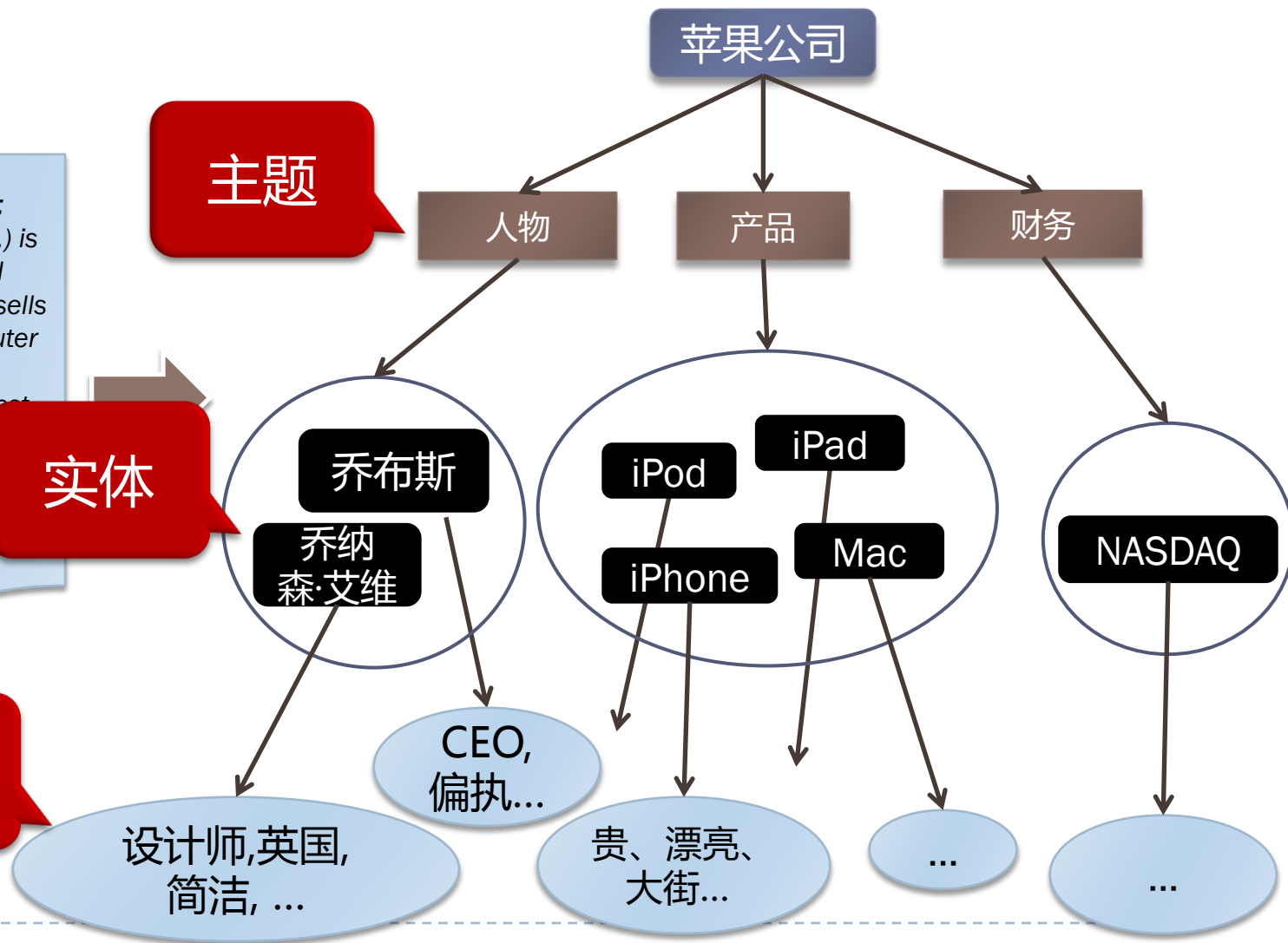
Document

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 the Macintosh line of computers, the iPod, the iPhone and the iPad, and the Mac OS X software ...

主题

实体

词



总体性能

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

Table 1. The overall results on IITB data set

局部
推理

全局
推理



提纲

- ▶ 背景
- ▶ 应用场景示例
- ▶ 关键技术
 - ▶ 引用表构建
 - ▶ 实体知识构建
 - ▶ 链接推理算法
- ▶ 总结



总结与展望

- ▶ 实体链接是一项解决自然语言歧义和多样性的有效技术，性能在一定程度上达到实用水平
- ▶ 加强对长尾(long tail)实体的处理
 - ▶ 现有的消歧系统通常只对知名实体进行消歧
 - ▶ 将Web文本中实体与长尾上的实体，如sina微博、Facebook上的用户相连，可创造广阔的应用前景
- ▶ From **Extraction** to **Discovery**
 - ▶ 蝙蝠侠—三部曲之终章 → 这是新出来的电影，还是老电影、漫画、或其他？
 - ▶ 可以与知识库构建相互补充，从而滚动式的构建大规模知识图谱

百度技术沙龙

畅想

交流

争鸣

聚会

关注我们：t.baidu-tech.com

资料下载和详细介绍：infoq.com/cn/zones/baidu-salon

“畅想•交流•争鸣•聚会”是百度技术沙龙的宗旨。百度技术沙龙是由百度与InfoQ中文站定期组织的线下技术交流活动。目的是让中高端技术人员有一个相对自由的思想交流和交友沟通的平台。主要分讲师分享和OpenSpace两个关键环节，每期只关注一个焦点话题。

讲师分享和现场Q&A让大家了解百度和其他知名网站技术支持的先进实践经验，OpenSpace环节是百度技术沙龙主题的升华和展开，提供一个自由交流的平台。针对当期主题，参与者人人都可以发起话题，展开讨论。