政府组织机构图谱v1.0

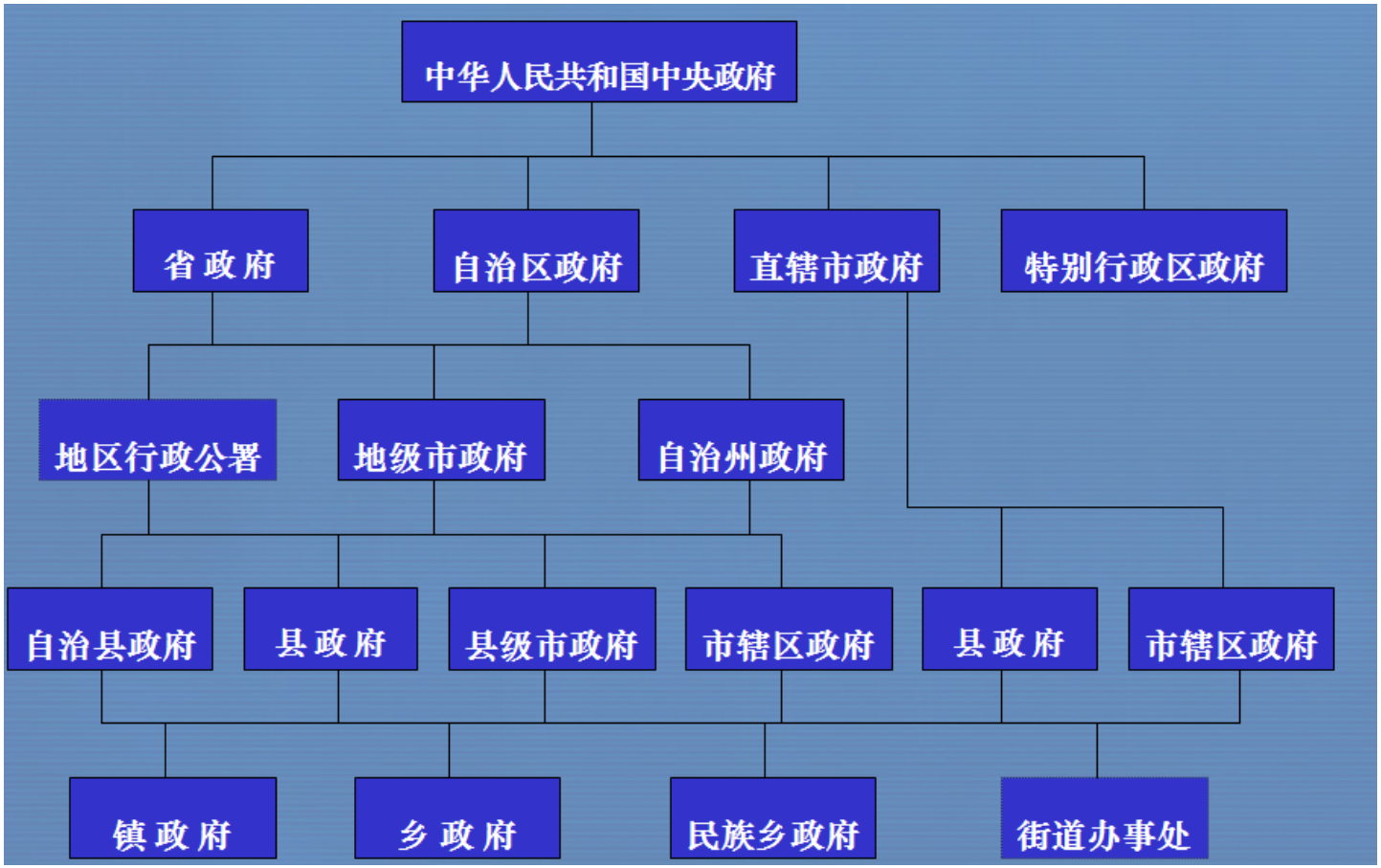
刘荣
清华大学信息技术学院语音和语言技术中心
北京市中科汇联信息技术有限公司
2015-01-26

目录

- 知识库框架
- 存储及查询方式
- 使用的工具
- 具体流程

政府组织结构图谱—知识库框架

1. 类型： 表示类别摘要概念，如政府，部，处，厅，局，科, 室



组织结构图谱—知识库框架

- 知识库框架设计（参照Freebase）

1. 类型： 表示类别摘要概念，如政府，部，处，厅，局，科, 室
2. 实体： 表示类别的具体实例。如济南市政府，山东省科技厅…
3. 属性：类别的一些属性。如电话，地址….

基于上述设计方案，整个知识库可以看做三元组的集合，即<实体, 属性, 属性值>的集合，其中属性值也可以是一个实体。如<济南市政府,地址，济南市**街道>
<济南市政府, 分支，济南市教育局>

组织结构图谱—知识库框架

- 知识库属性设计

1. 客户数据库，需要人工进行整理或D2R数据转换

2. 百度知道/企业搜素

2.1 利用专业知识库中的实体列表，确定每个用户查询Q中的实体类别，并将其中的实体名用类别名的英文名进行代替，得到Q'。

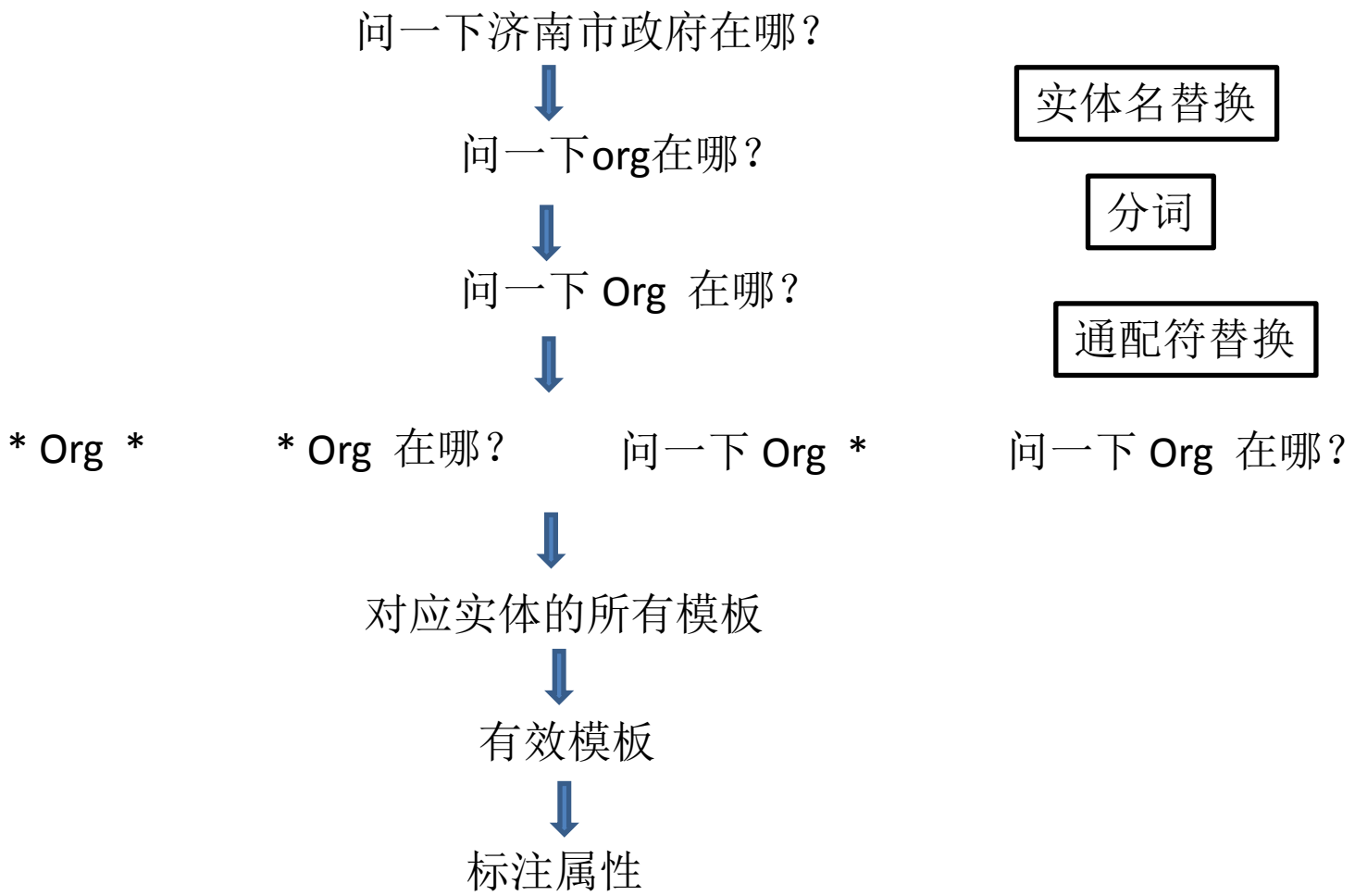
2.2 对替换后的用户查询Q'进行分词，记分词后除类别的英文名外，一共有k个词。分别枚举这k个词用通配符“*”替换，或者不用通配符替换的所有情况，一共可以生成$2k$个不同的模板

2.3 对于每个类别，统计包含该类别的英文名的所有模板，并按出现次数从高到低进行排序。

2.4对于每个类别，标注最常见的模板。如果模板为“有效模板”，则继续标注其对应的属性。如果该属性不包含在专业知识库的属性中，则将此用户关注的属性添加到对应类别的概要中。

需要说明的是，上述方法第4点中的“有效模板”是指能够明确表达用户查询意图的模板。比如“org的地址在哪?”就是“有效模板”，因为符合该模板的用户查询，可以确定其询问的为“机构”的“地址”。有些模板虽然出现次数较多，但却并不是“有效模板”。比如模板“org好呀！”，由于信息过少，无效模板。

组织结构图谱—知识库框架



组织结构图谱—知识库存储

- 基于关系数据库的存储和查询

组织结构图谱—知识库存储

- 基于关系数据库的存储和查询

1. 简单三列表

Prefix: y= http://en.wikipedia.org/wiki/

主体	属性	客体
y:Abraham_Lincoln	hasName	“Abraham Lincoln”
y:Abraham_Lincoln	BornOnDate	“1809–02–12”
y:Abraham_Lincoln	DiedOnDate	1865–04–15
y:Abraham_Lincoln	DiedIn	y:Washington_D.C
y:Washington_D.C	hasName	“Washington D.C.”
y:Washington_D.C	FoundYear	1790
y:Washington_D.C	rdf:type	y:city
y:United_States	hasName	“United States”
y:United_States	hasCapital	y:Washington_D.C
y:United_States	rdf:type	Country
y:Reese_Witherspoon	rdf:type	y:Actor
y:Reese_Witherspoon	BornOnDate	“1976–03–22”
y:Reese_Witherspoon	BornIn	y:New_Orleans,_Louisiana
y:Reese_Witherspoon	hasName	“ReeseWitherspoon”
y:New_Orleans,_Louisiana	FoundYear	1718
y:New_Orleans,_Louisiana	rdf:type	y:city
y:New_Orleans,_Louisiana	locatedIn	y:United_States

```
SELECT ?name                                  //查询返回的变量值
WHERE
{ ?m  <hasName>   ?name.                      //查询条件
  ?m <BornOnDate>  “1809–02–12” .
  ?m <DiedOnDate>  “1865–04–15” .
}
```

```
SELECT  T3.Subject
FROM  T as T1, T as T2, T as T3
WHERE T1. Property = “BornOnDate”
and T1.Object= “1809–02–12”
and T2.Property= “DiedOnDate”
and T2.Object= “1865–04–15”
and T3. Property = “hasName”
and T1.Subject = T2.Subject
and T2. Subject= T3.subject
```

图4　转换以后的SQL查询

组织结构图谱—知识库存储

- 基于关系数据库的存储和查询

2. 水平存储

Subject	rdf:type	hasName	BornOnDate	DiedOnDate	DiedIn	FoundYear	hasCapital	locatedIn	bornIn
y:Abraham_Lincoln		Abraham Lincoln	1809–02–12	1865–04–15	y:Washington_D.C				
y:Washington_D.C	y:city	Washington D.C.				1790			
y:United_States	y:country	United States					y:Washington_D.C		
y:Reese_Witherspoon	y:actor	Reese Witherspoon	1976–03–22						y:New_Orleans,_Louisiana
y:New_Orleans,_Louisiana	y:city					1718		y:United_States	

图5　水平存储

```
SELECT hasName from T WHERE
BornOnDate = “1809-02-12” and
DiedOnDate = “1865-04-15”.
```

图6　水平存储上的SQL查询

组织结构图谱—知识库存储

- 基于关系数据库的存储和查询

3. 属性表

Prefix: y= http://en.wikipedia.org/wiki/

People

Subject	hasName	BornOnDate	DiedOnDate	DiedIn	BornIn	rdf:type
y:Abraham_Lincoln	"Abraham Lincoln"	1809–02–12	1865–04–15	y:Washington_D.C		
y:Reese_Witherspoon	"ReeseWitherspoon"	1976–03–22		y:Washington_D.C	y:New_Orleans,_ Louisiana	y:Actor

City

Subject	FoundYear	rdf:type	locatedIn	hasName
y:New_Orleans,_Louisiana	1718	y:city	y:United_States	
y:Washington_D.C	1790	y:city	y:United_States	"Washington D.C."

Country

Subject	hasName	hasCapital	rdf:type
y:United_States	"United States"	y:Washington_D.C	Country

图7 聚类属性表

组织结构图谱—知识库存储

- 基于关系数据库的存储和查询

4. 二元存储

Prefix: y= http://en.wikipedia.org/wiki/

hasName

Subject	Object
y:Abraham_Lincoln	"Abraham Lincoln"
y:Washington_D.C	"Washington D.C."
y:Reese_Witherspoon	"ReeseWitherspoon"
y:United_States	"United States"

FoundYear

Subject	Object
y:Washington_D.C	1790
y:New_Orleans,_Louisiana	1718

rdf:type

Subject	Object
y:Washington_D.C	y:city
y:United_States	Country
y:Reese_Witherspoon	y:Actor
y:New_Orleans,_Louisiana	y:city

BornOnDate

Subject	Object
y:Abraham_Lincoln	"1809-02-12"
y:Reese_Witherspoon	"1976-03-22"

DiedOnDate

Subject	Object
y:Abraham_Lincoln	"1865-04-15"

BornIn

Subject	Object
y:Reese_Witherspoon	y:New_Orleans,_Louisiana

DiedIn

Subject	Object
y:Abraham_Lincoln	y:Washington_D.C

LocatedIn

Subject	Object
y:New_Orleans,_Louisiana	y:United_States

图8 二元垂直分割表

组织结构图谱—知识库存储

- 基于图数据库的SPARQL的查询

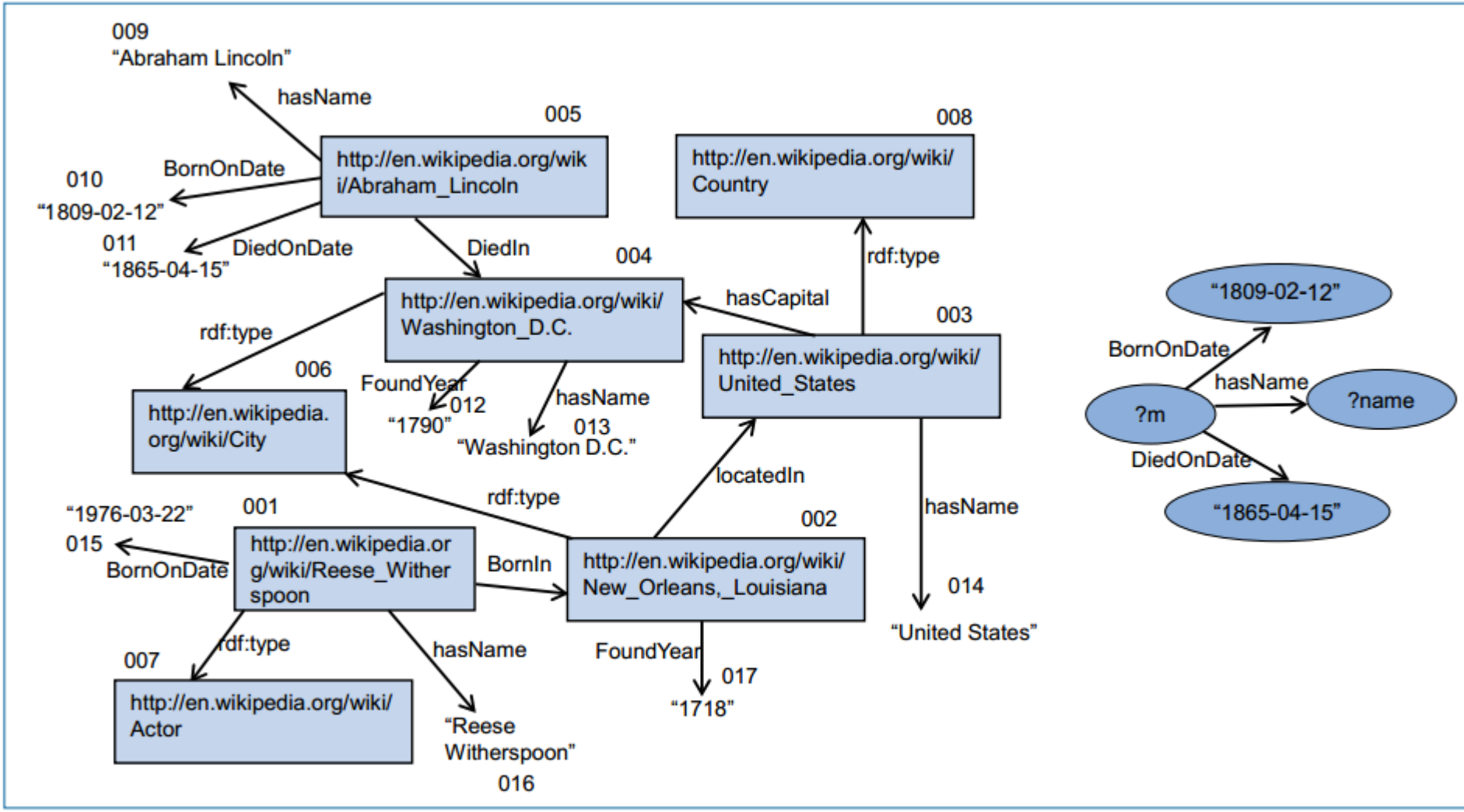


图3　RDF图和SPARQL查询图

组织结构图谱—工具

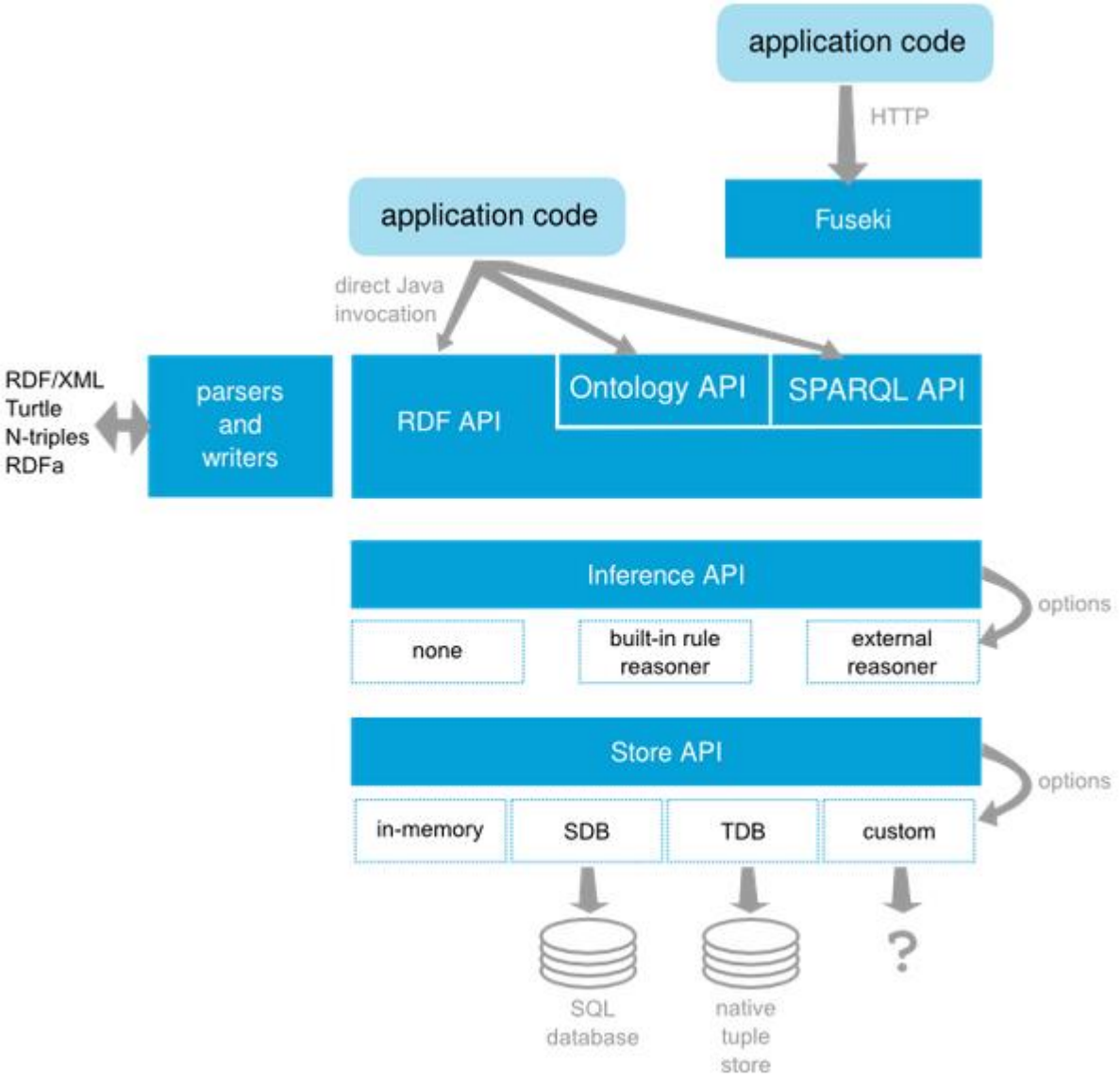
- Jena

1. 支持RDF创建及存储

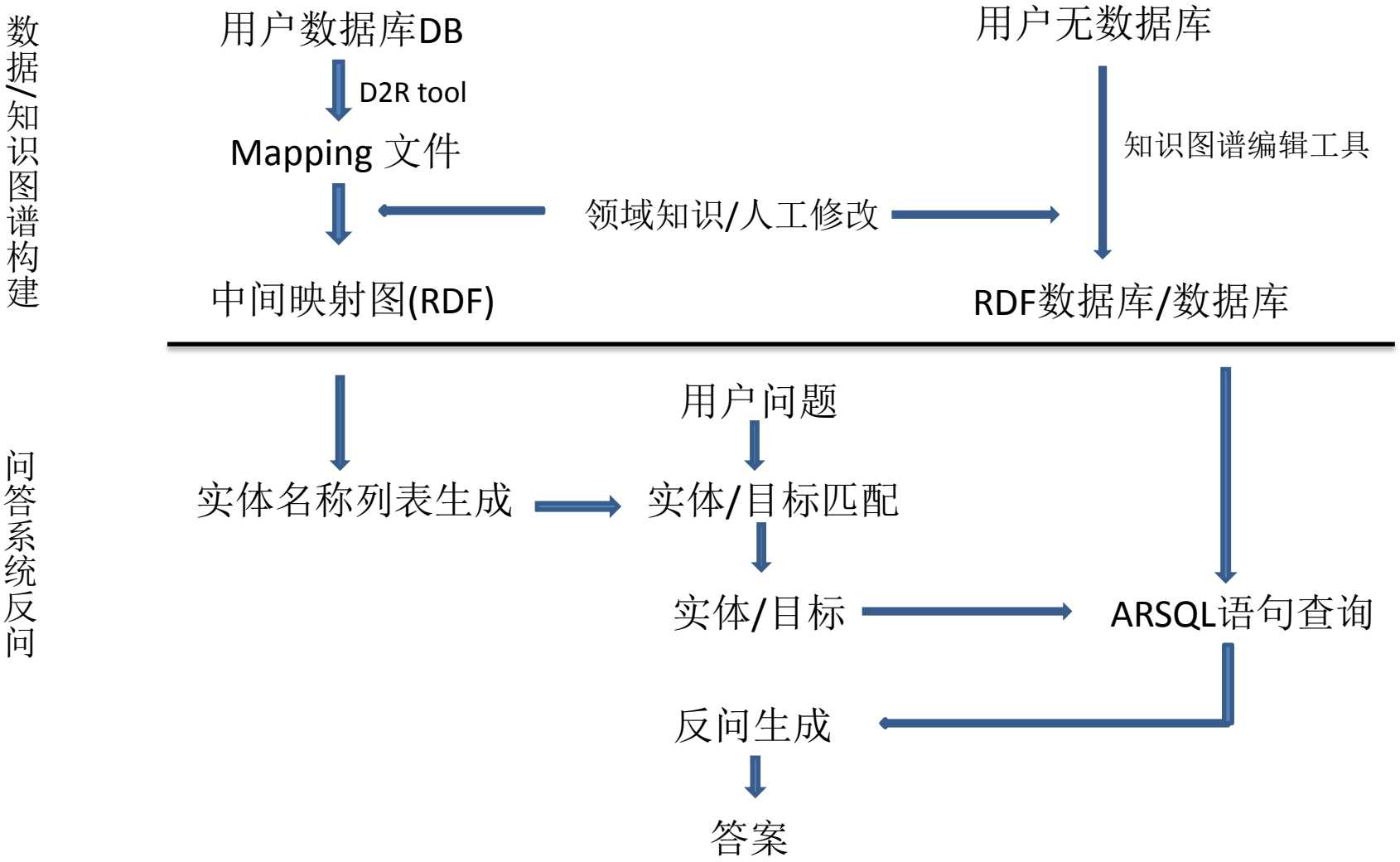
2. 支持SPARQL查询

3. 支持HBASE分布式存储

- D2R

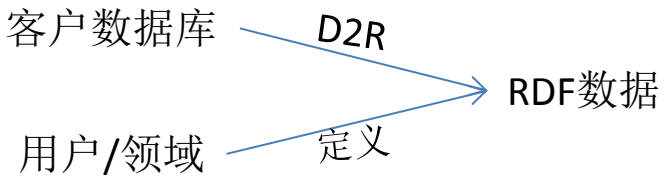


具体流程


数据/知识图谱构建
用户数据库DB
D2R tool
Mapping 文件
中间映射图(RDF)
领域知识/人工修改
用户无数据库
知识图谱编辑工具
RDF数据库/数据库
问答系统反问
实体名称列表生成
用户问题
实体/目标匹配
实体/目标
ARSQL语句查询
反问生成
答案

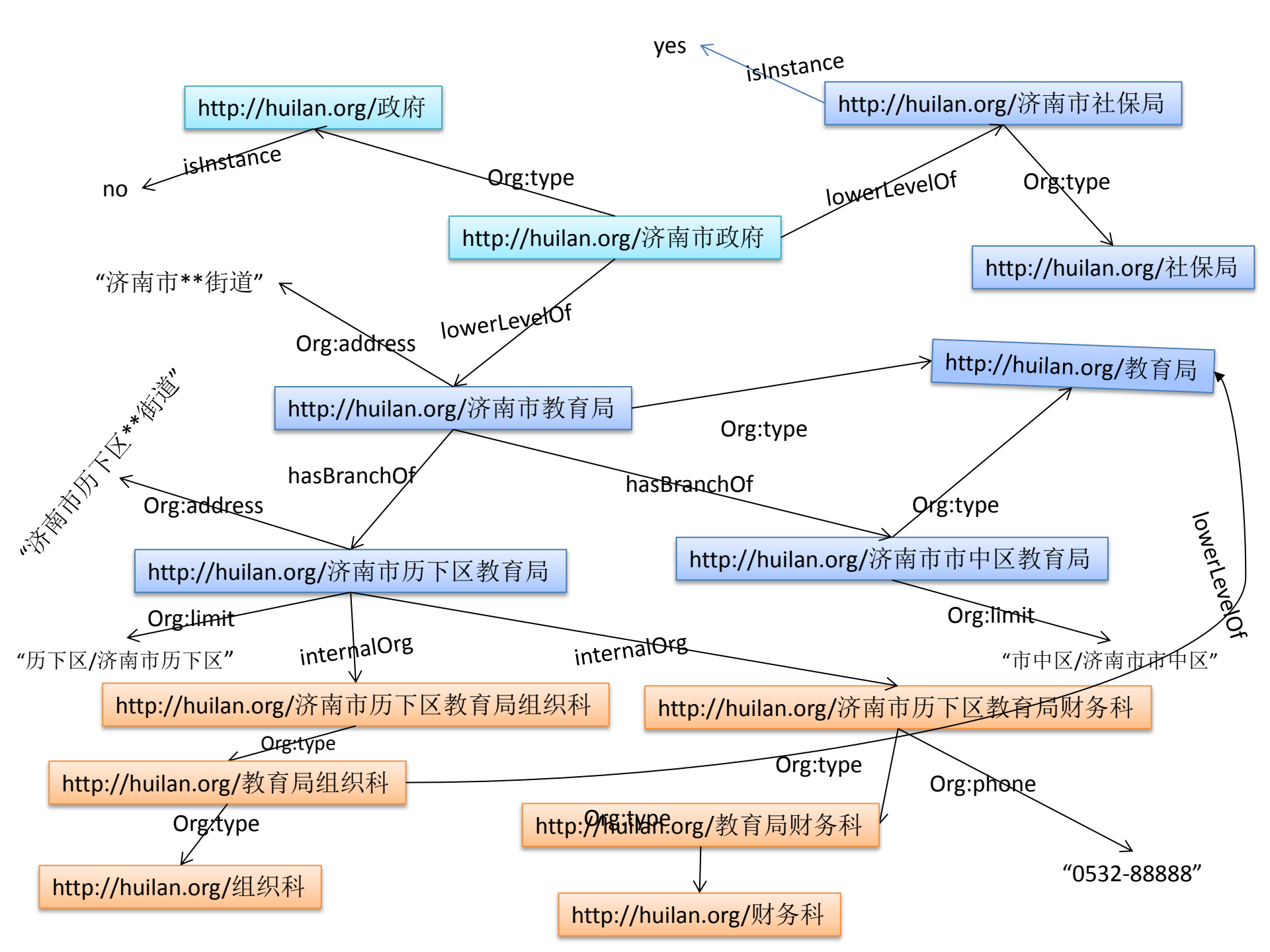
具体流程

- 数据准备



自上而下定义：

1. 类型： 表示类别摘要概念，如政府，部，处，厅，局，科, 室
2. 实体： 表示类别的具体实例。如济南市政府，山东省科技厅…
3. 属性：类别的一些属性。如电话，地址….
4. 关系：实体之间的关系，如<济南市政府,下级，济南市教育局>.


yes
isInstance
http://huilan.org/济南市社保局
http://huilan.org/政府
isInstance
no
Org:type
lowerLevelOf
Org:type
http://huilan.org/济南市政府
http://huilan.org/社保局
“济南市**街道”
lowerLevelOf
Org:address
http://huilan.org/教育局
http://huilan.org/济南市教育局
Org:type
“济南市历下区**街道”
hasBranchOf
hasBranchOf
Org:address
Org:type
lowerLevelOf
http://huilan.org/济南市历下区教育局
http://huilan.org/济南市市中区教育局
Org:limit
Org:limit
“历下区/济南市历下区”
internalOrg
internalOrg
“市中区/济南市市中区”
http://huilan.org/济南市历下区教育局组织科
http://huilan.org/济南市历下区教育局财务科
Org:type
Org:type
Org:phone
http://huilan.org/教育局组织科
Org:type
Org:type
http://huilan.org/教育局财务科
“0532-88888”
http://huilan.org/组织科
http://huilan.org/财务科

具体流程

- 数据准备RDF
- 实体名称列表生成

1. 将知识图谱中的实例实体名和别名作为实体名称列表，如济南市政府，济南市教育局
2. 将知识图谱中的非实例实体名作为虚实体名列表，如教育局，组织科
3. 将知识图谱中的限制名作为限制名列表，如济南市，历下区
4. 属性作为正则表达式中的target. 如电话，地址…

具体流程

- 数据准备RDF
- 实体名称列表生成
- 反问

市中区社保局组织科电话是多少？

市中区社保局组织科电话是多少？

目标： 电话

实例实体1: 市中区社保局组织科

搜索领域图谱 

```
SELECT ? Phone
WHERE{
   市中区社保局 Org:phone ?phone
}
```

<电话,市中区社保局组织科>

社保局组织科电话是多少？

社保局组织科电话是多少？

目标： 电话

虚实体1: 社保局组织科

搜索领域图谱

反问：请问是哪个社保局组织科

限制词：市中区

<I电话,市中区社保局组织科>

```
SELECT ?x
WHERE{
    ?x Org:type社保局组织科.
}
```

济南市历下区社保局组织科
济南市社保局组织科
济南市市中区社保局组织科

Type:社保局组织科

```
SELECT ?x
WHERE{
    ?x Org:type 社保局组织科组织科.
    ?x  Org:limit 市中区.
}
```

组织科电话是多少？

组织科电话是多少？

目标： 电话
虚实体1: 组织科

搜索领域图谱

反问：请问是哪个组织科

市中区教育局
限制词：市中区
虚实体：教育局

<I电话,市中区社保局组织科>

```
SELECT ?x
WHERE{
   ?x Org:type 组织科
}
```

社保局组织科
教育局组织科

Type:组织科

```
SELECT ?x
WHERE{
   ?x Org:type 教育局组织科
   ?x Org:limit 市中区
}
```

济南市市中区教育局组织科