

SEU 知识抽取-问题和方法

2019年11月11日 14:14

一、问题分析

• 知识抽取的场景（数据源）

- **（半）结构化文本数据**：百科知识中的Inforbox、规范的表格、数据库、社交网络.....
- **非结构化的文本数据**：网页、新闻、社交媒体、论文.....
- **多媒体数据**：图片、视频

• 从信息抽取到知识抽取

- **区别**：信息抽取获得结构化数据，知识抽取获得机器可理解和处理的知识（知识表示）
- **关系**：知识抽取建立在信息技术抽取基础上，都普遍利用到自然语言处理技术、基于规则的包装器和机器学习技术

• 知识抽取的挑战性

- 知识的不明确性 (ambiguous)
- 知识的不完备性 (incomplete)
 - 关系缺失
 - 标签/属性缺失
 - 实体缺失
- 知识的不一致性 (inconsistent)

二、知识抽取场景和方法

• 从关系数据库中抽取知识

◦ 抽取原理

- 表 (Table) --类 (Class)
- 列 (Column) --属性 (Property)
- 行 (Row) --资源/实例 (Resource/Instance)
- 单元 (Cell) --属性值 (Property Value)
- 外键 (Foreign Key) --指代 (Reference)

根据上述规则可将关系数据库转化为一个知识库。

◦ 抽取标准

- Direct Mapping
- R2RML

◦ 抽取工具

- D2R、Virtuoso、Orcle SW、Morph等

- **R2BML映射语言**

- 输入：数据库表 视图 SQL查询
- 输出：三元组
- 实例展示：

“员工”和“部门”两个关系数据库表

EMP			
EMPNO	ENAME	JOB	DEPTNO
INTEGER PRIMARY KEY	VARCHAR(100)	VARCHAR(20)	INTEGER REFERENCES DEPT (DEPTNO)
7369	SMITH	CLERK	10

DEPT		
DEPTNO	DNAME	LOC
INTEGER PRIMARY KEY	VARCHAR(30)	VARCHAR(100)
10	APPSERVER	NEW YORK

“员工”和“部门”两个关系数据库表映射的RDF

Example output data

```
<http://data.example.com/employee/7369> rdf:type ex:Employee.  
<http://data.example.com/employee/7369> ex:name "SMITH".  
<http://data.example.com/employee/7369> ex:department <http://data.example.com/department/10>.  
  
<http://data.example.com/department/10> rdf:type ex:Department.  
<http://data.example.com/department/10> ex:name "APPSERVER".  
<http://data.example.com/department/10> ex:location "NEW YORK".  
<http://data.example.com/department/10> ex:staff 1.
```

- **四个步骤：**
 - 抽取类
 - 抽取属性
 - 抽取实例
 - 建立类之间的关系
- **优点**
 - 转换规则简单
- **缺点**
 - 直接转换得到的知识库语义信息不足
 - 需要熟悉原数据库设计的专家进行知识库的优化

- **面向半结构化数据知识的抽取**

- **百科知识的抽取**
 - 大规模多语言百科知识图谱，维基百科的结构化版本，linked data核心数据集
 - 覆盖127种语言，2800w个实体，数亿三元组，支持数据集的完全下载
 - 固定模式对实体信息进行抽取，包括 abstract、infobox、category、pagelink等
- **YAGO**
 - YAGO整合了WikiPedia与WordNet
 - 覆盖多种语言，100w实体，1.2b三元组
 - 在YAGO2整合了GeoNames，增加了对时空信息的支持

- 通过规则对实体信息进行抽取与推断
- Infobox启发式规则
 - 人工定义映射规则，将同义属性统一表示
 - 每个属性定义domain和range，用于进一步推断和清洗
 - 类型推断（type heuristics）：优先首字母是名词且可数的推断

• 面向无结构化数据知识抽取

○ 问题

- 挑战：是当前知识图谱构建的技术瓶颈
- 关键技术
 - 实体识别
 - 关系抽取
 - 事件抽取
- pipeline的抽取过程会迅速降低知识的质量

○ 实体识别

- 抽取文本中的原子信息

- 人名
- 组织/机构
- 地理位置
- 时间日期
- 字符
- 金额
-

— 金额 时间 人名

—

北京时间3月23日0时50分许，美国总统特朗普在白宫正式签署对华贸易备忘录。特朗普当场宣布，将有可能对600亿美元中国出口商品征收关税。

地理位置

○ 关系抽取

- 关系抽取指实体间的语义关系

中国联通与中兴通讯正式签署SDN/NFV战略合作协议

2016-03-23

近日，中国联通与中兴通讯共同宣布，双方正式签署SDN/NFV战略合作协议。双方旨在通过协同创新更好地把握以SDN/NFV为代表的网络发展趋势，实现SDN/NFV的落地应用，推动中国联通新一代网络的构建，快速消

中兴通信中标联通规模最大100G项目

作者：李雁争 来源：上海证券报 2015-06-12 09:56 | 评论 |

- 事件定义

具体事件、地点、参与者等基本元素，可由某个动作触发或者状态改变而发生的一个图结构知识片段

- 事件抽取

从数据中抽取事件信息，并以结构化和语义化形式展现，例如事件发生时间、地点、原因、参与者等

外媒称中印边界对峙开始撤军 披露双方对峙缘由

[正文](#)[我来说两句 \(9961人参与\)](#)

2014-09-29 08:38:32 来源：环球网 作者：王斯图

[用手机看新闻](#)[保存到博客](#)[图](#)[品](#)

【环球网军事报道】据《印度快报》27日报道，印度外交部长苏亚斯在纽约表示，她与中国外长王毅会谈后决定，中印双方开始从边界撤军，到下周二撤完，双方军队将恢复到9月1日所在的位置，“糟糕的时期即将结束”。《纽约时报》称，至此，持续近三周的中印边界对峙宣告结束，双方兵力

多家西方媒体披露了此次对峙的起因。路透社25日发文称，本月初，印度军方在边界地区建起一个观察屋，从这里可以观察对面中国士兵的举动。这一行动激怒了中方，中国士兵之后开始铺设通往印度声称主权领土的公路，并要求印方拆除观察屋。印方拒绝了中国的要求，坚持了中国建造的公路，并增加了在这一地区的驻兵。“莫迪曾在竞选中承诺更加强硬的国家安全政策，中印上千名士兵在边界地区对峙说明，即便是强势的中国，也并非无所不能。”《纽约时报》26日称，拉达克地区官员坚持称，观察屋仅用于民用，但建设活动已经停止。

简单事件抽取

事件类型	会谈
触发词	会谈
参与者	苏亚斯、王毅
时间	-
地点	-
事件类型	撤军
触发词	撤军
参与者	中印双方
时间	周五
事件类型	建造
触发词	建起
参与者	印度军方
时间	-
事件类型	破坏
触发词	毁坏
参与者	印方
时间	-
地点	-