基于图的RDF数据管理

报告人：邹磊
zoulei@pku.edu.cn

提纲

1. 知识图谱概述
2. 从不同角度和学科研究
3. 从数据管理层面的讨论
4. 一些开放性问题
5. 系统应用
6. 总结

提纲

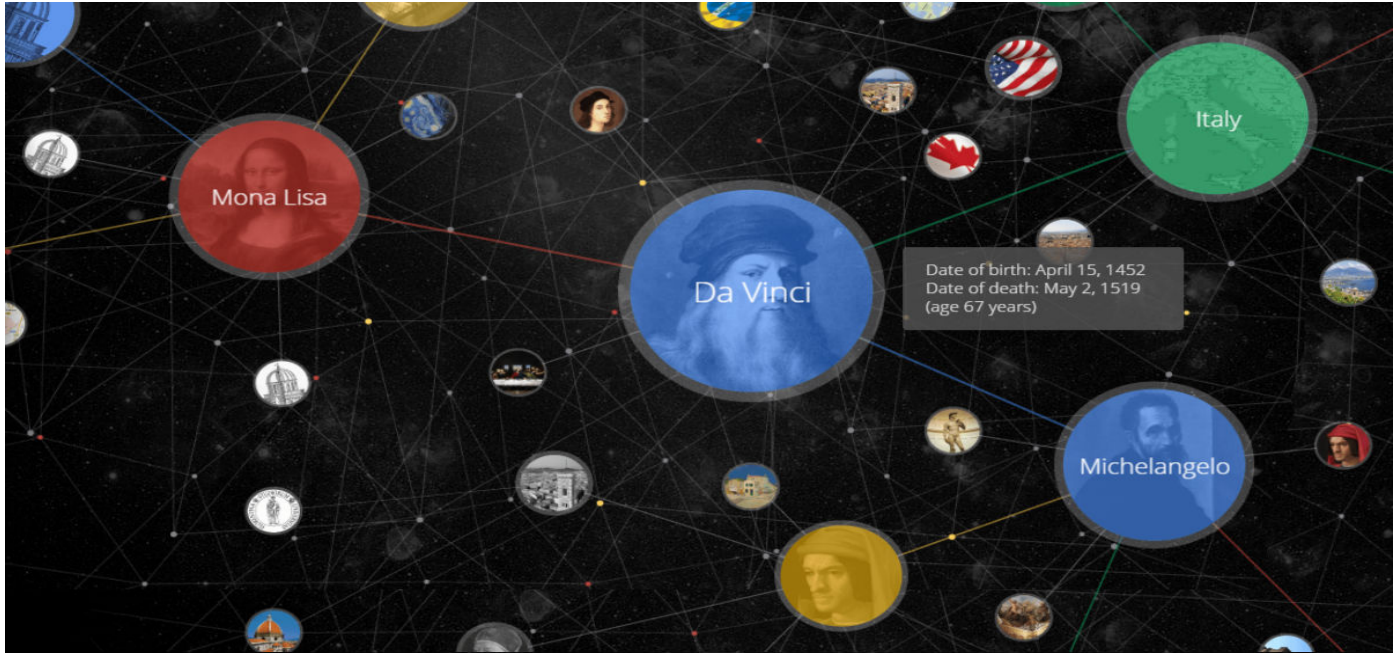
1. 知识图谱概述
2. 从不同角度和学科研究
3. 从数据管理层面的讨论
4. 一些开放性问题
5. 系统应用
6. 总结

知识图谱 (Knowledge Graph)

2012年5月16日, Google发布“知识图谱”的新一代“智能”搜索功能。

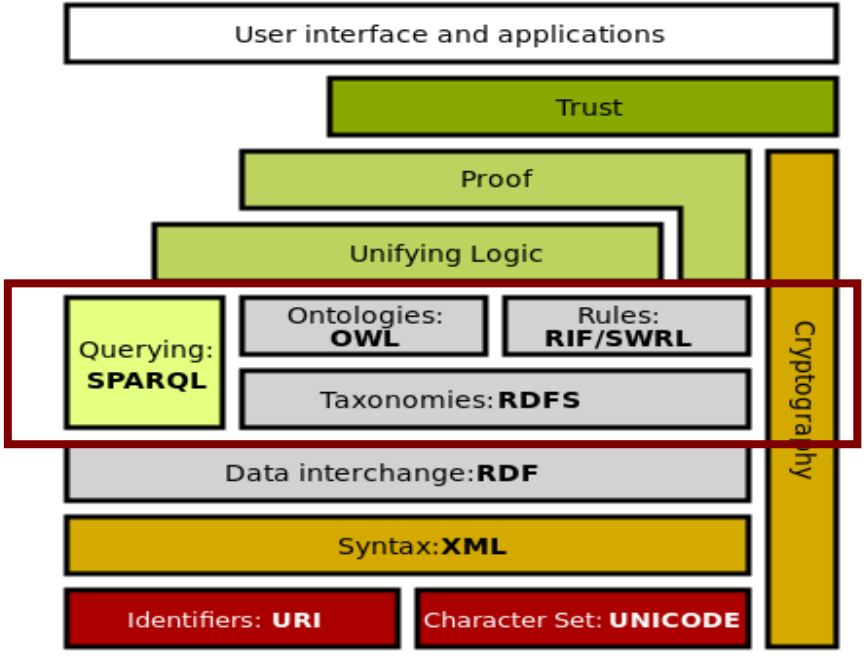
知识图谱 (Knowledge Graph)

本质上是基于图的语义网络，表示实体和实体之间的关系！



资源描述框架(RDF)数据

- RDF是知识图谱数据的事实标准
- RDF是由W3C组织提出的一种描述资源概念模型的语言
- RDF是语义网的一个基石（Building Block）
- 语义网的目标是网络上的资源是“机器可理解”(Machine understandable)



工业应用


THE INDUSTRY


Google发布
知识图谱
搜狗发布
“知立方”
Facebook
社交知识图谱的
图搜索功能上线
百度发布
知识图谱
2012-05-16
2012-11-12
2013-01-16
2013-02-08

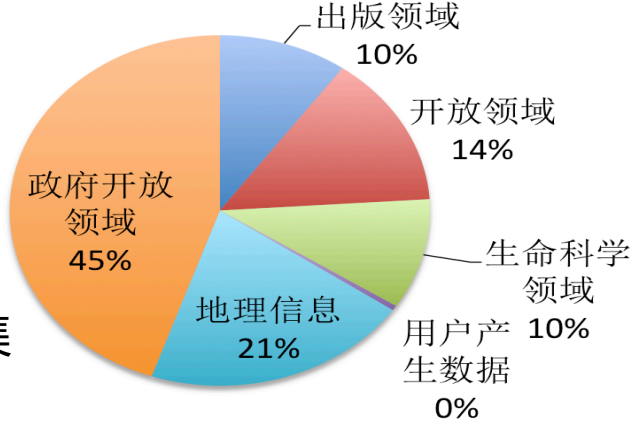
Google发布
知识图谱
2012-05-16

搜狗发布
“知立方”
2012-11-12

Facebook
社交知识图谱的
图搜索功能上线
2013-01-16

百度发布
知识图谱
2013-02-08

RDF数据集
类别统计



BioModels, Biosamples, ChEMBL,
Ensembl, Atlas, Reactome and UniProt

语义网？------一个简单例子(RDFa)

传统的HTML只是考虑网页的显示，例如字体、段落格式等；而不是网页中的内容的语义。

```
<html>
    <font size="3" color="red"> Lei Zou </font>
    </br>
    Email:<a href= "mailto: zoulei@pku.edu.cn">
zoulei@pku.edu.cn  </a>
    <p>
       <font size="3" color="black">Publications: </font>
    </p>
    <div>
        Lei Zou, Jinhui Mo, Lei Chen, M. Tamer Ozsu,
Dongyan Zhao, gStore: Answering SPARQL Queries Via
Subgraph Matching, VLDB, 2011
    </div>
</html>
```

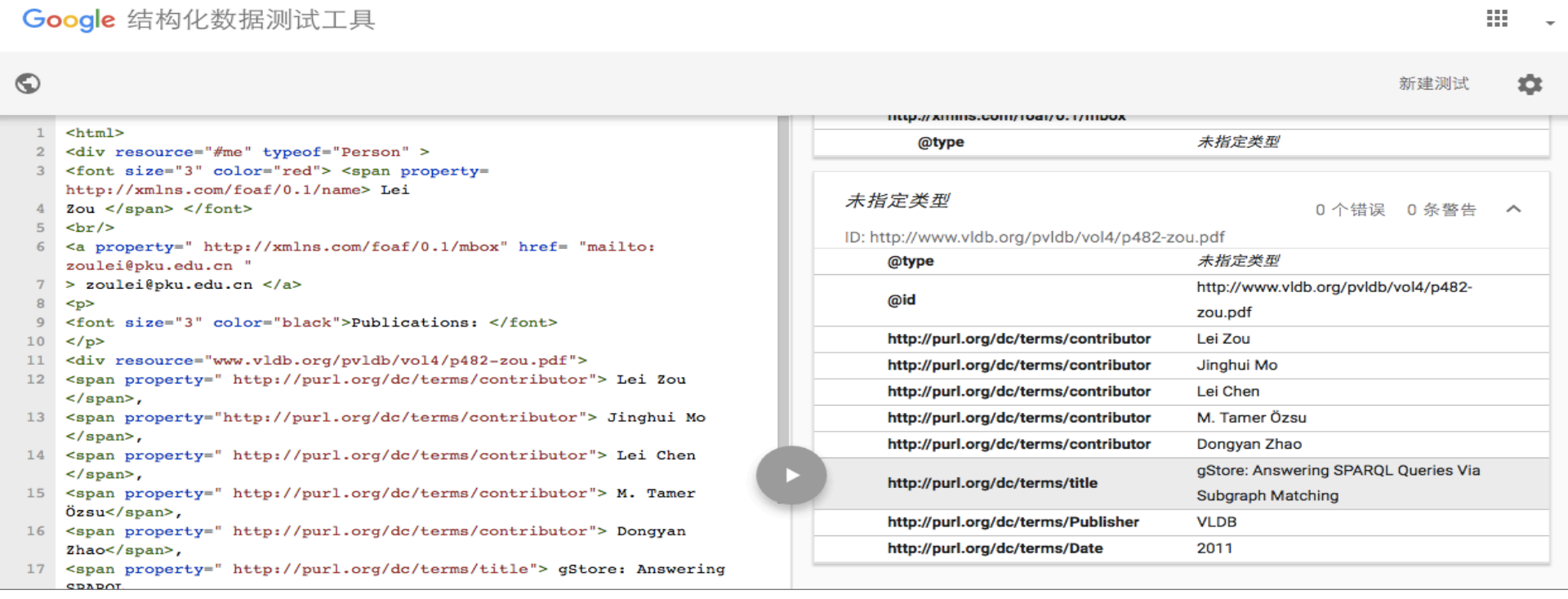
语义网？------一个简单例子(RDFa)

语义网考虑的是内容的语义。

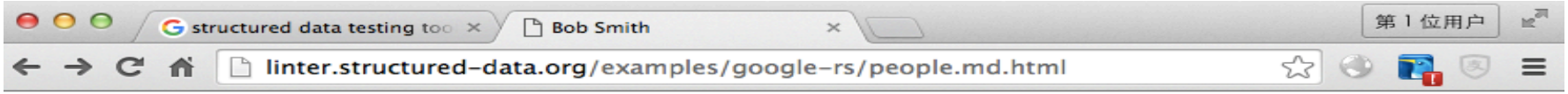
```
<html>
<div resource="#me" typeof="Person" >
<font size="3" color="red"> <span property= http://xmlns.com/foaf/0.1/name> Lei
Zou </span> </font>
<br/>
<a property=" http://xmlns.com/foaf/0.1/mbox" href= "mailto: zoulei@pku.edu.cn "
> zoulei@pku.edu.cn </a>
<p>
<font size="3" color="black">Publications: </font>
</p>
<div resource="www.vldb.org/pvldb/vol4/p482-zou.pdf">
    <span property=" http://purl.org/dc/terms/contributor"> Lei Zou </span>,
    <span property="http://purl.org/dc/terms/contributor"> Jinghui Mo </span>,
    <span property=" http://purl.org/dc/terms/contributor"> Lei Chen </span>,
    <span property=" http://purl.org/dc/terms/contributor"> M. Tamer Özsu</span>,
    <span property=" http://purl.org/dc/terms/contributor"> Dongyan Zhao</span>,
    <span property=" http://purl.org/dc/terms/title"> gStore: Answering SPARQL
Queries Via Subgraph Matching </span>,
    <span property=" http://purl.org/dc/terms/Publisher"> VLDB </span>
    <span property=" http://purl.org/dc/terms/Date">2011</span>
</div>
</html>
```

语义网？------一个简单例子(RDFa)

Google结构化数据测试工具



语义网？------一个简单例子(RDFa)



My name is Bob Smith, but people call me Smithy. Here is my homepage: www.example.com. I live in Albuquerque, NM and work as an engineer at ACME Corp. My friends: Darryl, Edna

RDFa (source lint)

```
<div xmlns:v="http://rdf.data-vocabulary.org/#" typeof="v:Person">
   My name is <span property="v:name">Bob Smith</span>,
   but people call me <span property="v:nickname">Smithy</span>.
   Here is my homepage:
   <a href="http://www.example.com" rel="v:url">www.example.com</a>.
   I live in
   <span rel="v:address">
      <span typeof="v:Address">
         <span property="v:locality">Albuquerque</span>,
         <span property="v:region">NM</span>
      </span>
   </span>
   and work as an <span property="v:title">engineer</span>
   at <span property="v:affiliation">ACME Corp</span>.
   My friends:
   <a href="http://darryl-blog.example.com" rel="v:friend">Darryl</a>,
   <a href="http://edna-blog.example.com" rel="v:friend">Edna</a>
</div>
```

语义网？------一个简单例子(RDFa)

Enhanced search result preview

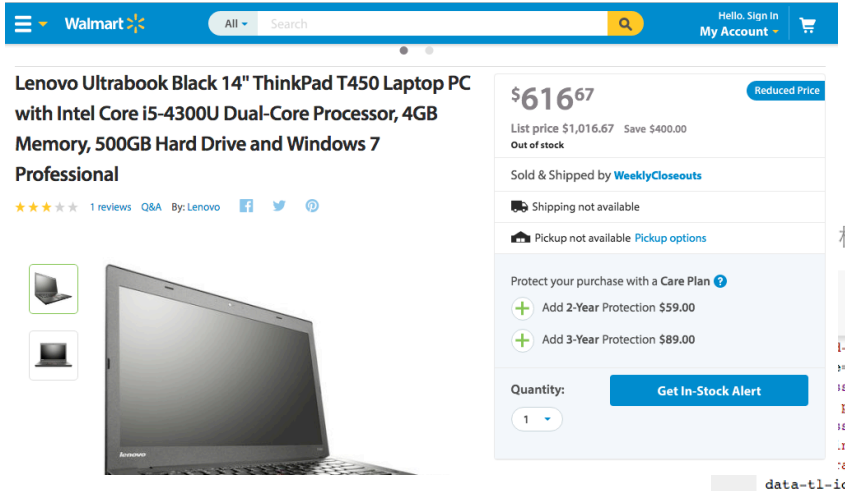
*Disclaimer: this preview is only shown as a example of what a search engine **might** display. It is to the discretion of each search engine provider to decide whether your page will be displayed as an enhanced search result or not in their search results pages.*

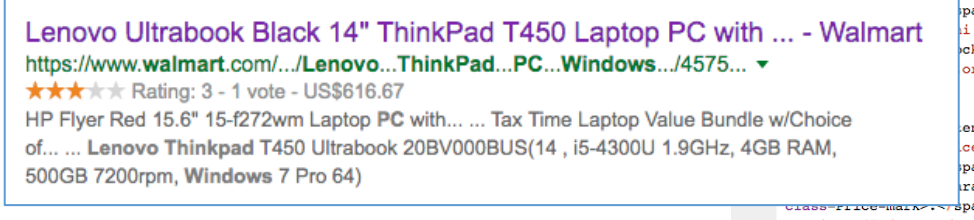
Bob Smith
linter.structured-data.org/examples/google-rs/people.rdfa.html
Albuquerque, NM - engineer, ACME Corp
an **actual** search result **may** display other content **relating** to your search terms here.

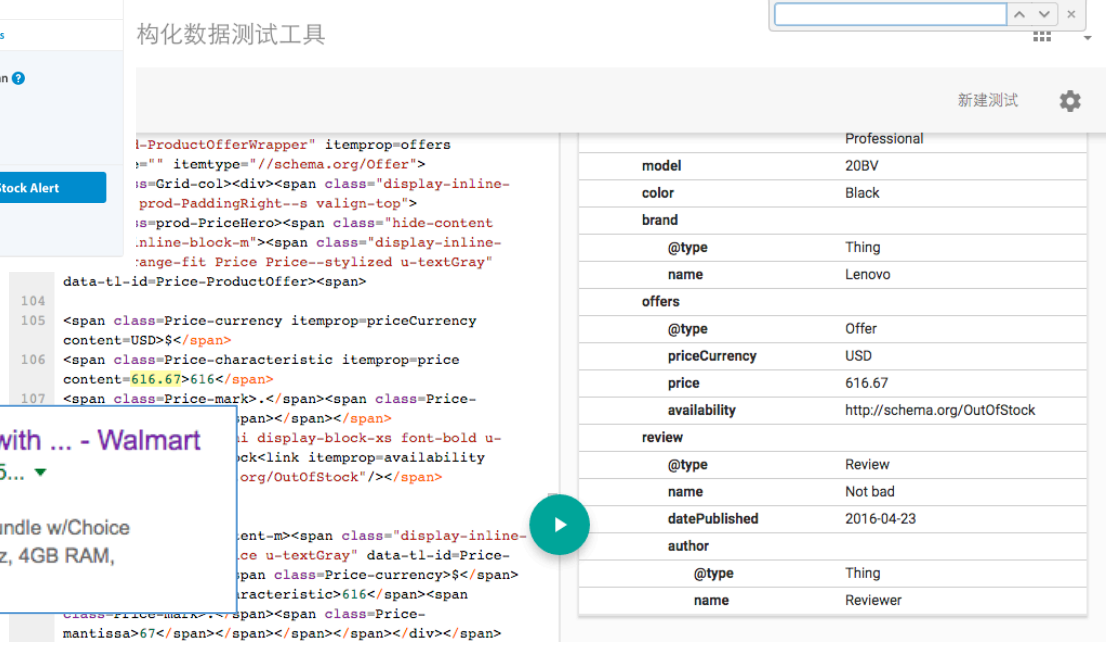
Raw structured data extracted from the page:

@id	http://rdf.data-vocabulary.org/#Address(1)
rdf:type	rdfs:Class
@id	http://rdf.data-vocabulary.org/#Person(1)
rdf:type	rdfs:Class
rdf:type	v:Person rdfs:Resource
v:address	**rdf:type** v:Address rdfs:Resource **v:locality** Albuquerque **v:region** NM
v:affiliation	ACME Corp
v:friend	• http://edna-blog.example.com • http://darryl-blog.example.com
v:name	Bob Smith
v:nickname	Smithy

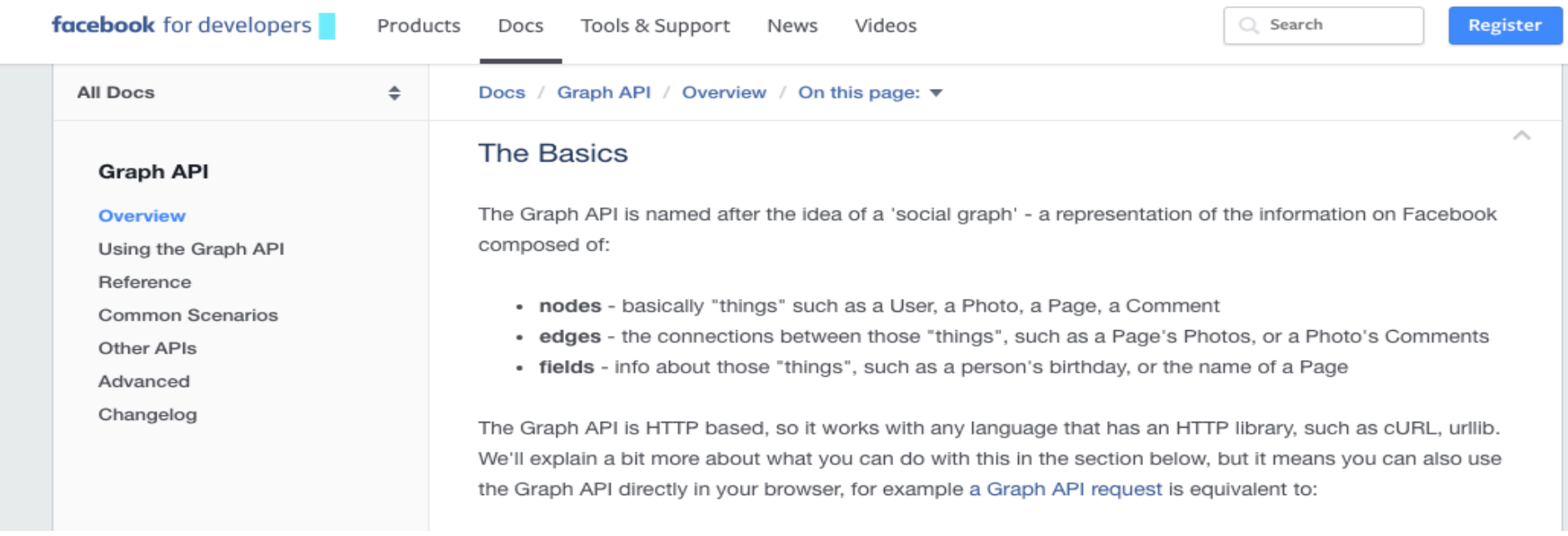
What is Semantic Web ? ---A Simple Example







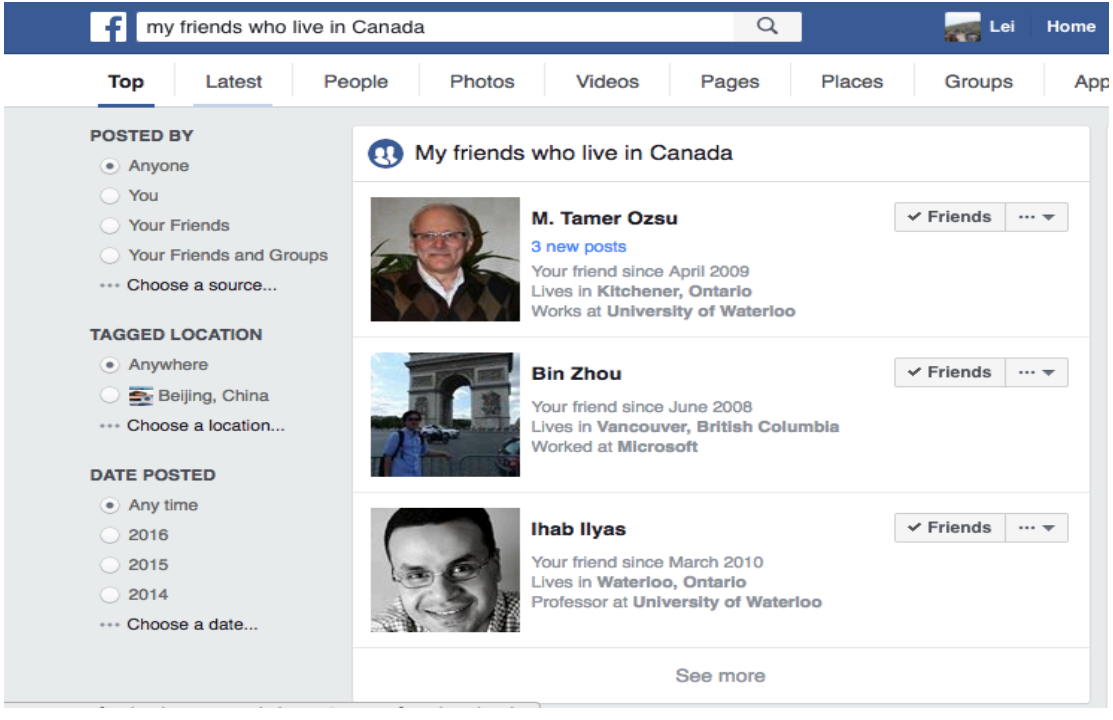
Facebook Social Graph

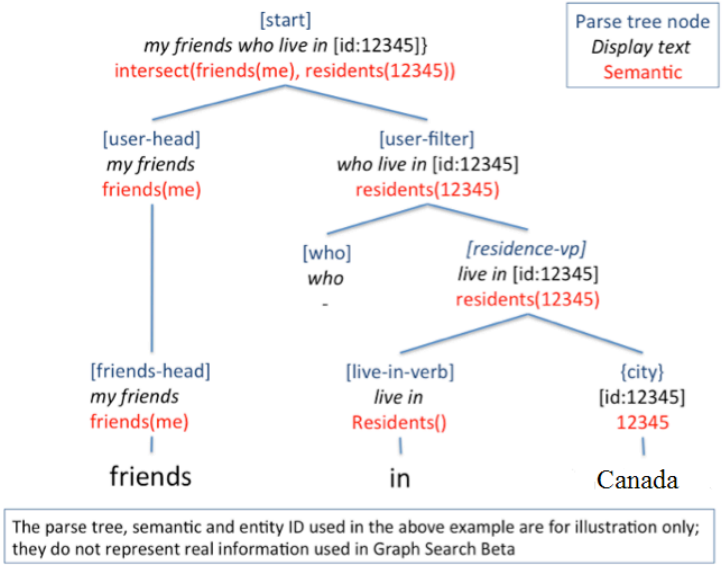


Facebook Graph Search

2013年1月16日 Facebook Graph Search 产品发布会---Mark Zuckerberg

“My friends who live in Canada”

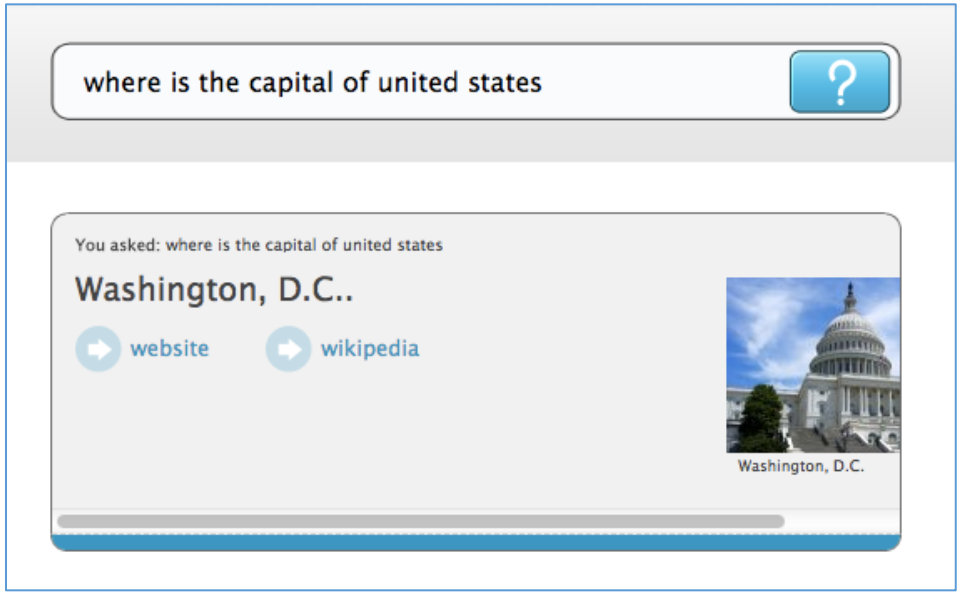




Facebook Graph Search

"Photos of my friends who live in Canada"

EVI---原名True Knowledge

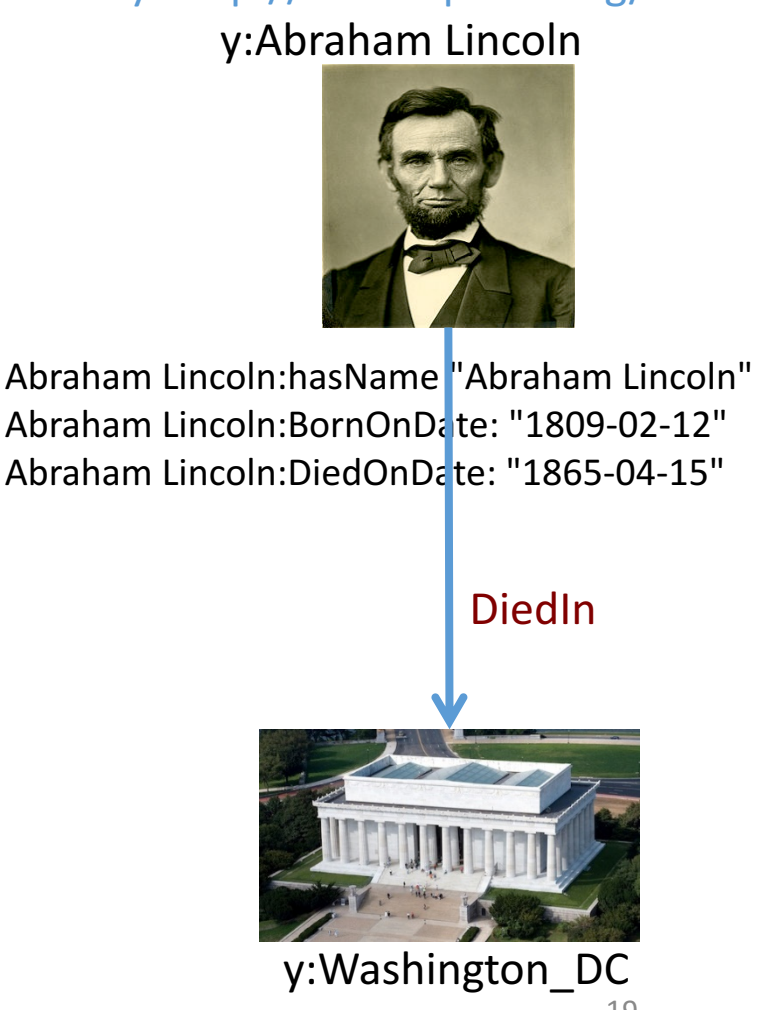


年度	获得风投
2007-09	120万 USD
2008-07	400万 USD
2012-01	被亚马逊收购

William Tunstall-Pedoe: *True Knowledge: Open-Domain Question Answering using Structured Knowledge and Inference*. AI Magazine 31(3): 80-92 (2010)

RDF 数据模型

- RDF中任何实体都被称之为资源(Resource)，用URI来表示。
- 实体的属性需要被定义
- 实体间关系需要被定义
- 不同数据集直接互相链接构成海量的关联数据
  - 一个集成的Web”数据库”



RDF 数据 & SPARQL查询语言

RDF 数据库

主语	谓词	宾语
Abraham_Lincoln	hasName	“Abraham Lincoln"
Abraham_Lincoln	BornOnDate	“1809-02-12"
Abraham_Lincoln	DiedOnDate	“1865-04-15"”
Abraham_Lincoln	DiedIn	Washington_DC
Abraham_Lincoln	bornIn	Hodgenville KY
Reese_Witherspoon	bornOnDate	"1976-03-22"
Reese_Witherspoon	bornIn	New_Orleans_LA
New_Orleans_LA	foundingYear	“1718”
New Orleans LA	locatedIn	United_States
United_States	hasName	“United States ”
United_States	hasCapital	Washington_DC
United_States	foundingYear	“1776”

“找到出生1976年生的，并且出生地是1718年构建的城市的人有哪些？”

SPARQL

```
SELECT ?name
WHERE {
?m <bornIn> ? c i t y .
?m <hasName> ?name .
?m <bornOnDate> ?bd .
? c i t y <foundingYear> ``1718 '' .
FILTER( regex (str (?bd ), “1 9 7 6 '' ) )
}
```

提纲

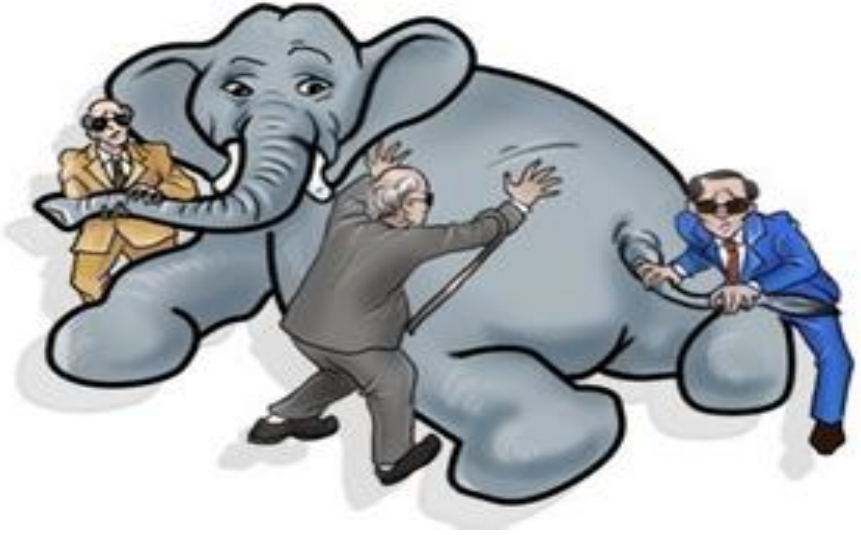
1. 知识图谱概述
2. 从不同角度和学科研究
3. 从数据管理层面的讨论
4. 一些开放性问题
5. 系统应用
6. 总结

交叉研究

数据库

RDF数据库系统

数据集成、知识融合



自然语言处理

关系抽取

语义解析

(Semantic Parsing)

机器学习

知识图谱数据的知识表示

(Graph Embedding)

知识工程

知识库构建

基于规则的推理

知识工程

知识库构建 [Mendes et al. 12; Suchanek et al. 07; Bollacker]

Leipzig University
University of Mannheim
OpenLink Software

11亿三元组

Max-Planck-Institute

1.8亿三元组

Metaweb Company
2010年被Google收购

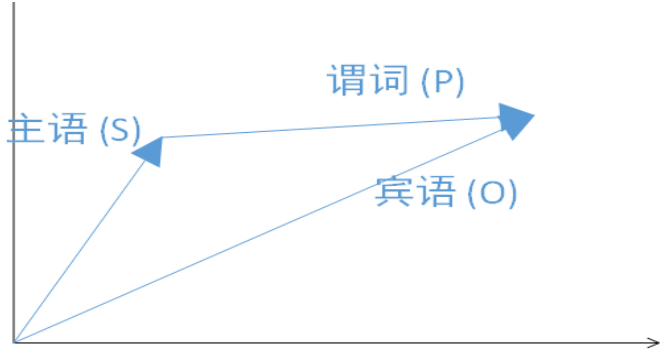
25.3亿三元组

机器学习

知识表示代表模型：TransE [Bordes et al., NIPS 13]

- 对每个事实(Subject, Predicate, Object)，将其中的predicate作为从subject到object的翻译操作
- 每个Subject/Predicate/Object，都映射成一个高纬向量
- 优化目标: S+P=O

$$\Gamma=\sum_{(s,p,o)\in s}\sum_{(s',p',o')\notin s}[r+d(s+p,o)-d(s'+p',o')]_+$$



S	P	O
China	Capital	Beijing
Canada	Capital	Ottawa
……	……	……

$$\text{Beijing} - \text{China} \approx \text{Ottawa} - \text{Canada} = \text{Capital}$$

自然语言处理

语义解析Semantic Parsing [Zettlemoyer et al., UAI 05]

语义解析就是将自然语言映射成机器可以表达的形式。

E.g., “Which states borders New Mexico ?”

Lambda表达式[Alonzo Church, 1940]

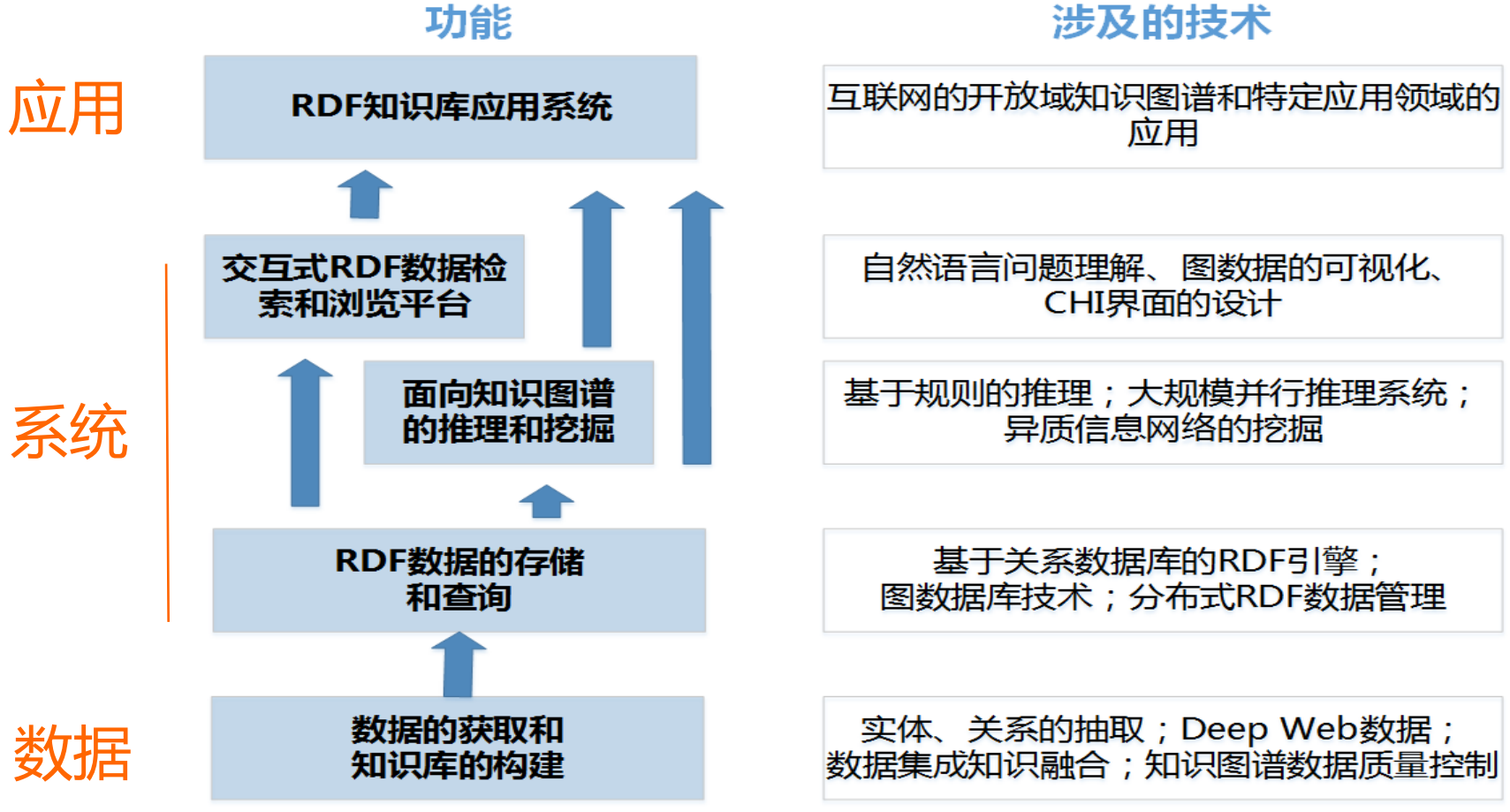
$$\lambda x.state(x) \wedge borders(x, new\_mexico)$$

“**Simply typed-calculus** can express varies database query languages such as **relational algebra**, fixpoint logic and the complex object algebra." [Hillebrand et al., 1996]

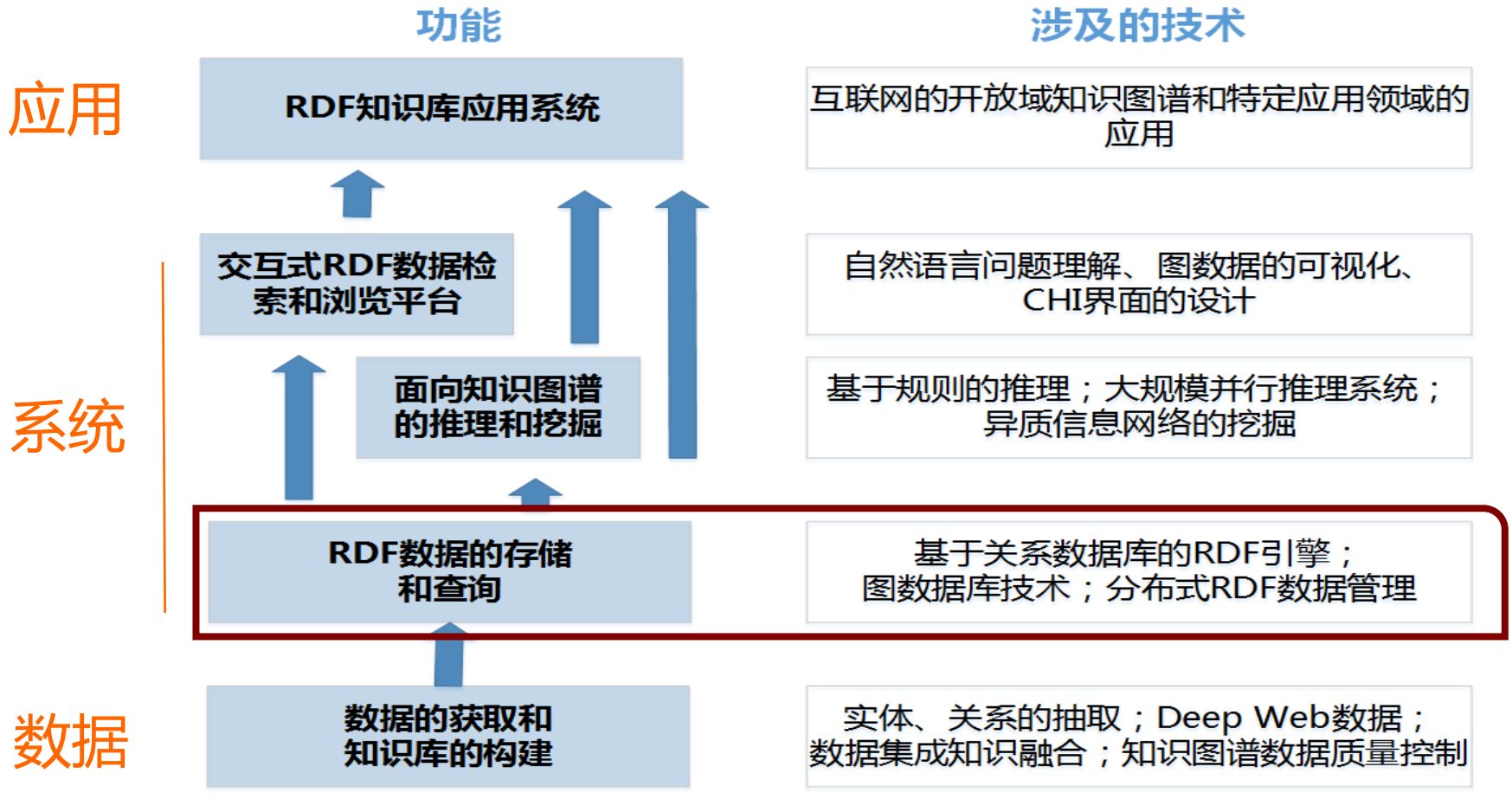
提纲

1. 知识图谱概述
2. 从不同角度和学科研究
3. 从数据管理层面的讨论
4. 一些开放性问题
5. 系统应用
6. 总结

RDF数据管理问题


功能
涉及的技术
应用
系统
数据
RDF知识库应用系统
交互式RDF数据检索和浏览平台
面向知识图谱的推理和挖掘
RDF数据的存储和查询
数据的获取和知识库的构建
互联网的开放域知识图谱和特定应用领域的应用
自然语言问题理解、图数据的可视化、CHI界面的设计
基于规则的推理；大规模并行推理系统；异质信息网络的挖掘
基于关系数据库的RDF引擎；图数据库技术；分布式RDF数据管理
实体、关系的抽取；Deep Web数据；数据集成知识融合；知识图谱数据质量控制

RDF数据管理问题


功能
涉及的技术
应用
系统
数据
RDF知识库应用系统
互联网的开放域知识图谱和特定应用领域的应用
交互式RDF数据检索和浏览平台
自然语言问题理解、图数据的可视化、CHI界面的设计
面向知识图谱的推理和挖掘
基于规则的推理；大规模并行推理系统；异质信息网络的挖掘
RDF数据的存储和查询
基于关系数据库的RDF引擎；图数据库技术；分布式RDF数据管理
数据的获取和知识库的构建
实体、关系的抽取；Deep Web数据；数据集成知识融合；知识图谱数据质量控制

一个基本的问题：如何存储RDF数据和回答SPARQL查询

主语	谓词	宾语
Abraham_Lincoln	hasName	“Abraham Lincoln"
Abraham_Lincoln	BornOnDate	“1809-02-12"
Abraham_Lincoln	DiedOnDate	“1865-04-15"”
Abraham_Lincoln	DiedIn	Washington_DC
Abraham_Lincoln	bornIn	Hodgenville KY
Reese_Witherspoon	bornOnDate	"1976-03-22"
Reese_Witherspoon	bornIn	New_Orleans_LA
New_Orleans_LA	foundingYear	“1718”
New Orleans LA	locatedIn	United_States
United_States	hasName	“United States”
United_States	hasCapital	Washington_DC
United_States	foundingYear	“1776”

SPARQL

```
SELECT ?name
WHERE {
?m <bornIn> ? c i t y .
?m <hasName> ?name .
?m <bornOnDate> ?bd .
? c i t y <foundingYear> ` `1718 ' ' .
FILTER( regex (str (?bd ), “1 9 7 6 ' ' ) )
}
```

怎样快速回答SPARQL？

现有方法：求助于关系数据库技术

主语	谓词	宾语
Abraham_Lincoln	hasName	“Abraham Lincoln"
Abraham_Lincoln	BornOnDate	“1809-02-12"
Abraham_Lincoln	DiedOnDate	“1865-04-15"”
Abraham_Lincoln	DiedIn	Washington_DC
Abraham_Lincoln	bornIn	Hodgenville KY
Reese_Witherspoon	bornOnDate	"1976-03-22"
Reese_Witherspoon	bornIn	New_Orleans_LA
New_Orleans_LA	foundingYear	“1718”
New Orleans LA	locatedIn	United_States
United_States	hasName	“United States”
United_States	hasCapital	Washington_DC
United_States	foundingYear	“1776”

SPARQL

```
SELECT ?name
WHERE {
?m <bornIn> ? c i t y .
?m <hasName> ?name .
?m <bornOnDate> ?bd .
? c i t y <foundingYear> ``1718 ' ' .
FILTER( regex (str (?bd ), “1 9 7 6 ' ' ) )
}
```

SQL

```
SELECT T2 . o b j e c t
FROM T
T as T4
WHERE T1.prope        bornIn "
AND T2.proper         hasName"
AND T3.property= "bornOnDate "
AND T1.subject=T2.subject
AND T2.subject=T3.subject
AND T1.object=T4.subject
AND T4.propety=“foundingYear “
AND T4.object=" 1718 "
AND T3.object LIKE '%1976%'
```

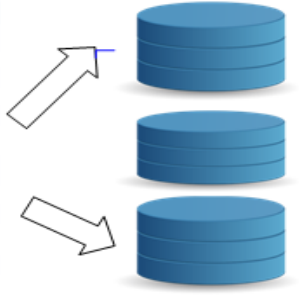
多步的自连接操作？

三种典型基于关系数据库的优化策略

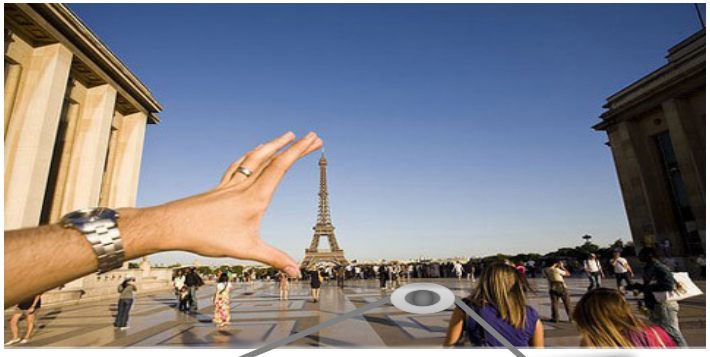
- **属性表方法** Jena [Wilkinson et al., 2003] ,FlexTable [Wang et al., 2010] , DB2-RDF [Bornea et al., 2013]
- **垂直划分方法** SW-store [Abadi et al., 2009]
- **全索引方法** RDF-3X [Neumann and Weikum, 2008], Hexastore [Weiss et al., 2008]

基本思路： 划分三元组表、生成更加简单的查询。

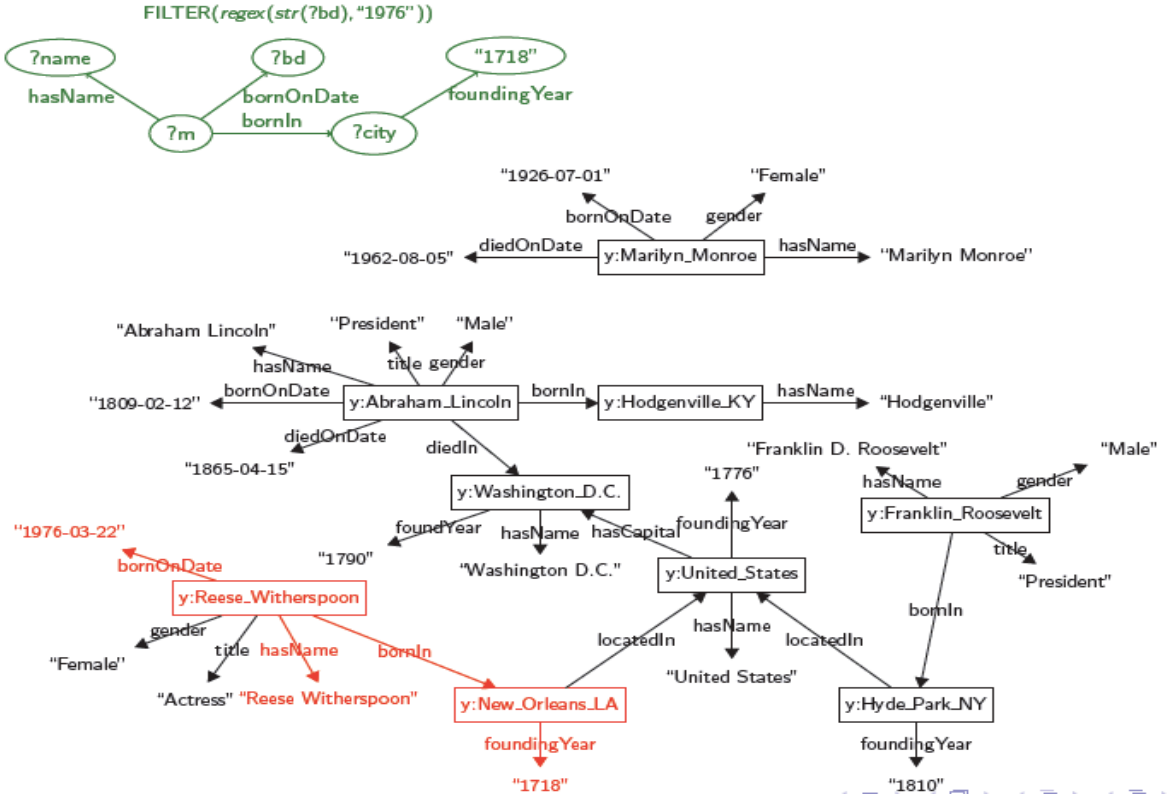
主语	谓词	宾语
Abraham_Lincoln	hasName	"Abraham Lincoln"
Abraham_Lincoln	BornOnDate	"1809-02-12"
Abraham_Lincoln	DiedOnDate	"1865-04-15"^^
Abraham_Lincoln	DiedIn	Washington_DC
Abraham_Lincoln	bornIn	Hodgenville KY
Reese_Witherspoon	bornOnDate	"1976-03-22"
Reese_Witherspoon	bornIn	New_Orleans_LA
New_Orleans_LA	foundingYear	"1718"
New Orleans LA	locatedIn	United_States
United_States	hasName	"United States"
United_States	hasCapital	Washington_DC
United_States	foundingYear	"1776"



- M. T. Özsu. "A Survey of RDF Data Management Systems", Front. Comp. Sci., 2016.
- Lei Zou, M. T. Özsu. “Graph-based RDF Data Management”, Data Science and Engineering, 2(1): 56-70 (2017)



我们的方法---gStore [Zou et al., VLDB 11; VLDB J 14]



回答SPARQL查询
== 子图匹配

我们基于图的RDF数据管理研究路线图

gAnswer: 图匹配驱动的RDF知识图谱自然语言**检索平台**

gStore-D: **分布式RDF**图数据管理系统

gStore：基于子图匹配的**SPARQL查询**系统

研究主题

子图模式匹配查询

建模

基于结构感知的图数据库索引和子图匹配查询**优化理论**

主语	谓词	宾语
Abraham_Lincoln	hasName	"Abraham Lincoln"
Abraham_Lincoln	BornOnDate	"1809-02-12"
Abraham_Lincoln	DiedOnDate	"1865-04-15"^^
Abraham_Lincoln	DiedIn	Washington_DC
Abraham_Lincoln	bornIn	Hodgenville KY
Reese_Witherspoon	bornOnDate	"1976-03-22"
Reese_Witherspoon	bornIn	New_Orleans_LA
New_Orleans_LA	foundingYear	"1718"
New Orleans LA	locatedIn	United_States
United_States	hasName	"United States"
United_States	hasCapital	Washington_DC
United_States	foundingYear	"1776"

RDF数据

SPARQL

```
SELECT ?name
WHERE {
?m <bornIn> ?city .
?m <hasName> ?name .
?m <bornOnDate> ?bd .
?city <foundingYear> ''1718'' .
FILTER( regex (str (?bd ), "1976" ) )
}
```

自然语言问题

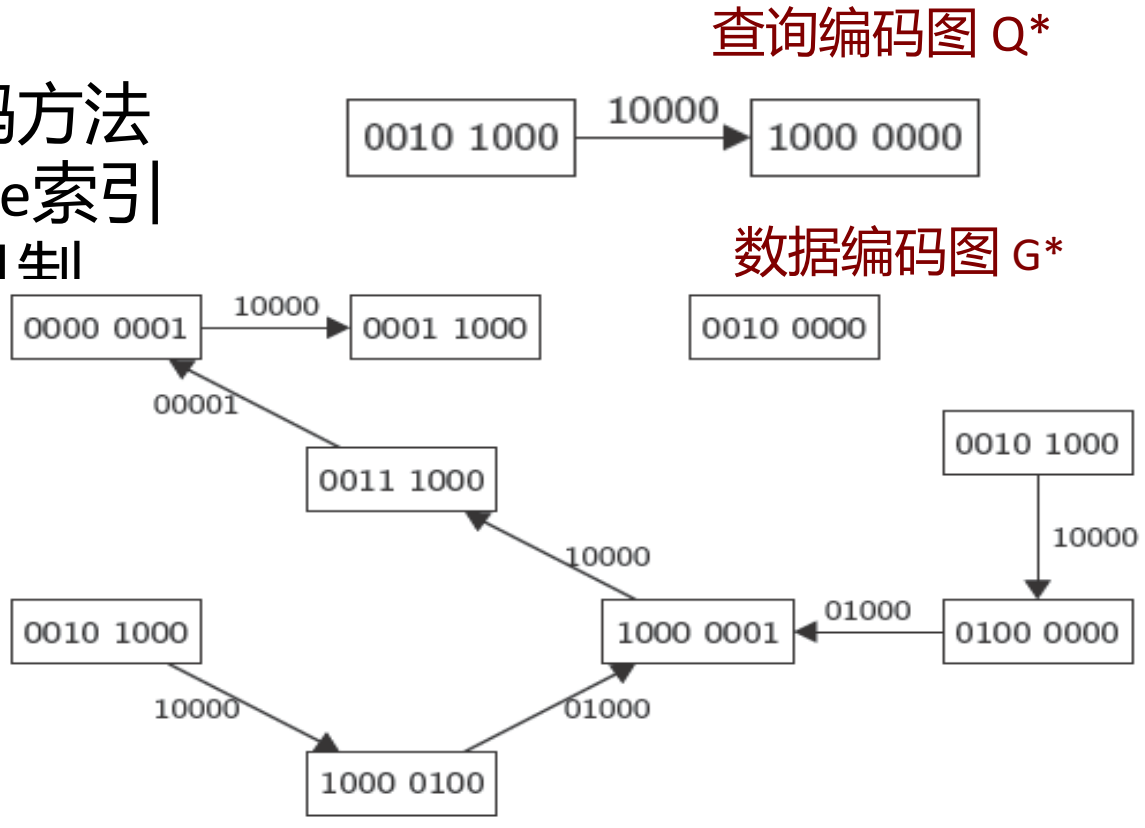
"杨绛爱人的著作有哪些？"

"Who was married to an actor that play in Philadelphia ?"

gStore---一种基于图的RDF存储和查询系统

主要技术手段：

- 统一的结构和内容编码方法
- 一种高度平衡树 VS-tree索引
- 基于索引的多级过滤机制



找到Q*在G*上的匹配 → 测试每个这样的匹配

构建的系统

gStore

代码：除了SPARQL语法解析器外均为独立开发的，目前有14万行C++，完成自主知识产权；共计6人年, 目前是版本v 0.3.0 .

开源地址：https://github.com/Caesar11/gStore/

包括全部的系统代码；详细的用户手册；与目前最好的开源和工业系统在多个Benchmark数据集上的对比测试报告；系统使用演示视频。

开源协议：基于BSD 协议

部署方法：单机和C/S方式部署

接口：C++, Java, python, PHP等API接口；接收标准的RDF文件格式（N3，Turtle等格式）

对比测试

实验环境

- Linux服务器(CentOS 7.2)
- CPU:　E5-2640 v3@2.60GHz
- RAM: 128GB RDIMM,2133Mt/s
- gStore版本：0.4.0
- 对比系统：Apache-Jena 3.0.1, Virtuoso-openlinksw 7.2
- 数据集：WatDiv, LUBM, BSBM, DBpedia

对比系统	系统性质	所属单位
Virtuoso	目前最好的、使用最广泛的**商业**RDF数据库系统	Openlink公司 (美国伯灵顿、1992成立)
Apache Jena	目前最有影响力的 **开源**RDF数据库项目	来源于HP研究院 (HP Lab) 2000年开始的项目； 2010年以后成为Apache开源项目

对比测试

数据集规模

Dataset	Size	Triple	Predicate	Entity	Literal
DBpedia 1B	172GB	1,111,481,066	124,034	139,493,254	94,130,070
WatDiv 100M	15GB	109,795,918	86	5,212,745	5,077,247
LUBM 500M	85GB	500,000,000	18	81,342,489	41,804,418
BSBM 300M	82GB	311,957,992	40	46,514,164	25,176,573
Freebase 2.5B	342GB	2,530,199,503	770,349	178,312,621	278,393,451

性能比较（国外同行的第三方测试）

【Vijay Ingalalli, Dino Ienco, Pascal Poncelet, Serena Villata: Querying RDF Data Using A Multigraph-based Approach. **EDBT 2016**: 245-256】

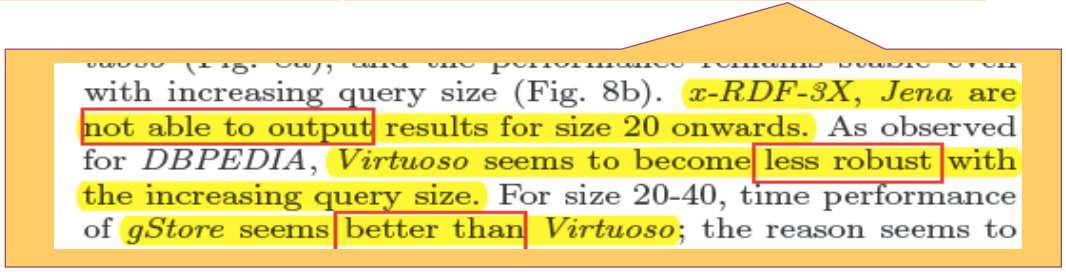
- 法国LIRMM实验室（录属于蒙彼利埃大学和法国国家科学研究中心CNRS）
- I3S实验室 CNRS(法国国家科学研究中心)

DBpedia 2014		
1.7亿 三元组	7百万 实体	23.8 GB 文件大小

对比系统	系统性质	同行实验总结
Apache Jena	目前最有影响力的**开源**RDF数据库系统	“当查询图大小超过20以后，系统就不能输出查询结果了”
x-RDF-3x	影响力较大的**学术界**系统	
Virtuoso	目前最好的、使用最广泛的**商业**RDF数据库系统	“对于查询图大小增长，系统性能的鲁棒性较差”
gStore（我们的系统）	Github上开源【Zou et al., VLDB 2011】	“gStore性能要好于其他对比系统”

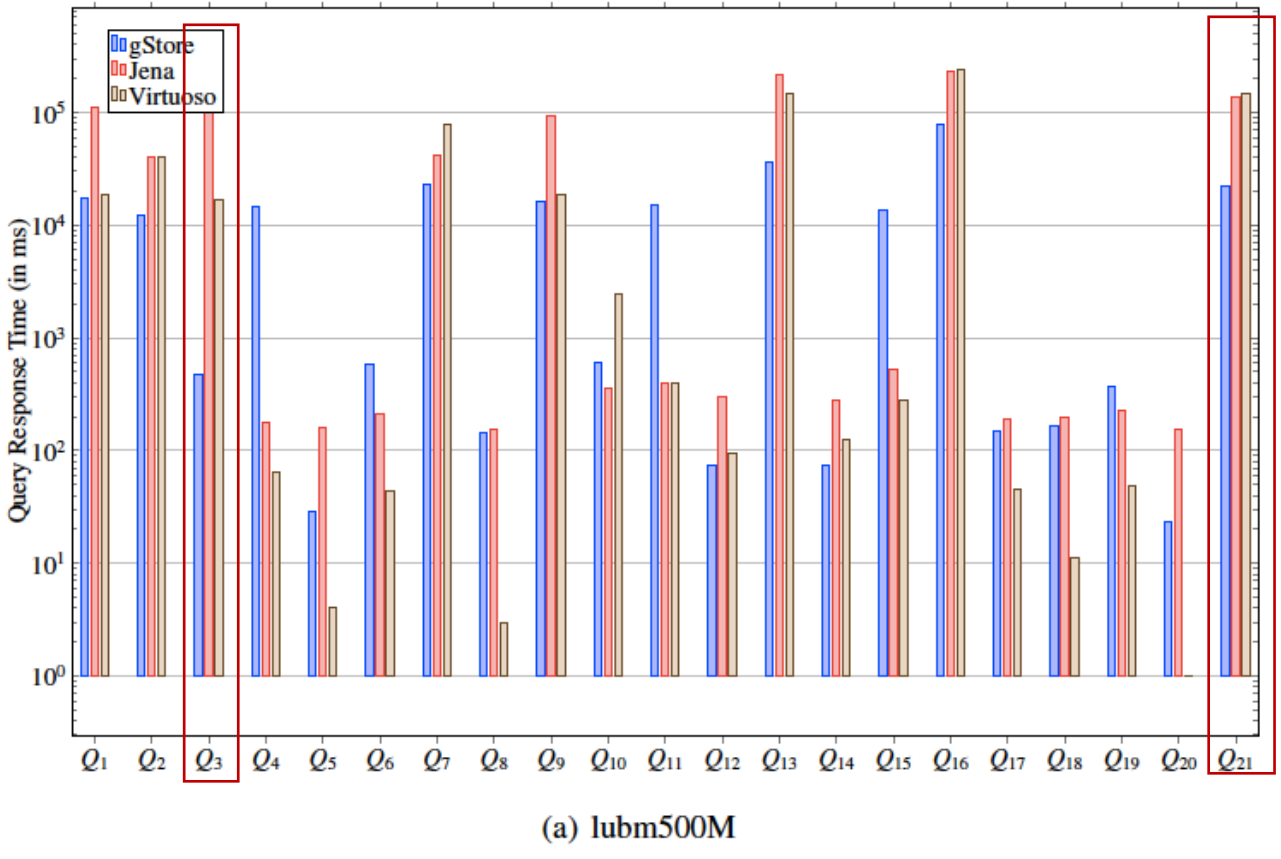
gStore	Virtuoso	RDF-3x
11.96 (秒)	20.45 （秒）	>60 （秒）

对20个针对Dbpedia数据集的查询的平均查询时间 --- 摘自 Ingalalli et. EDBT 16


with increasing query size (Fig. 8b). *x-RDF-3X*, *Jena* are not able to output results for size 20 onwards. As observed for *DBPEDIA*, *Virtuoso* seems to become less robust with the increasing query size. For size 20-40, time performance of *gStore* seems better than *Virtuoso*; the reason seems to

Testing

Query Performance over LUBM 500M

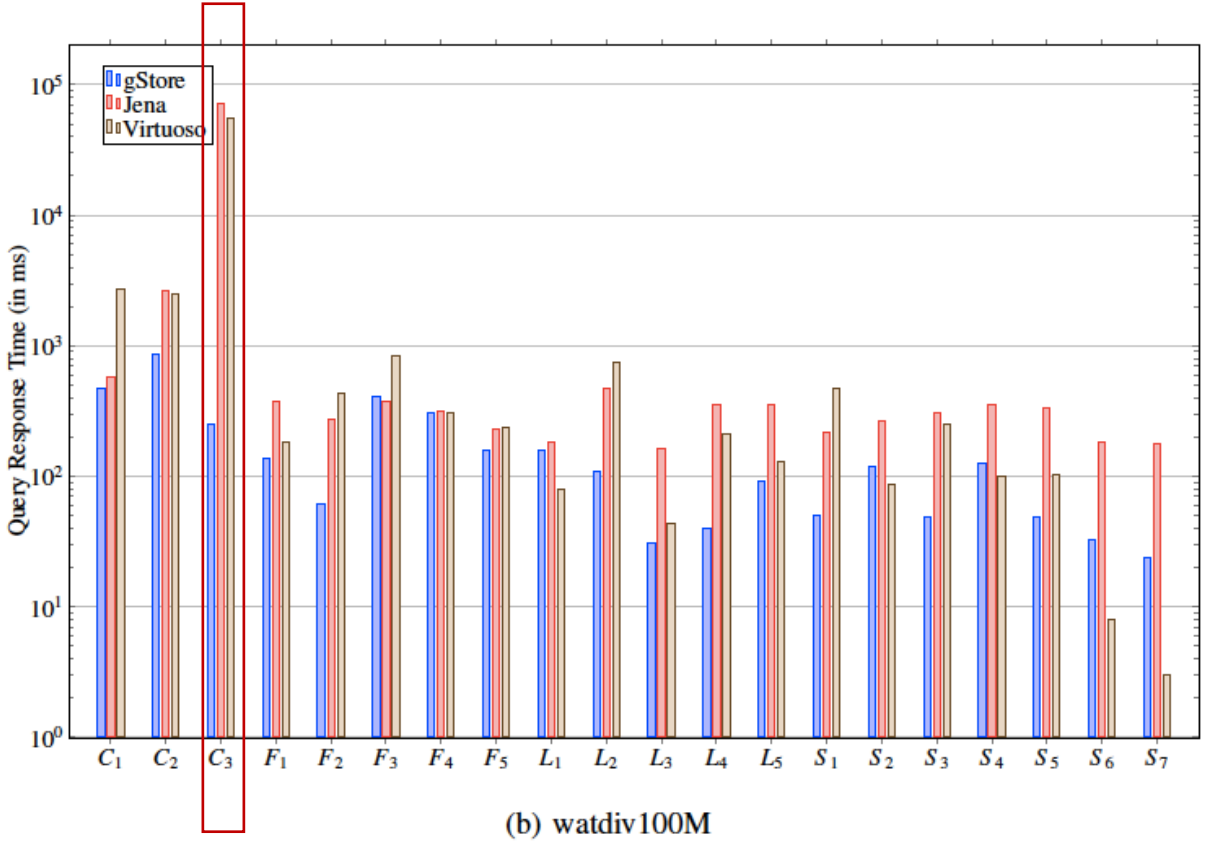


(a) lubm500M

查询	三元组数	返回列数
Q1,Q2,Q3	6,2,6	1
Q4	5	4
Q5	2	1
Q6	4	2
Q7	6	3
Q8	2	1
Q9	6	3
Q10	2	1
Q11	5	3
Q12,Q13	1	1
Q14	4	2
Q15,Q16	5	1,3
Q17-Q21	≤3	≤2

Testing

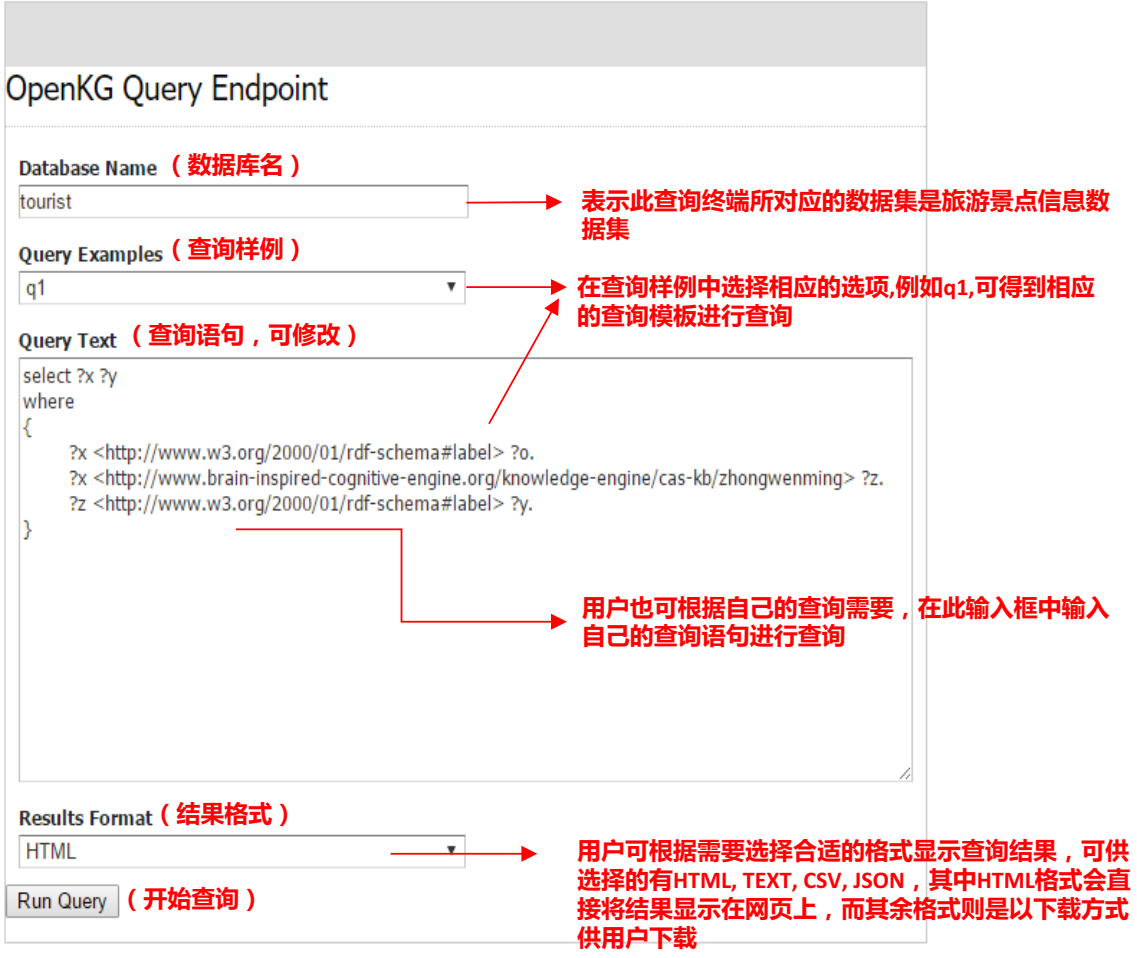
Query Performance over WatDiv 100M



(b) watdiv100M

查询	三元组数	返回列数
C1,C2	8,10	4
C3	6	1
F1-F3	6-8	5-7
F4	9	8
F5	6	6
L1-L5	≤3	≤3
S1	9	9
S2-S7	3-4	3-4

OpenKG查询终端主页： http://openkg.gstore-pku.com
http://sparql.openkg.cn



- 目前主页上共包括9个RDF数据集查询终端，分别是旅游信息、音乐、紧急事故、中文地理信息以及五个乳腺癌相关的数据集。
- 我们将为OpenKG**长期维护并不断改善**这些SPARQL Endpoint，有任何问题请及时向我们提出: bookug@qq.com

Java API样例程序：

```
String url = "http://" + "127.0.0.1" + ":" + "9000";
StringBuffer result = new StringBuffer();
BufferedReader in = null;
String param = "?operation=load&db_name=lubm";
String urlNameString = url + "/" +
    URLEncoder.encode(param, "UTF-8");

URL realUrl = new URL(urlNameString);
URLConnection connection = realUrl.openConnection();
connection.setRequestProperty("accept", "*/*");
connection.setRequestProperty("connection", "Keep-Alive");
connection.setRequestProperty("user-agent",
  "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1;SV1)");
connection.connect();

Map<String, List<String>> map = connection.getHeaderFields();
in = new BufferedReader(new
    InputStreamReader(connection.getInputStream()));
String line;
while ((line = in.readLine()) != null) {
    result.append(line);
}
```

Ip and port

http request: build, load, unload, query…

URL encode

Set connection

Receive the header and data from db server

Java API样例程序：

```java
import jgsc.GstoreConnector;
public class JavaAPIExample
{
    public static void main(String[] args)
    {
        GstoreConnector gc = new GstoreConnector("127.0.0.1", 9000);

        gc.build("LUBM10", "data/LUBM_10.n3");
        gc.load("LUBM10");

        String sparql = "select ?x where "
                + "{"
                + "?x    <rdf:type>    <ub:UndergraduateStudent>. "
                + "?y    <ub:name> <Course1>. "
                + "?x    <ub:takesCourse>  ?y. "
                + "?z    <ub:teacherOf>    ?y. "
                + "?z    <ub:name> <FullProfessor1>. "
                + "?z    <ub:worksFor>    ?w. "
                + "?w    <ub:name>    <Department0>. "
                + "}";
        String answer = gc.query(sparql);
        System.out.println(answer);

        gc.unload("LUBM10");

        gc.load("LUBM10");
        answer = gc.query(sparql);
        System.out.println(answer);
        gc.unload("LUBM10");
    }
}
```

initialize the GStore server's IP address and port.
for sparql endpoint, URL can also be used here, like tourist.gstore-pku.com:80

```
GstoreConnector gc = new GstoreConnector("tourist.gstore-pku.com", 80);
```

build a new database by a RDF file.
note that the relative path is related to gse

execute SPARQL query on this database.

unload this database.

also, you can load some exist database directly and then query.

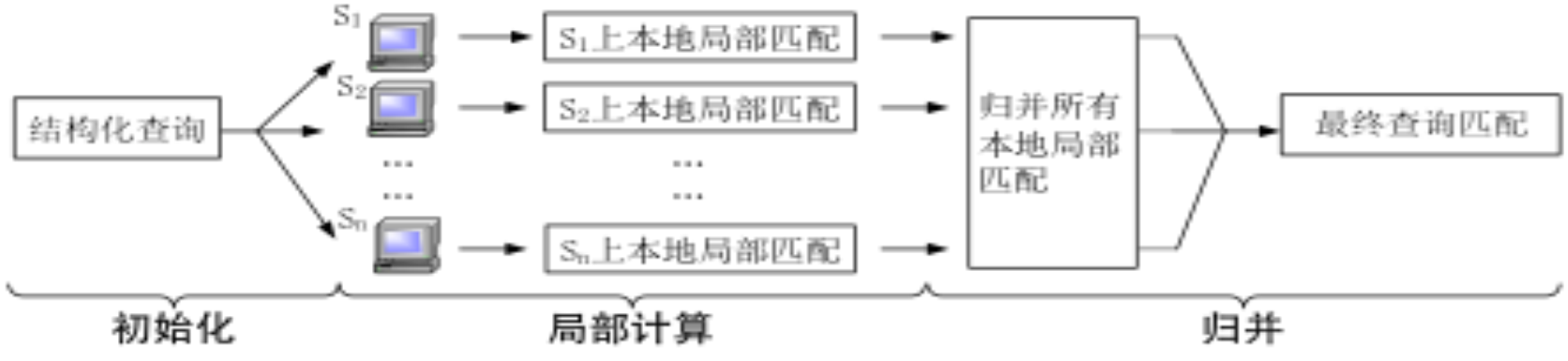
gStore---API 接口（C++; Java; python, PH

```
// initialize the Gstore server's IP address and port.
GstoreConnector gc("127.0.0.1", 3305);

// build a new database by a RDF file.
// note that the relative path is related to gserver.
gc.build("LUBM10.db", "example/LUBM_10.n3");

// then you can execute SPARQL query on this database.
std::string sparql = "select ?x where \
                     { \
                      ?x    <rdf:type>    <ub:UndergraduateStudent>. \
                      ?y    <ub:name> <Course1>. \
                      ?x    <ub:takesCourse>  ?y. \
                      ?z    <ub:teacherOf>    ?y. \
                      ?z    <ub:name> <FullProfessor1>. \
                      ?z    <ub:worksFor>    ?w. \
                      ?w    <ub:name>    <Department0>. \
                      }";
std::string answer = gc.query(sparql);

// unload this database.
gc.unload("LUBM10.db");

// also, you can load some exist database directly and then query.
gc.load("LUBM10.db");

// query a SPARQL in current database
answer = gc.query(sparql);
```

gStore-D: 分布式系统 [Peng; Zou, et al., VLDB J 16]

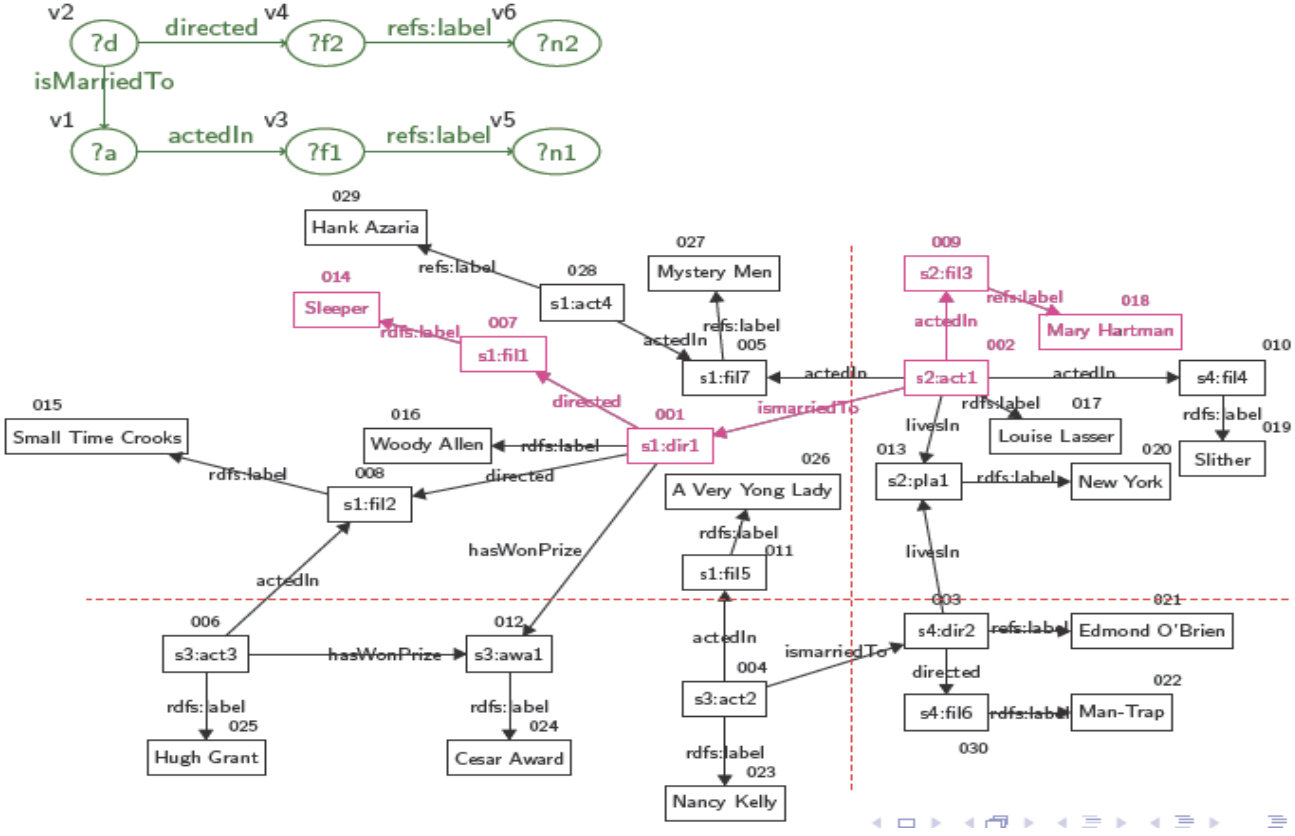
主要技术手段：

- 利用Partial Evaluation and Assembly方案来解决分布式SPARQL匹配；
- 分布式环境下的优化归并策略



gStore-D：分布式系统

主要技术问题：如何找到“跨界匹配”



gStore-D：分布式系统

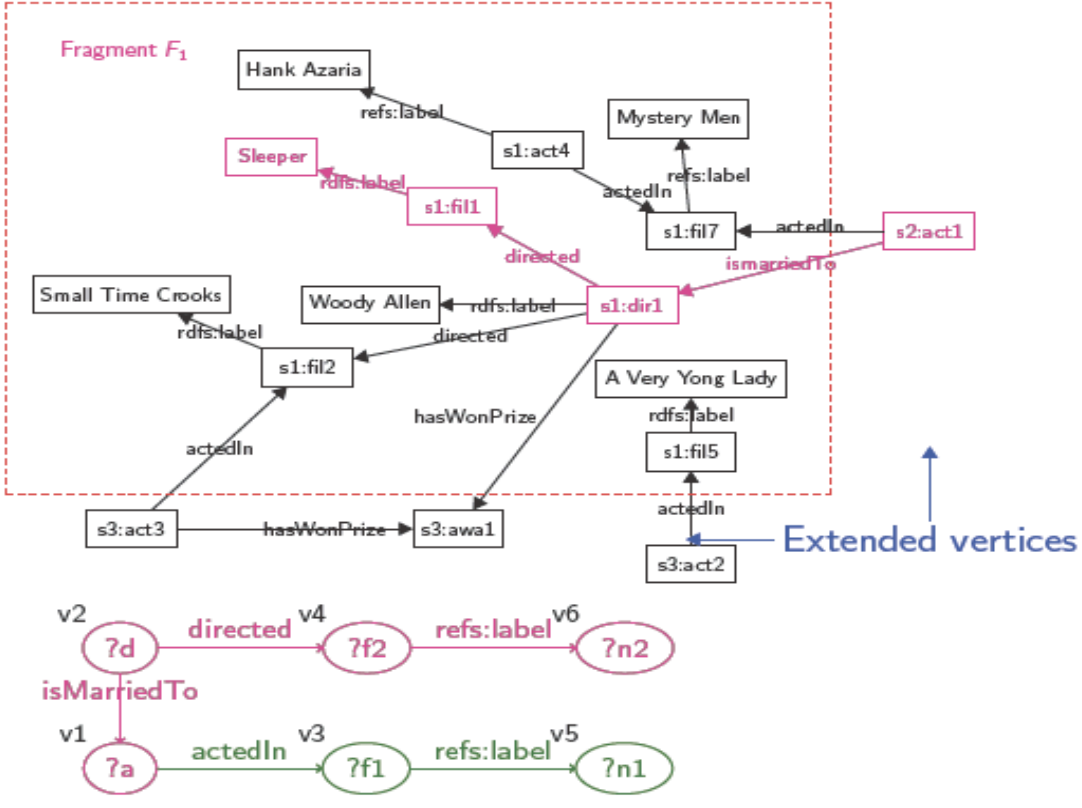
背景: 部分执行(Partial Evaluation) [Jones, 1996; Fan et al., 06; Shuai et al., 2012]

$$f(x) \Rightarrow f(s,d) \Rightarrow f''(f'(s),d) \Rightarrow \text{最终结果}$$

已知输入　未知输入　部分解

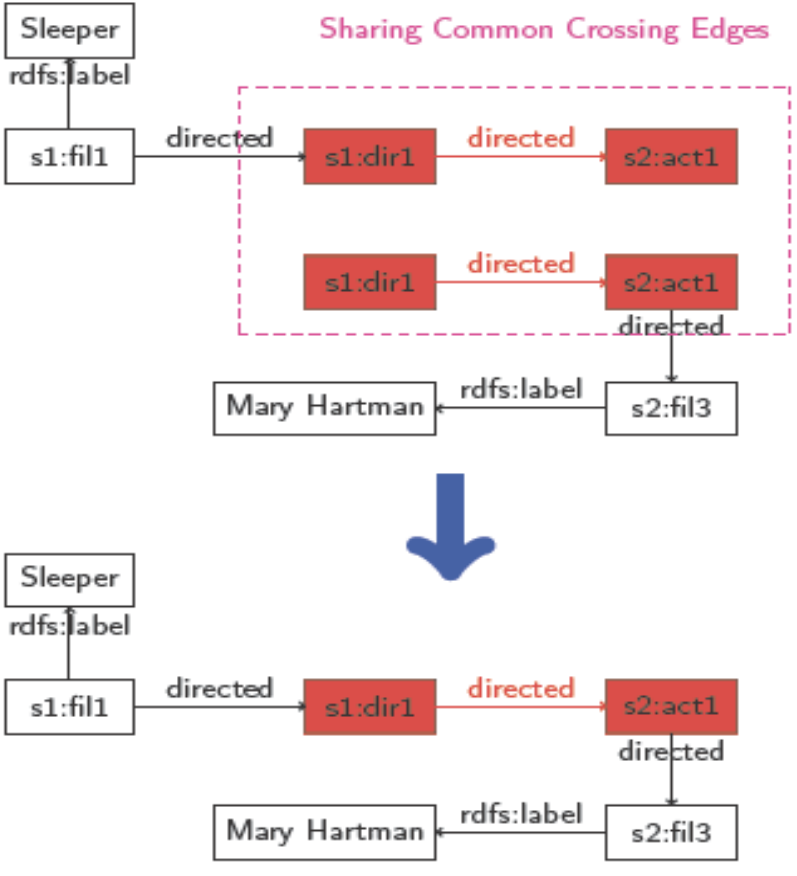
gStore-D：分布式系统

哪些是“已知输入”和“部分解”？



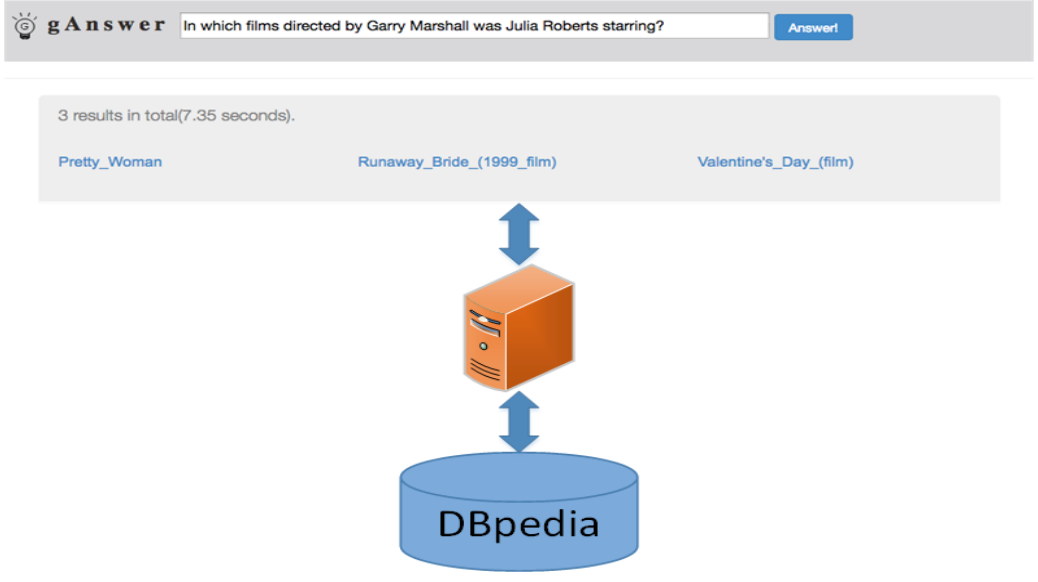
gStore-D：分布式系统

部分解合并



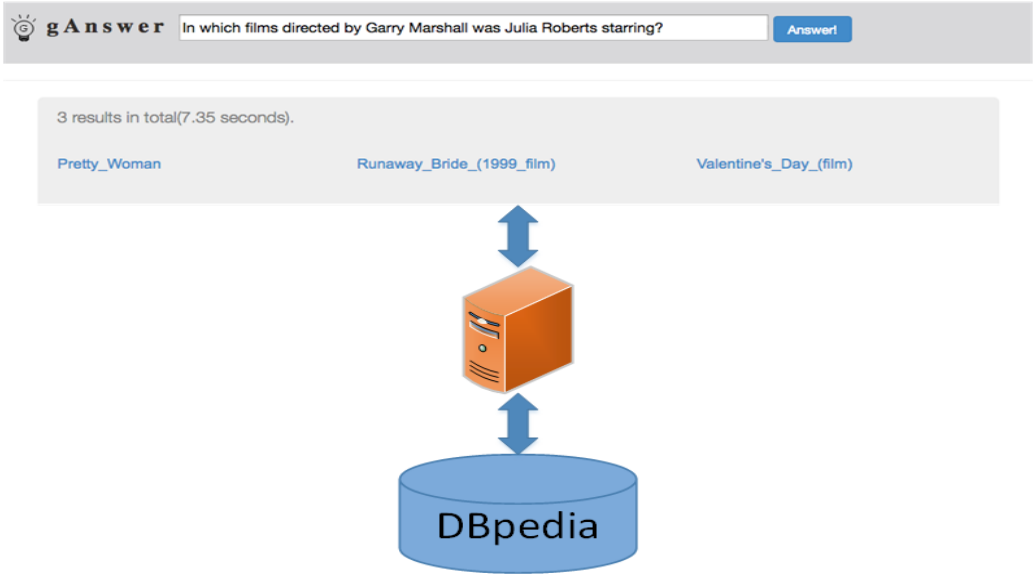
面向RDF知识图谱的问答系统

- 用户没有计算机学科背景，无法掌握SPARQL等计算机查询语言
- RDF知识图谱是“弱模式”数据，不同于关系数据是基于模式的数据。



面向RDF知识图谱的问答系统

- 提供方便的用户访问接口
- 数据库和自然语言处理的交叉研究
- 学术界和工业界共同关心的问题



面向RDF知识图谱的问答系统

COMMENT

GENETICS Growth in genome screening could cause dangerous meddling p.27 | EVOLUTION How genes and culture have shaped our ability to cooperate p.29 | CHEMISTRY Debating how life got going on the early Earth p.30 | EXHIBITION Wildlife paintings from Yukon to Yellowstone p.32

Search needs a shake-up

On the twentieth anniversary of the World Wide Web's public release, Oren Etzioni calls on researchers to think outside the keyword box and improve Internet trawling.

Two decades after Internet pioneer Tim Berners-Lee introduced his World Wide Web project to the world using the alt.hypertext newsgroup, web search is on the cusp of a profound change — from simple document retrieval to question answering. Instead of poring over long lists of documents that contain requested keywords, users need direct answers to their questions. With sufficient scientific and financial investment, we could soon view today's keyword searching with the same nostalgia and amusement reserved for bygone technologies such as electric typewriters and vinyl records.

But this transformation could be unreasonably delayed. As a community, computer scientists have underinvested in tools that can synthesize sophisticated answers to questions, and have instead focused on incremental progress in lowest-common-denominator search. The classic keyword search box exerts a powerful gravitational pull. Academics and industry researchers need to achieve the intellectual 'escape velocity' necessary to revolutionize search. They must invest much more in bold strategies that can achieve natural-language searching and answering, rather than providing the electronic equivalent of the index at the back of a reference book.

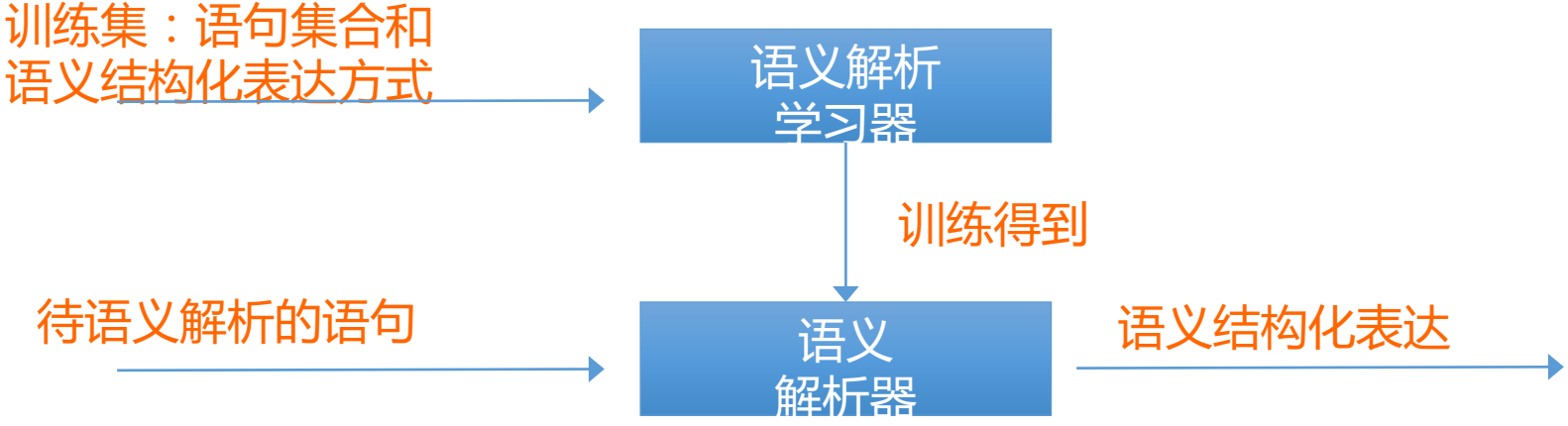
Today, that 'book' is distributed over billions of web pages of uneven quality, and much effort has been directed at ranking the most useful results. Such engines readily index billions of documents, but overwhelm their users with millions of results in response to simple queries. This quandary only worsens as the number of web pages ▸

Oren Etzioni, AAAI Fellow

译文：“工业界和学术界的研究者应该更加大胆地研究自然语言的搜索和问答”
---《Search needs a shake up（搜索需要重塑）》, NATURE, Vol 476, p25-26, 2011.

面向RDF知识图谱的问答系统

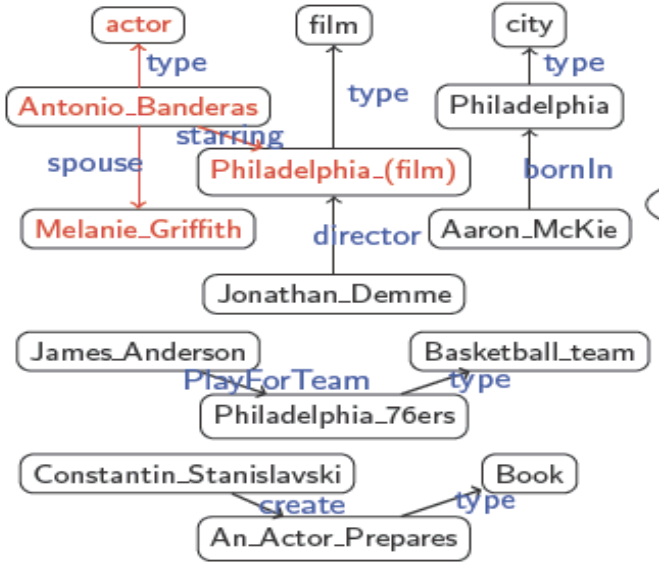
- 基于自然语言的语法规则推导的方法, e.g., CDG [Zettlemoyer and Collins, 2005]
- 基于手工书写规则的方法， e.g., [Unger, 2006]
- 基于机器学习的方法，e.g., 基于监督学习 [Wong, 2006]

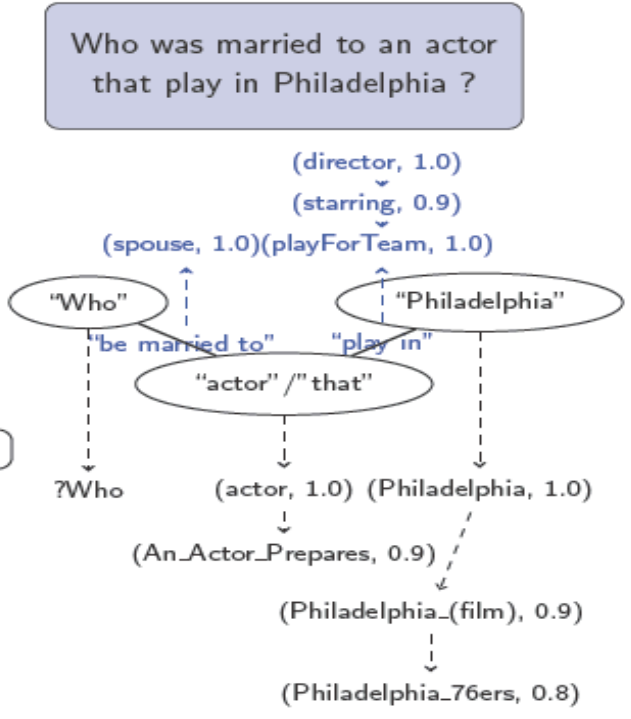


我们的方法-数据驱动策略

gAnswer [Zou et al, SIGMOD 14]

- 将自然语言问答转换为子图匹配问题
- 基于子图匹配结果的消歧
- 基于结构的查询图生成策略

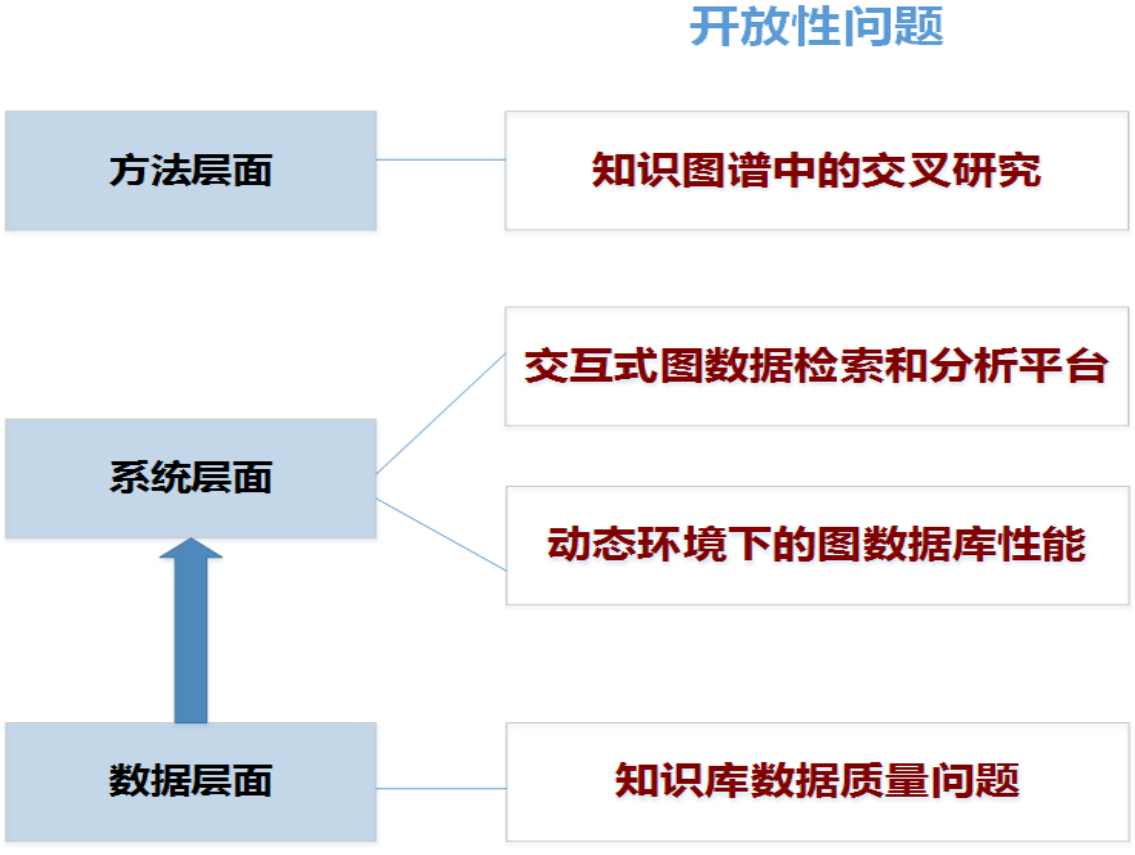




提纲

1. 知识图谱概述
2. 从不同角度和学科研究
3. 从数据管理层面的讨论
4. 一些开放性问题
5. 系统应用
6. 总结

一些开放性问题


开放性问题
方法层面
知识图谱中的交叉研究
系统层面
交互式图数据检索和分析平台
动态环境下的图数据库性能
数据层面
知识库数据质量问题

知识库数据质量问题

1. Wenfei Fan, Yinghui Wu, Jingbo Xu, Functional Dependencies for Graphs, SIGMOD, 2016.
2. Binbin He, Lei Zou, Dongyan Zhao: Using Conditional Functional Dependency to Discover Abnormal Data in RDF Graphs. SWIM 2014: 43:1-43:7

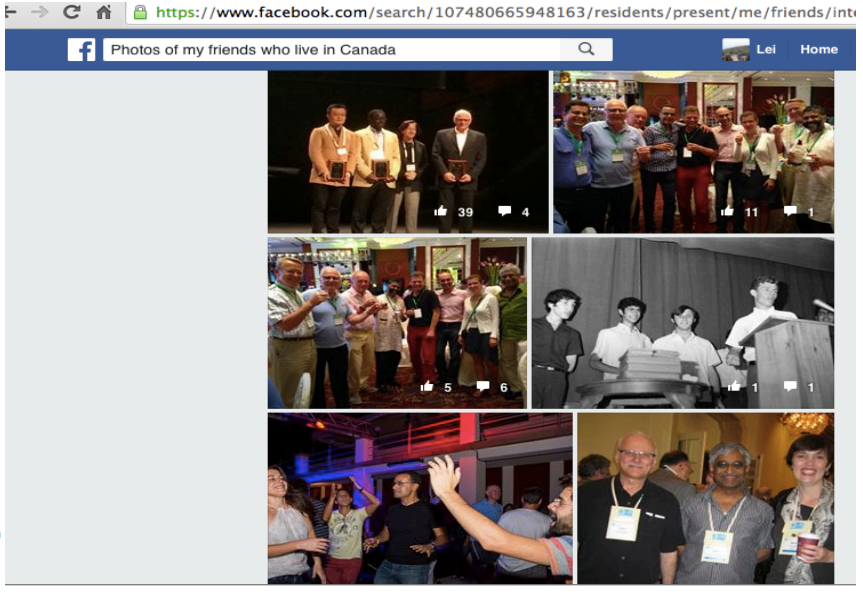
百度2014年7月的“吉林大学”词条

目前的“吉林大学”词条

动态环境下的图数据库性能

回顾一下这个例子！

“Photos of my friends who live in Canada”



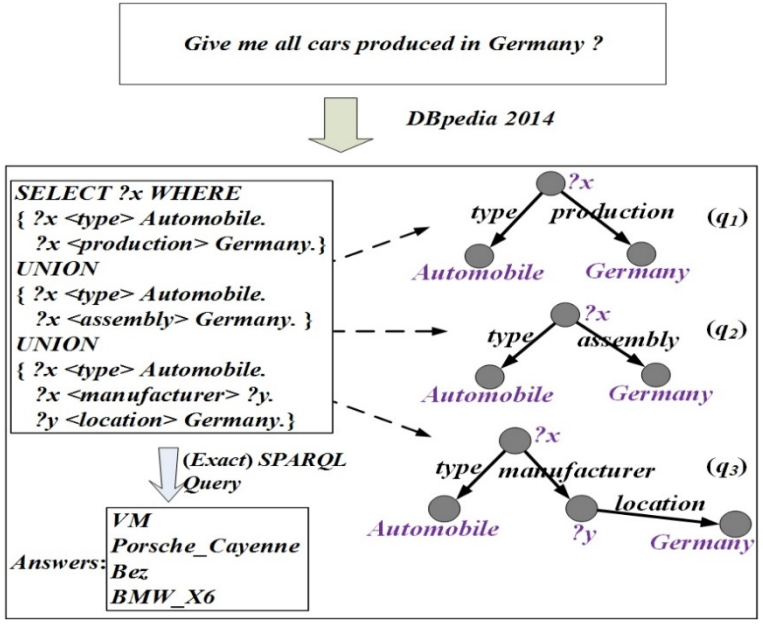
每秒中有多少照片和用户关系会被插入和删除？

交互式图数据检索和分析平台

问题由来： 知识图谱中的数据的异构性更强：

问题举例：在Dbpedia知识图谱中，
“德国生产的汽车”有5种以上表达模式！

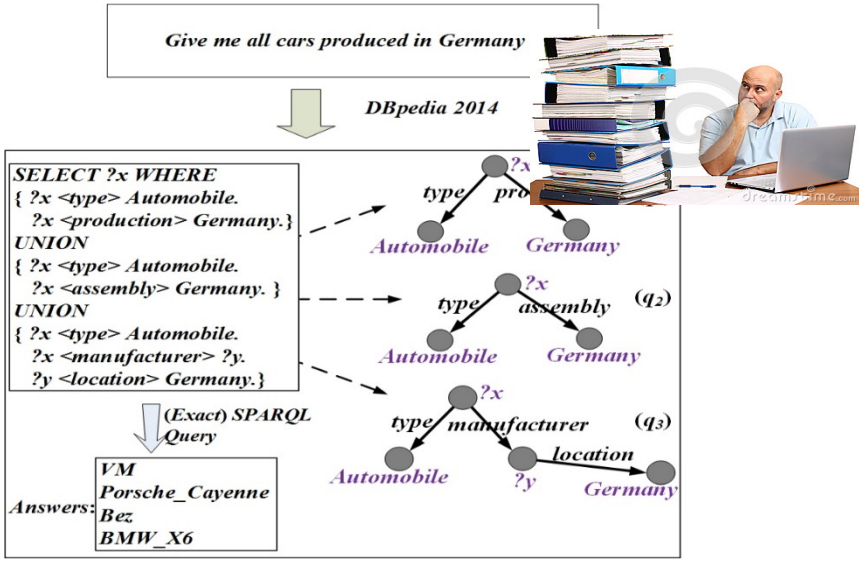
穷举所有的可能性？



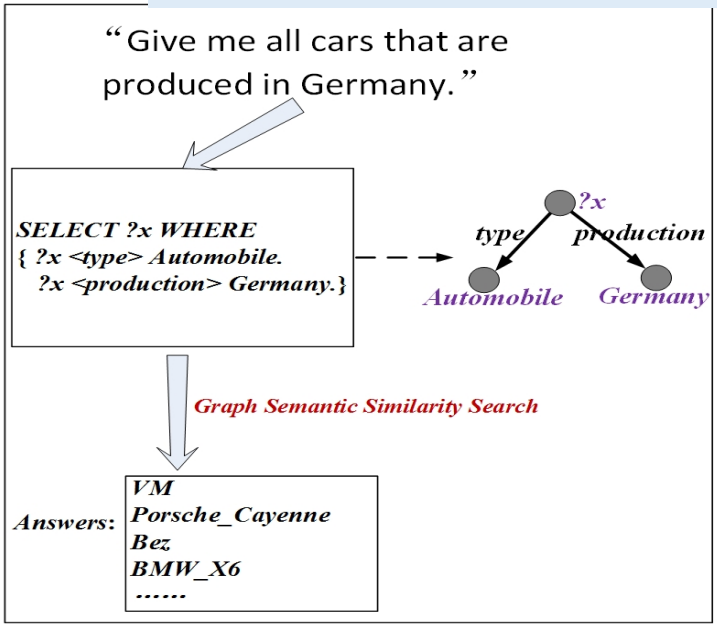
交互式图数据检索和分析平台

问题由来： 知识图谱中的数据的异构性更强：

问题举例：在Dbpedia知识图谱中，
“德国生产的汽车”有5种以上表达模式！



在结构近似的基础上，
提出 “**语义结构近似**”



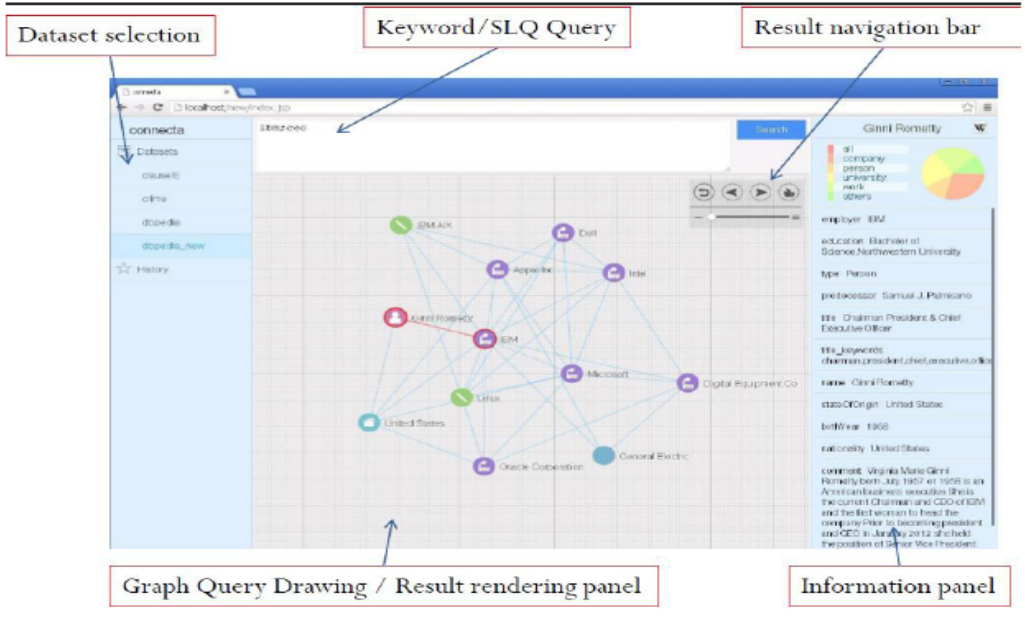
我们的方法

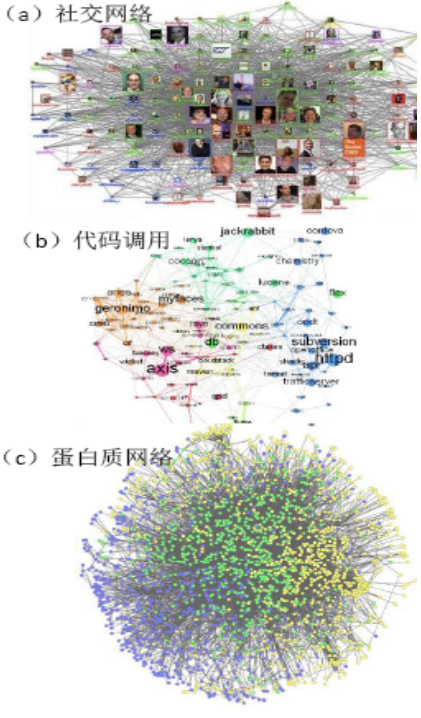
Weiguo Zheng (指导学生), **Lei Zou**, et al., Semantic SPARQL Similarity Search Over RDF Knowledge Graphs, in **VLDB** 2016.

交互式图数据检索和分析平台

“机器的归机器，人的归人！”

Mengxia Jiang, Yueguo Chen, Jinchuan Chen, Xiaoyong Du: Interactive Predicate Suggestion for Keyword Search on RDF Graphs. ADMA (2) 2011: 96-109

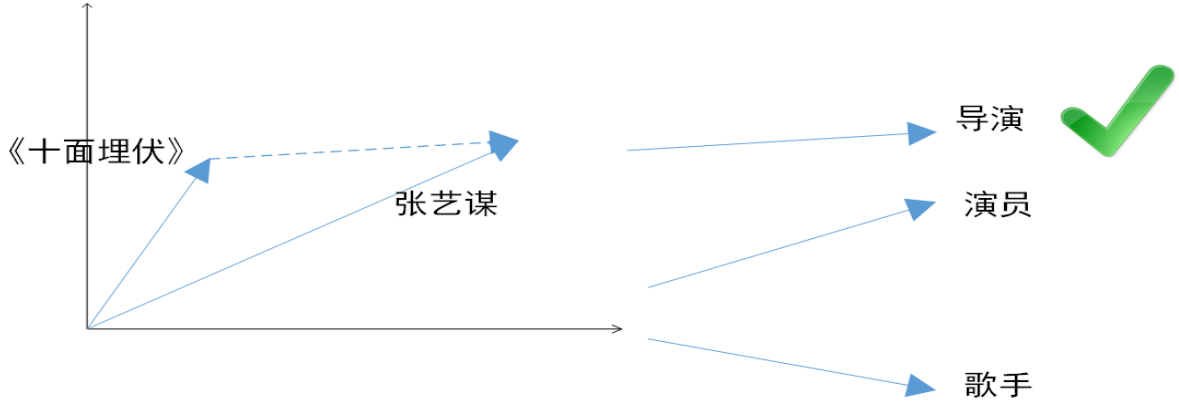




知识图谱中的交叉研究

表示学习（机器学习）+ 数据质量（数据管理）

主语	谓词	宾语
《美国队长3》	导演	乔·卢素
《重庆森林》	导演	王家卫
《十面埋伏》	歌手	张艺谋



知识图谱中的交叉研究

语义解析（自然语言处理）+查询执行（数据管理）

这附近最近的邮局在哪儿？

$$\arg\min(\lambda x.POST(x) \wedge dis(HERE, x))$$

SPARQL

```
SELECT ?x WHERE {
?x rdf:type Post.
?x :longitude ?o.
?x :latitude ?a. }
ORDERY BY Dist(HERE, [?o, ?a])
LIMIT 1
```

知识图谱中的交叉研究

语义解析（自然语言处理）+查询执行（数据管理）

这附近最近的邮局在哪儿？

$$\arg\min(\lambda x.POST(x) \wedge dis(HERE, x))$$

SPARQL

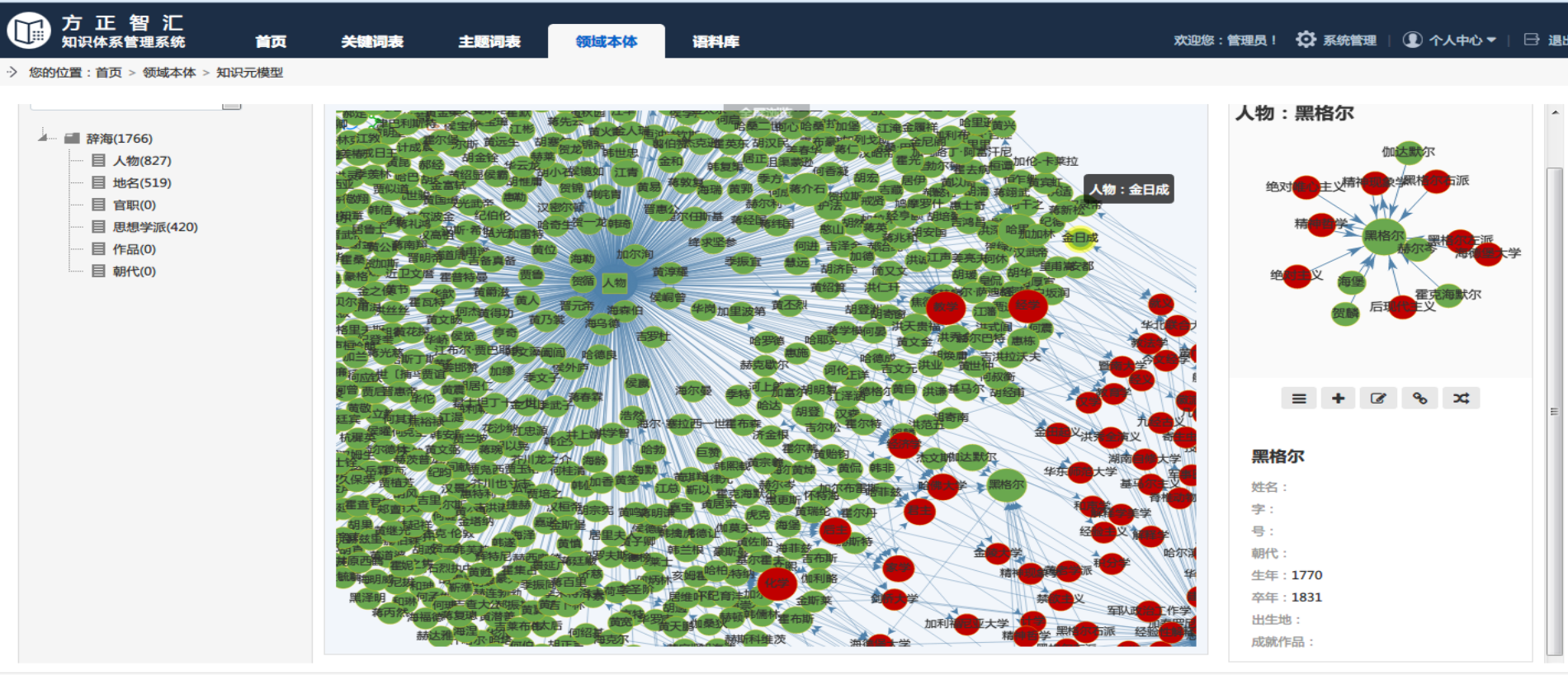
```
SELECT ?x WHERE {
?x rdf:type Post.
?x :longitude ?o.
?x :latitude ?a. }
ORDERY BY Dist(HERE, [?o, ?a])
LIMIT 1
```

提纲

1. 知识图谱概述
2. 从不同角度和学科研究
3. 从数据管理层面的讨论
4. 一些开放性问题
5. 系统应用
6. 总结

gStore应用

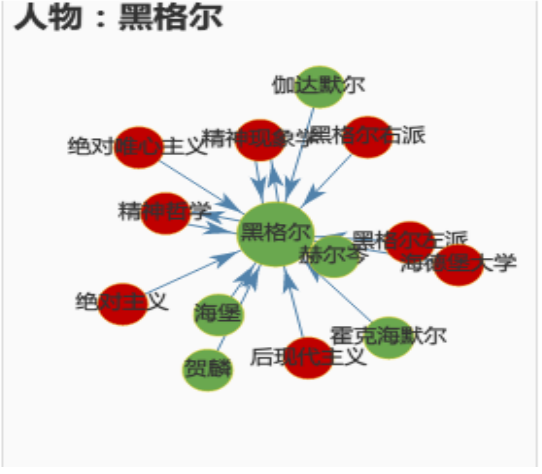
- 方正电子知识出版系统



gStore应用

- 方正电子知识出版系统

```
SELECT *
WHERE
{
   ?s <http://www.founder.106.attr:name> ?name.
   {     ?s <http://www.founder.106.link:12855> ?o. //思想学派受影响于
   }
   UNION
   {     ?s <http://www.founder.106.link:12855> ?o. //哲学家受影响于
   }
   ?o <http://www.founder.106.attr:name> "黑格尔".
}
```




```
//删除所有与黑格尔有关的三元组
DELETE {?s ?p ?o}
WHERE
{
   ?s ?p ?o.
   {   ?s <http://www.founder.106.attr:name> "黑格尔".}
   UNION
   {   ?o <http://www.founder.106.attr:name> "黑格尔".}
}
```

gStore应用

- 中科院微生物所-全球微生物中心 gStore

# of Triples	# of Entities
3,594,457,749	414,953,654

Bacteria > Terrabacteria group > Actinobacteria > Actinobacteria > Micrococcales > Micrococcaceae > Micrococcus > Micrococcus luteus

细菌 陆生菌 放线菌门 放线菌纲 微球菌目 微球菌科 微球菌属 藤黄微球菌

Overview Taxonomy Genome Feature GO Pathway Literature

Species Information

Taxonomy	Bacteria > Terrabacteria group > Actinobacteria > Actinobacteria > Micrococcales > Micrococcaceae > Micrococcus > Micrococcus luteus
NCBI taxonomy ID	1270
Scientific Name	Micrococcus luteus
Children	Micrococcus luteus CD1_FAA_NB_1 Micrococcus luteus J28 Micrococcus luteus Mu201 Micrococcus luteus NCTC 2665 Micrococcus luteus SK58 Micrococcus luteus str. modasa More
Reference Title In IJSEM	
Type Strains	
Strains	

```
PREFIX annotation:
<http://gcm.wdcm.org/ontology/gcmAnnotation/v1/>
PREFIX taxonomy:
<http://gcm.wdcm.org/data/gcmAnnotation1/taxonomy/>

SELECT ?taxonId ?name
WHERE
{
        ?taxonId  annotation:parentTaxid  taxonomy:1270.
        ?nameId  annotation:taxid  ?taxonId.
        ?nameId  annotation:nameclass  'scientificName'.
        ?nameId  annotation:taxname  ?name.
}
```

"查询藤黄微球菌下面的菌株"

gStore应用

- 中科院微生物所-全球微生物中心 gStore

# of Triples	# of Entities
3,594,457,749	414,953,654

```
PREFIX annotation:
<http://gcm.wdcm.org/ontology/gcmAnnotation/v1/>
PREFIX taxonomy:
<http://gcm.wdcm.org/data/gcmAnnotation1/taxonomy/>

SELECT (COUNT(?geneid) AS ?num)
WHERE
{
        {       ?taxonid  annotation:ancestorTaxid  taxonomy:1270.
                ?geneid  a  annotation:GeneNode.
                ?geneid  annotation:x-taxon  ?taxonid.
        }UNION
        {       ?geneid  a  annotation:GeneNode.
                ?geneid  annotation:x-taxon  taxonomy:1270.
        }
}
```

Number of Gene	54824
Number of Protein	16229

Annotation summary

Proteins with PDB structures	15
Proteins with Pfam assignments	2008
Proteins with GO assignments	32453
Proteins with EC number assignments	680
Proteins with Pathway assignments	2398

Publications and Patents

Publications	
Patents	

"和藤黄微球菌物种或者下面的菌株相关的基因的个数"

gStore应用

- 中科院微生物所-全球微生物中心 gStore

# of Triples	# of Entities
3,594,457,749	414,953,654

Genome

Export Excel

Organism Name	Genome Accession	Description
Micrococcus luteus str. modasa	AMYK02000110	Micrococcus luteus str. modasa contig_110, whole genome shotgunsequence.
Micrococcus luteus NCTC 2665	CP001628	Micrococcus luteus NCTC 2665, complete genome.
Micrococcus luteus SK58	ADCD01000097	Micrococcus luteus SK58 ctg1119142780327, whole genome shotgunsequence.
Micrococcus luteus str. modasa	AMYK02000273	Micrococcus luteus str. modasa contig_273, whole genome shotgunsequence.
Micrococcus luteus str. modasa	AMYK02000081	Micrococcus luteus str. modasa contig_81, whole genome shotgunsequence.
Micrococcus luteus str. modasa	AMYK02000252	Micrococcus luteus str. modasa contig_252, whole genome shotgunsequence.
Micrococcus luteus str. modasa	AMYK02000060	Micrococcus luteus str. modasa contig_60, whole genome shotgunsequence.
		Micrococcus luteus str.

“查询藤黄微球菌下面的菌株相关的基因组、描述之类的信息”

```
PREFIX annotation:
<http://gcm.wdcm.org/ontology/gcmAnnotation/v1/>
PREFIX taxonomy:
<http://gcm.wdcm.org/data/gcmAnnotation1/taxonomy/>

SELECT ?taxonid ?name ?genomeid ?description ?strain
WHERE
{
        ?taxonid  annotation:ancestorTaxid  taxonomy:1270.
        ?nameId  a  annotation:TaxonName.
        ?nameId  annotation:taxid  ?taxonid.
        ?nameId  annotation:nameclass  ‘scientificName’.
        ?nameId  annotation:taxname  ?name.
        ?genomeid  a  annotation:GenomeNode.
        ?genomeid  annotation:x-taxon  ?taxonid.
        ?genomeid  annotation:definition  ?description.
        optional{?genomeid annotation:strain ?strain.}
}
```

提纲

1. 知识图谱概述
2. 从不同角度和学科研究
3. 从数据管理层面的讨论
4. 一些开放性问题
5. 系统应用
6. 总结

Take-home Message

1. 路线

基于图的RDF数据管理是可行的技术路线

2. 方法

图匹配查询是有效的技术手段

参考文献：

- Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, Oksana Yakhnenko: Translating Embeddings for Modeling Multi-relational Data. NIPS 2013: 2787-2795
- Luke S. Zettlemoyer, Michael Collins: Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars. UAI 2005: 658-666
- Pablo N. Mendes, Max Jakob, Christian Bizer: DBpedia: A Multilingual Cross-domain Knowledge Base. LREC 2012: 1813-1817
- Fabian M. Suchanek, Gjergji Kasneci and Gerhard Weikum, Yago - A Core of Semantic Knowledge, 16th international World Wide Web conference (WWW 2007)
- Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, Jamie Taylor: Freebase: a collaboratively created graph database for structuring human knowledge. SIGMOD Conference 2008: 1247-1250
- Peter Buneman, Gao Cong, Wenfei Fan, Anastasios Kementsietsidis: Using Partial Evaluation in Distributed Query Evaluation. VLDB 2006: 211-222
- Yuk Wah Wong, Raymond J. Mooney: Learning for Semantic Parsing with Statistical Machine Translation. HLT-NAACL 2006
- C. Unger, L. Bühmann, J. Lehmann, A.-C. N. Ngomo, D. Gerber, and P. Cimiano. Template-based question answering over RDF data. In WWW, pages 639–648, 2012
- Lei Zou, Ruizhe Huang, Haixun Wang, Jeffrey Xu Yu, Wenqiang He and Dongyan Zhao, Natural Language Question Answering over RDF ---- A Graph Data Driven Approach , SIGMOD (2014)
- Lei Zou, Jinghui Mo, Lei Chen,M. Tamer Özsu, Dongyan Zhao, gStore: Answering SPARQL Queries Via Subgraph Matching, in Proceedings of 37th International Conference on Very Large Databases (VLDB), 2011.
- Peng Peng, Lei Zou, Tamer Ozsu, Lei Chen, Dongyan Zhao, Processing SPARQL queries over distributed RDF graphs, accepted by VLDB Journal

谢 谢 ！

zoulei@pku.edu.cn