知识图谱：
大数据语义链接的基石

李涓子

清华大学

2014年10月17日

一段真实的经历

網易新闻 网易首页 > 新闻中心 > 滚动新闻 > 正文

法航罢工(图)

2014-09-17 08:31:28 来源: 京华时报(北京) 有0人参与 分享到

法航过半航班被迫取消，法国戴高乐机场法航柜台前显得十分冷清。图/东方IC

京华时报（微博）讯 （记者韩旭）由于法国航空公司旗下飞行员15日起开始为期一周的大罢工，该公司过半航班取消。

记者昨晚在首都机场官网看到，从巴黎出发的法航AF128次航班计划5：45抵达首都机场，AF382次航班从巴黎出发计划到港时间为15：15，但航班信息显示，这两架航班都未起飞，航班取消。截至发稿，尚无法证实取消航班是否与法航罢工相关。

法航罢工新闻

易达旅行
010-89678959
工作时间：周一至周日 08:30-23:00

单程： 广州 - 巴黎 约12小时50分钟

法国航空公司
AF107 波音777(大)

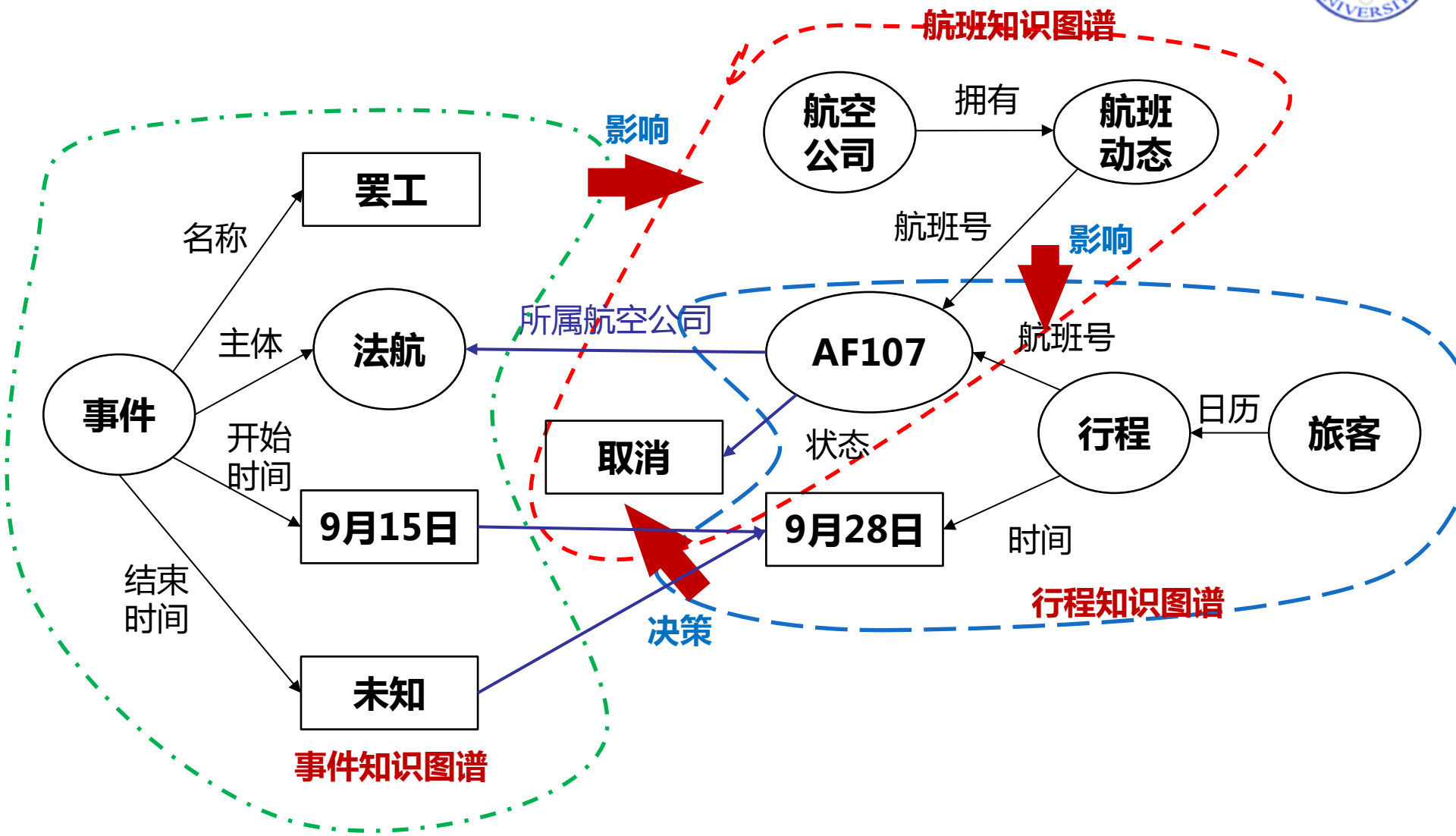
起飞 9月28日 23:00 白云机场
飞行 约12小时50分钟
到达 9月29日 05:50 戴高乐机场

成人退改签规则：
退票规定：不得退票
改期规定：不得改期
行李额规定：1pc
其他说明：根据航空公司规定执行
友情提示：团队签证无法单独购买机票，详情请咨询签证办理方。

旅客行程安排

背景：9月中旬，法国航空公司发生飞行员为期10天以上的大规模罢工，多次航班因此取消

语义链接与信息主动推送



主要内容

一、知识图谱基础

二、知识图谱类型

三、知识图谱构建方法及关键技术

四、基于知识图谱的语义链接及其应用

知识图谱基础

DBpedia

250概念
4M实例
6000属性
500M三元组
在线更新

yago
select knowledge

350K概念
10M实例
100属性
120M三元组

搜狗搜索

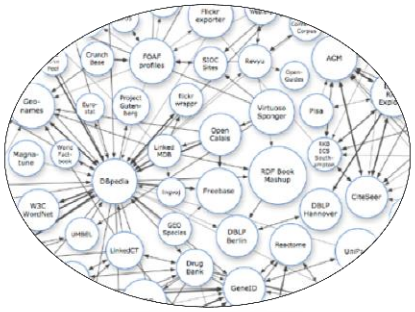
Baidu 百度

BabelNet 1.1.1
A very large multilingual ontology

50M义项
50+种语言
262M三元组

Freebase

15K概念
40M实例
4000属性
1B三元组
Google KB核心



WolframAlpha

NELL

OpenIE
(Reverb, OLLIE)

XLORE

850K概念
8M实例
70K属性

Google KG

15K概念
600M实例
20B三元组

WordNet
7种欧洲语言
跨语言链接

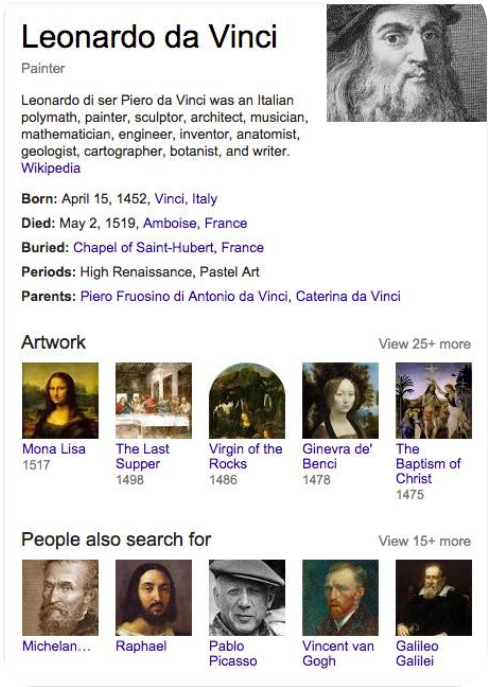
EuroWordNet

知识图谱基础

知识图谱

知识图谱，也称为科学知识图谱，它通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论与方法与计量学引文分析、共现分析等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构达到多学科融合目的的现代理论。为学科研究提供切实的、有价值的参考。

--- 百度百科



Google知识图谱

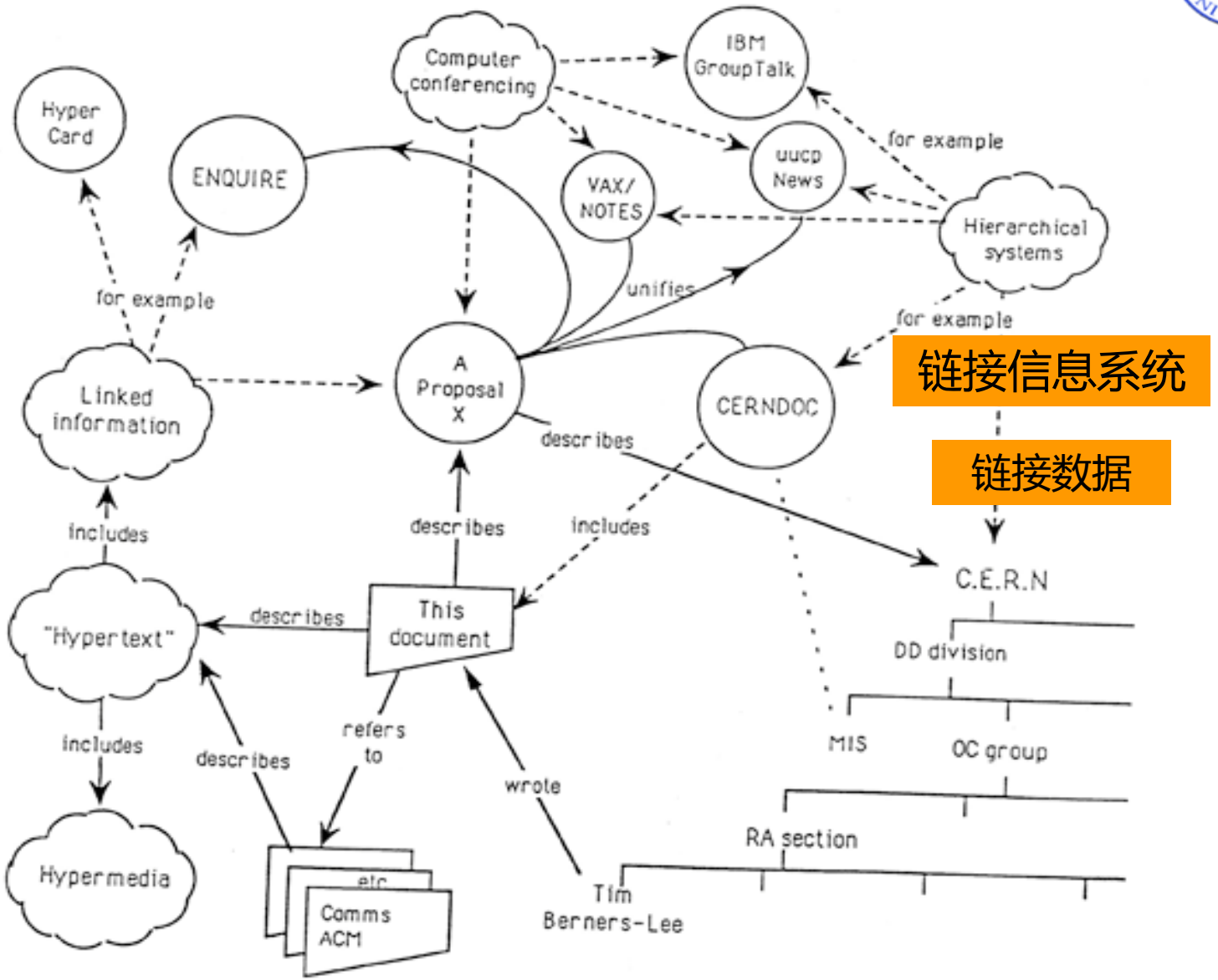
实体及其之间的关系图。

规模：5亿个对象，35亿个事实和关系

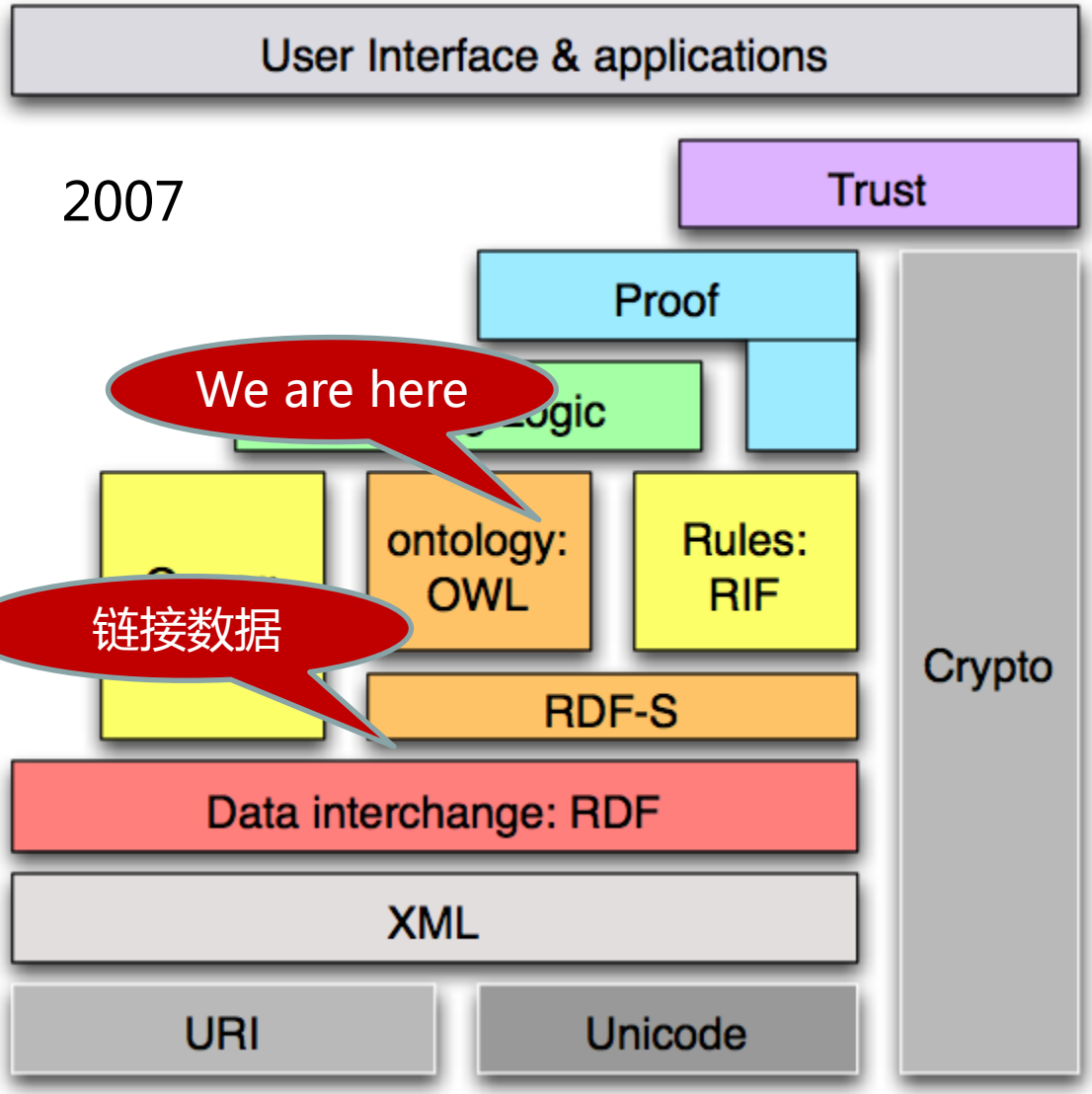
---维基百科

知识图谱的本质：知识库？语义网络？
知识图谱的形式：RDF？Graph？

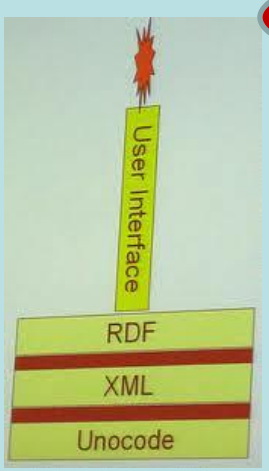
Tim Berners-Lee's Proposal 1989



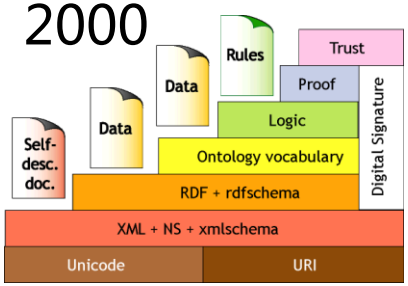
万维网信息描述语言塔



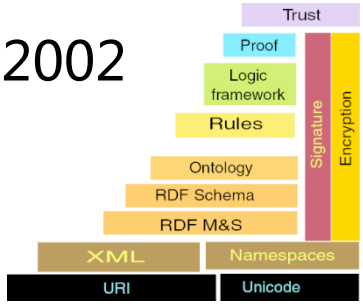
2006 swui Candle



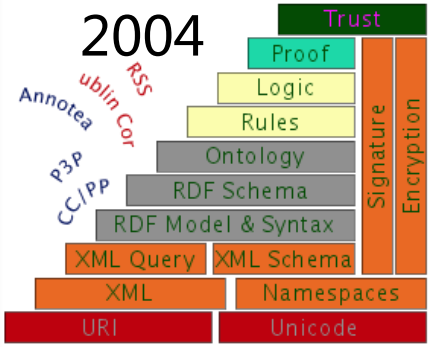
2000



2002



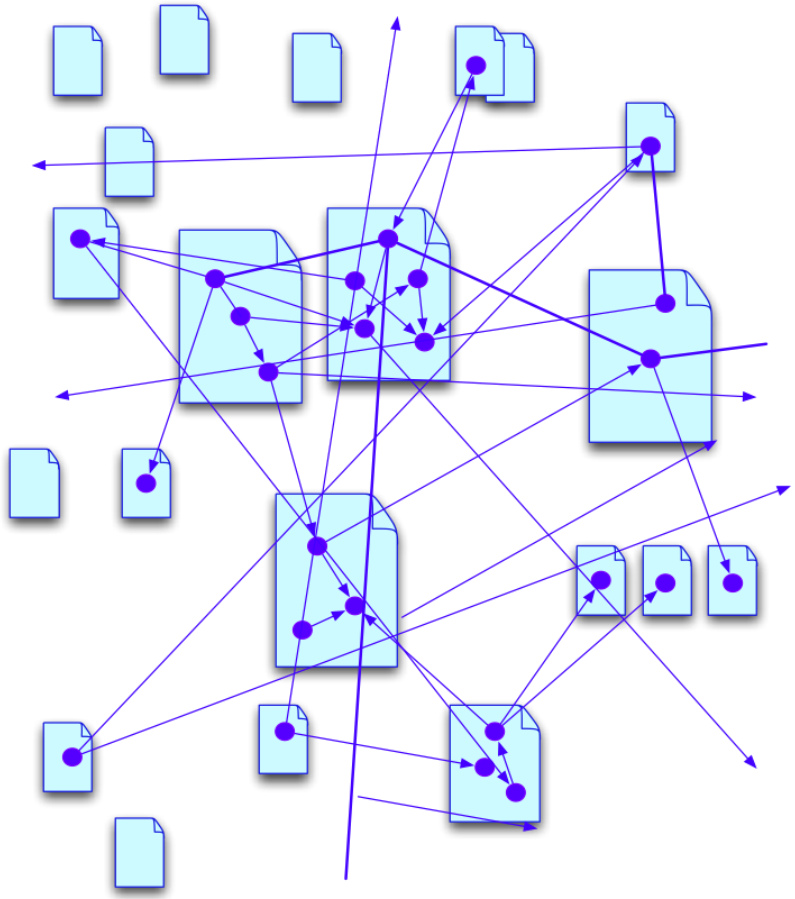
2004



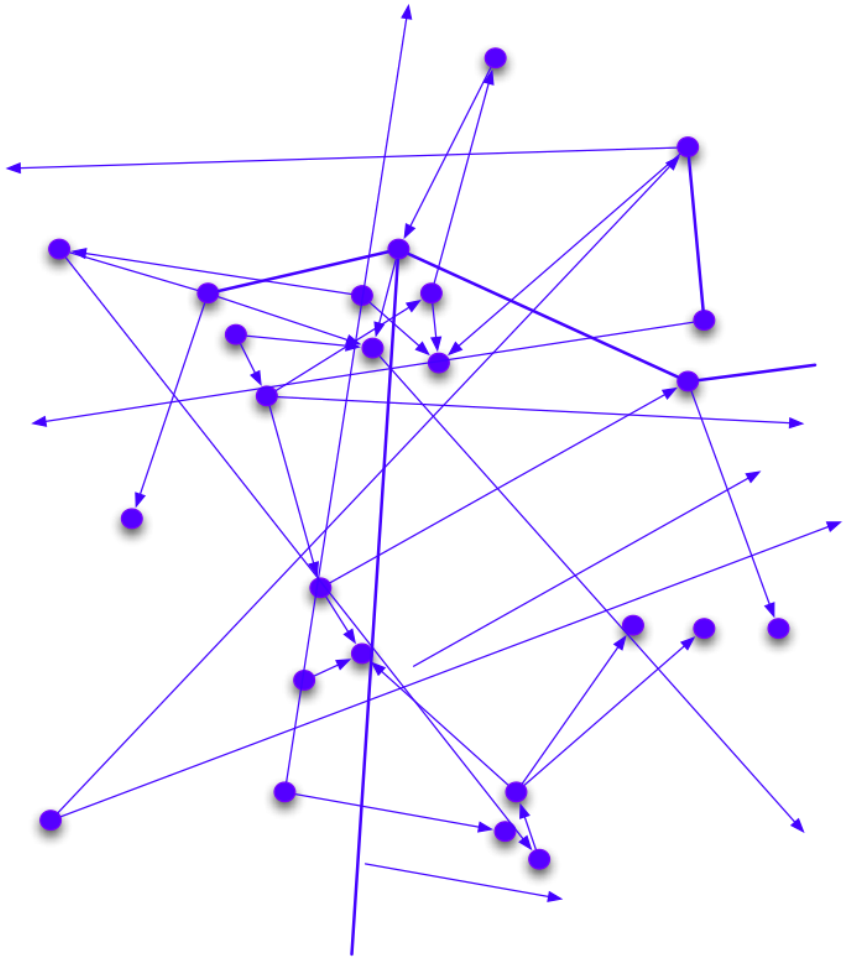
http://bbs.w3china.org/dispbbs.asp?boardID=2&ID=86430

从文档万维网到数据万维网

文档万维网



数据万维网

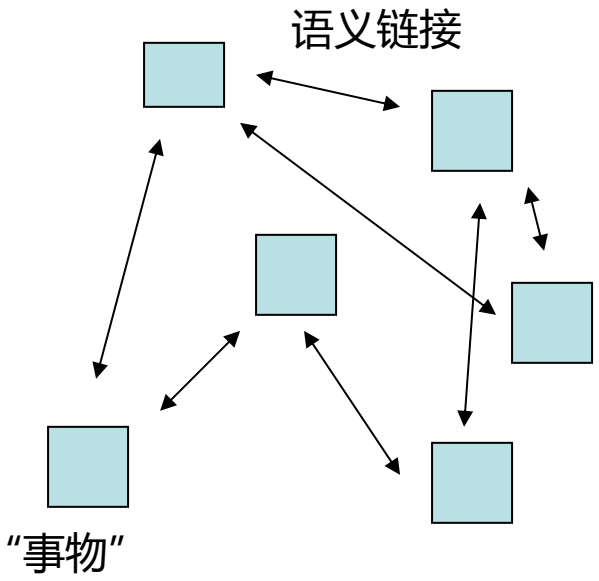


http://www.w3.org/2007/Talks/1211-whit-tbl/#%2828%29

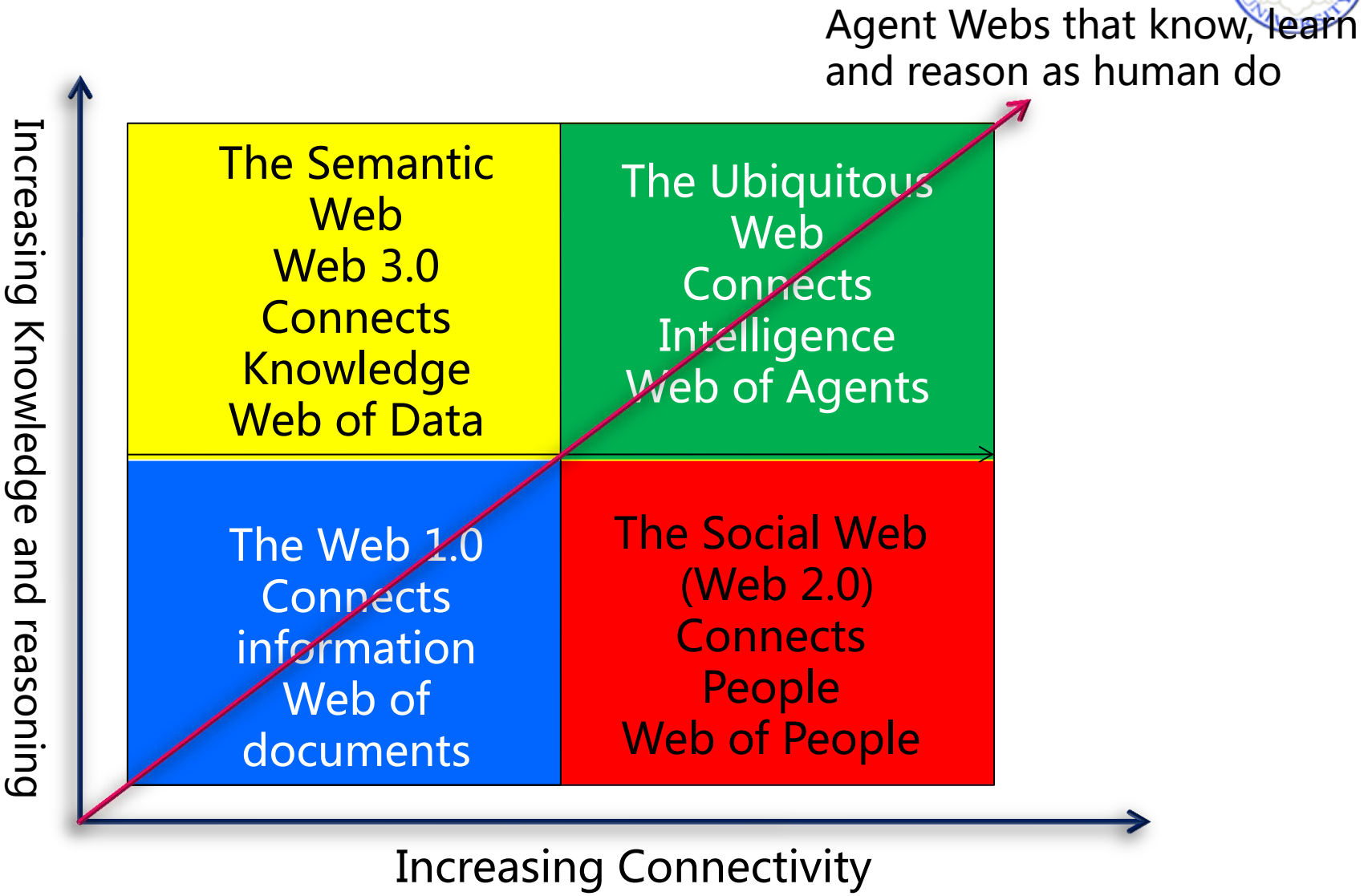
数据万维网

- **特征：**
  - Web上的事物拥有唯一的URI
  - 事物之间由链接关联（如人物、地点、事件、建筑物）
  - 事物之间链接显式存在并拥有类型
  - Web上数据的结构显式存在

- 全球开发的知识共享平台



万维网的发展

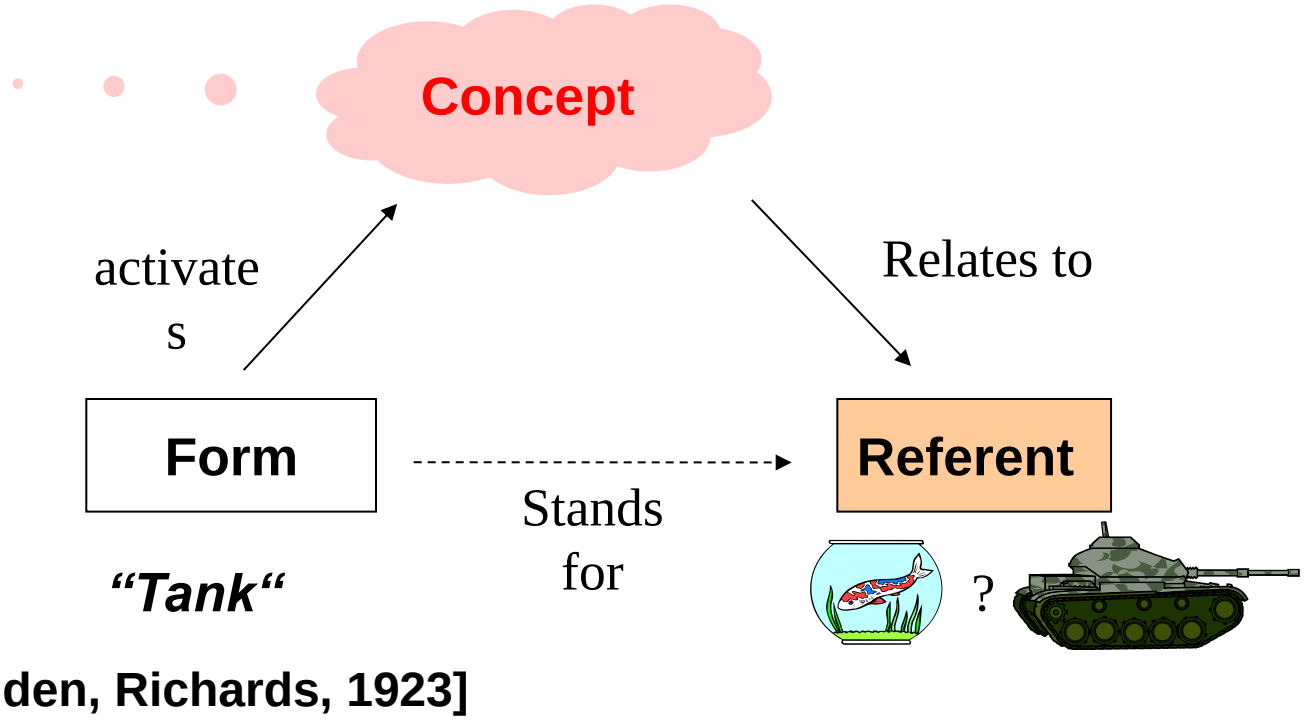

Agent Webs that know, learn and reason as human do
The Semantic Web
Web 3.0
Connects Knowledge
Web of Data
The Ubiquitous Web
Connects Intelligence
Web of Agents
The Web 1.0
Connects information
Web of documents
The Social Web (Web 2.0)
Connects People
Web of People
Increasing Knowledge and reasoning
Increasing Connectivity

Bring structure to the meaningful content of Web pages

The Semantic Web. Tim Berners-Lee, James Hendler, and Ora Lassila. Scientific American, 2001.

哲学中的本体

- 概念三角形



[Ogden, Richards, 1923]

Ontology is the philosophical study of the nature of ***being***, ***becoming***, ***existence***, *or* ***reality***, *as well as the* ***basic categories*** *of being and their relations.* --- Wikipedia

计算机领域本体定义

- **An ontology is a formal, explicit specification of a shared conceptualization** – Gruber 1993

  - Conceptualization: an abstract model of phenomena in the world by having identified the relevant concepts of those phenomena.
  - Explicit: the type of concepts used, and the constraints on their use are explicitly defined.
  - Formal: the fact that the ontology should be machine readable.
  - Shared: ontology should capture consensual knowledge accepted by the communities

本体形式化

- **五元组表示** $O = \{C, R, F, A, I\}$
  - C - *concepts*
    - **概念集合**，通常以Taxonomy形式组织
    - **球星**，**清华校友**
  - R - *relations*
    - **描述概念或者实例之间语义关系的集合**
    - **subClassOf**，**birthplace**
  - F - *functions*
    - 一组特殊的关系，关系中第n个元素的值由其他n-1个元素的值确定
    - Price-of-a-used-car 由 the car-model, manufacturing data 和 kilometers确定

本体的形式化

五元组表示

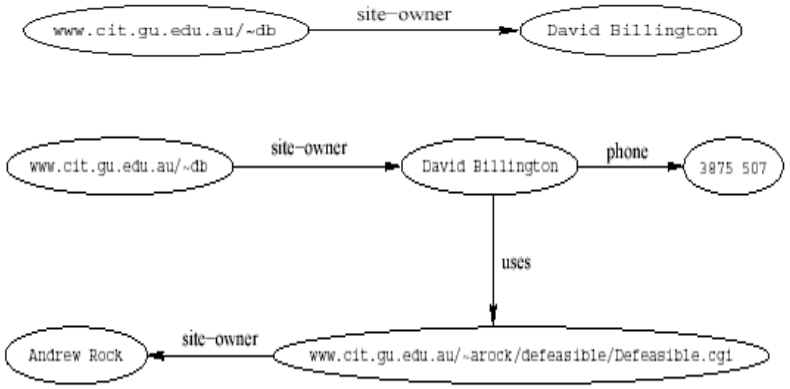
$$O = \{C, R, F, A, I\}$$

- *A - axioms*
  - 公理
  - **如果A是B的子女，B是C的子女，则A是C的子孙**

- *I - instances*
  - 描述具体的**个体**
  - 如：**Peter**是概念**学生**的实例

本体的描述方法

- 资源描述框架 RDF
  - Resource Description Framework



- RDF数据模式
  - 资源 Resource
    - 使用URI唯一标示一个资源
    - 一个资源通常表示一个事物(Thing)
  - 属性 Property
    - 一种特殊类型的资源，用以描述资源与资源见的关系
  - 语句 Statement
    - 由3种资源组成的三元组(Triple)
    - 主语rdf:subject，谓语rdf:predicate以及宾语rdf:object

一个形式化示例

本体的简化形式

$$O = \{C, I, T, P\}$$

- *C – concepts*
  - 描述领域或任务中的**抽象概念**，通常以Taxonomy形式组织
  - 如描述世界知识的本体中，**学生**和**老师**是两个概念
- *I - instances*
  - 描述具体的**实例**
  - **学生Peter**是概念学生的实例
- *T - ISA*
  - **概念与概念之间、实例与概念之间的关系**
  - **subClassOf关系**和**instanceOf关系**

Taxonomy知识

- *P – properties*
  - 本体中用于描述实例信息的**其他语义关系**
  - 如：**instance-attribute-value** (AVP)

AVP知识

What's in freebase? - Light type system

- Topic: one concept or one entity with globally unique ID
- Literal: string, numeric value, Boolean, or timestamp
- Type: properties are grouped into types, an object that is used to semantically group topics
- Property: attribute of a topic
- Schema: Each type has collection of zero or more properties, known as the schema of that type
- Domain: a collection of types which share namespace

本节总结

- 知识图谱实现对客观世界从字符串描述到结构化语义描述，是对客观世界的知识映射（mapping world knowledge）
- 本体可以作为知识图谱表示的概念模型和逻辑基础
- 知识图谱可以描述不同层次和粒度的概念抽象
- 知识图谱可以作为互联网资源组织的基础

虽然语义Web的愿景还尚未发生，知识图谱的发展是让互联网更好的具有世界知识的良好开端

知识图谱类型

一、领域无关知识图谱

- DBPedia, Yago, Freebase, Google KG, etc.

二、特定领域知识图谱

- FOAF, Geonames, Linked Movie Database, etc.

三、跨语言知识图谱

- DBPedia, Yago, Freebase, XLORE, etc.

领域无关知识图谱

类别	名称	其他
人工构建	ResearchCyc	http://www.cyc.com/platform/researchcyc
	WordNet	wordnet.princeton.edu
基于维基百科	DBPedia	dbpedia.org
	YAGO	yago-knowledge.org
	Freebase	freebase.com
	WikiTaxonomy	http://www.h-its.org/english/research/nlp/download/wikitaxonomy.php
	BabelNet	babelnet.org
开放知识抽取	KnowItAll	openie.cs.washington.edu
	NELL	rtw.ml.cmu.edu
	Probase	http://research.microsoft.com/en-us/projects/probase/
中文知识图谱	百度知心	www.baidu.com
	搜狗知立方	www.sogou.com

领域无关知识图谱

name	*# of concepts*	*# of isA pairs*
Freebase	1,450	24,483,434
WordNet	25,229	283,070
WikiTaxonomy	111,654	105,418
YAGO	352,297	8,277,227
DBPedia	259	1,900,000
ResearchCyc	≈ 120,000	< 5,000,000
KnowItAll	N/A	< 54,753
TextRunner	N/A	< 11,000,000
OMCS	173,398	1,030,619
NELL	123	< 242,453
Probase	**2,653,872**	**20,757,545**

概念数量对比 2010.12

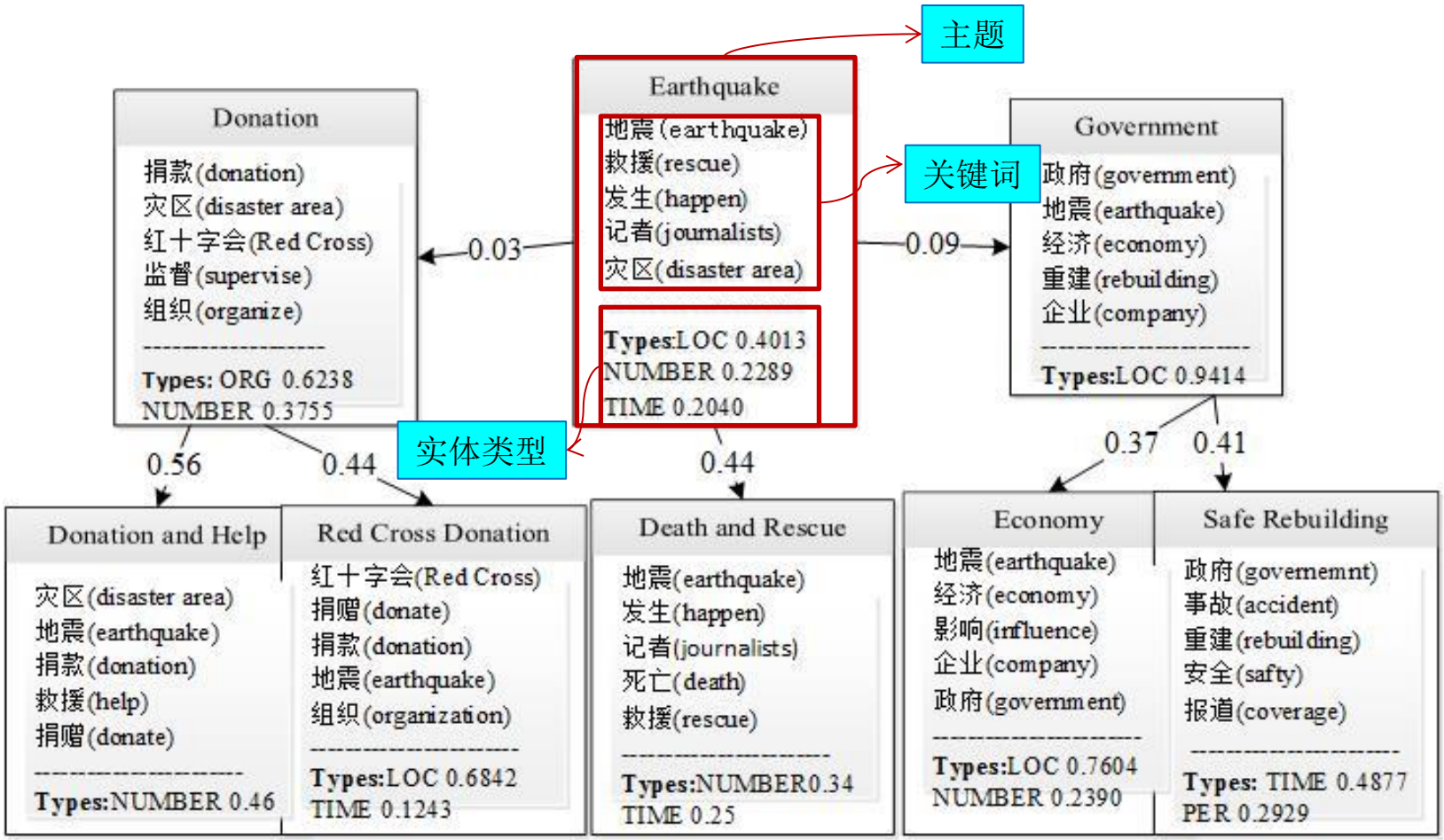
http://research.microsoft.com/en-us/projects/probase/

事件知识图谱

灾难> 自然灾害> 地震： 2014年鲁甸地震事件

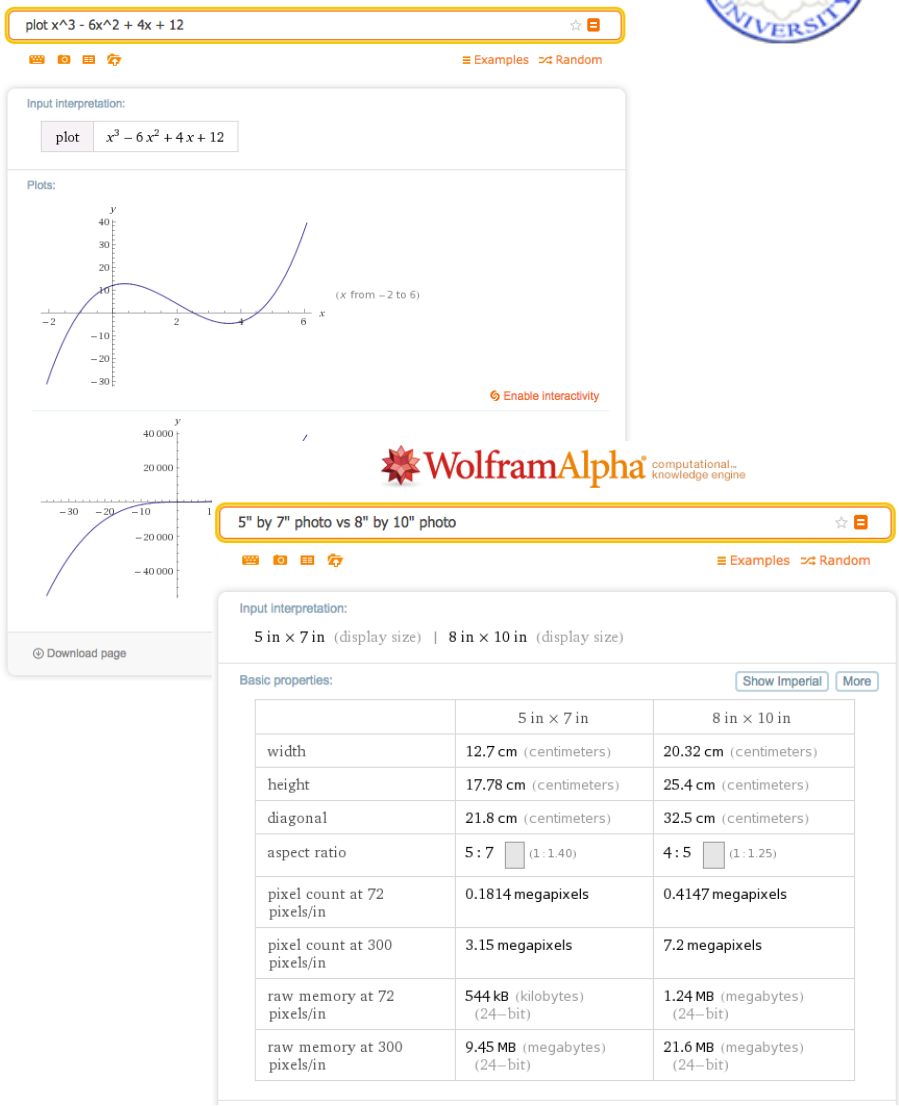
事件知识图谱

- 事件学习：从多个相似事件实例中学习层次主题模式



计算知识图谱

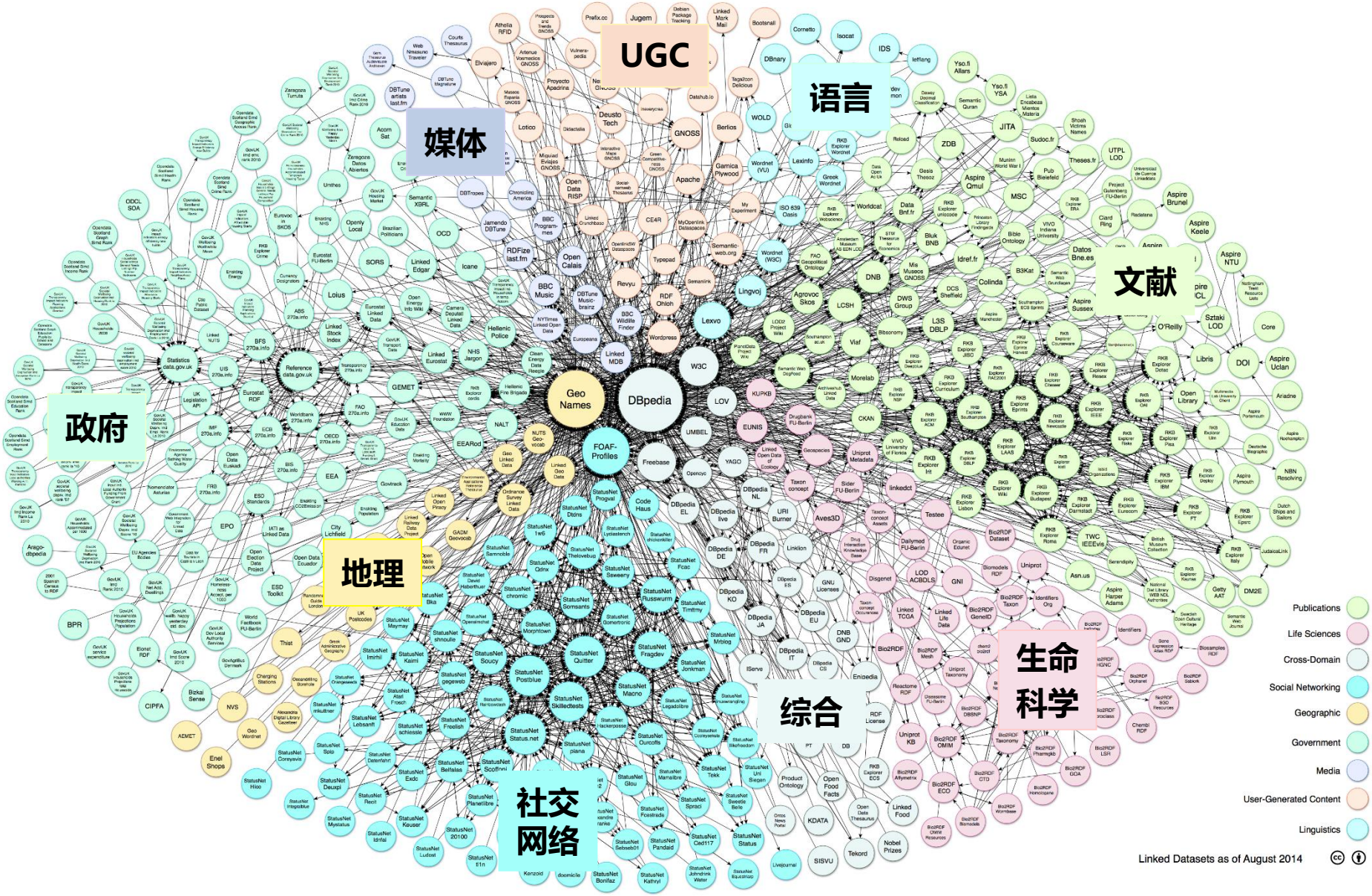
- 计算知识引擎WolframAlpha



	5 in × 7 in	8 in × 10 in
width	12.7 cm (centimeters)	20.32 cm (centimeters)
height	17.78 cm (centimeters)	25.4 cm (centimeters)
diagonal	21.8 cm (centimeters)	32.5 cm (centimeters)
aspect ratio	5 : 7 (1 : 1.40)	4 : 5 (1 : 1.25)
pixel count at 72 pixels/in	0.1814 megapixels	0.4147 megapixels
pixel count at 300 pixels/in	3.15 megapixels	7.2 megapixels
raw memory at 72 pixels/in	544 kB (kilobytes) (24–bit)	1.24 MB (megabytes) (24–bit)
raw memory at 300 pixels/in	9.45 MB (megabytes) (24–bit)	21.6 MB (megabytes) (24–bit)

http://www.wolframalpha.com/

特定领域知识图谱


UGC
语言
媒体
文献
政府
地理
生命科学
综合
社交网络
Geo Names
DBpedia
FOAF-Profiles
Publications
Life Sciences
Cross-Domain
Social Networking
Geographic
Government
Media
User-Generated Content
Linguistics
Linked Datasets as of August 2014

影视领域本体

多源影视知识

维基百科
自由的百科全书

后会无期 *The Continent*	
基本资料	
导演	韩　寒
监制	方励
编剧	韩　寒
主演	冯绍峰、陈柏霖、钟汉良、陈乔恩、王珞丹、袁　泉[1]
配乐作曲	小林武史
摄影	廖拟
剪辑	肖洋
片长	106 分钟
制片商	北京劳雷影业有限公司 杭州果麦文化传媒有限公司 博纳影业集团
产地	中国大陆
语言	现代标准汉语
上映及发行	
上映日期	中国大陆2014年07月24日 台湾2014年9月19日[1] 香港：2014年10月16日
发行商	天津博纳文化传媒有限公司 华夏电影发行责任公司

属性数据补充

链接补充

Baidu百科

中文名	后会无期
外文名	The Continent
导演	韩寒
制片人	方励
编剧	韩寒
主演	冯绍峰、陈柏霖、钟汉良、陈乔恩、王珞丹、袁泉
出品时间	2014年
制片地区	中国
出品公司	劳雷影业、果麦文化、博纳影业
片长	106 分钟
对白语言	普通话
上映时间	2014年07月24日
发行公司	天津博纳文化 华夏电影发行
类　型	喜剧，爱情，冒险
色　彩	彩色
制片成本	5000万人民币
拍摄地点	上海，四川西昌，内蒙古赤峰，浙江舟山普陀，东极岛
拍摄日期	2014年2月14日

影视领域知识图谱

影视知识融合

相关实体:

isA关系验证:

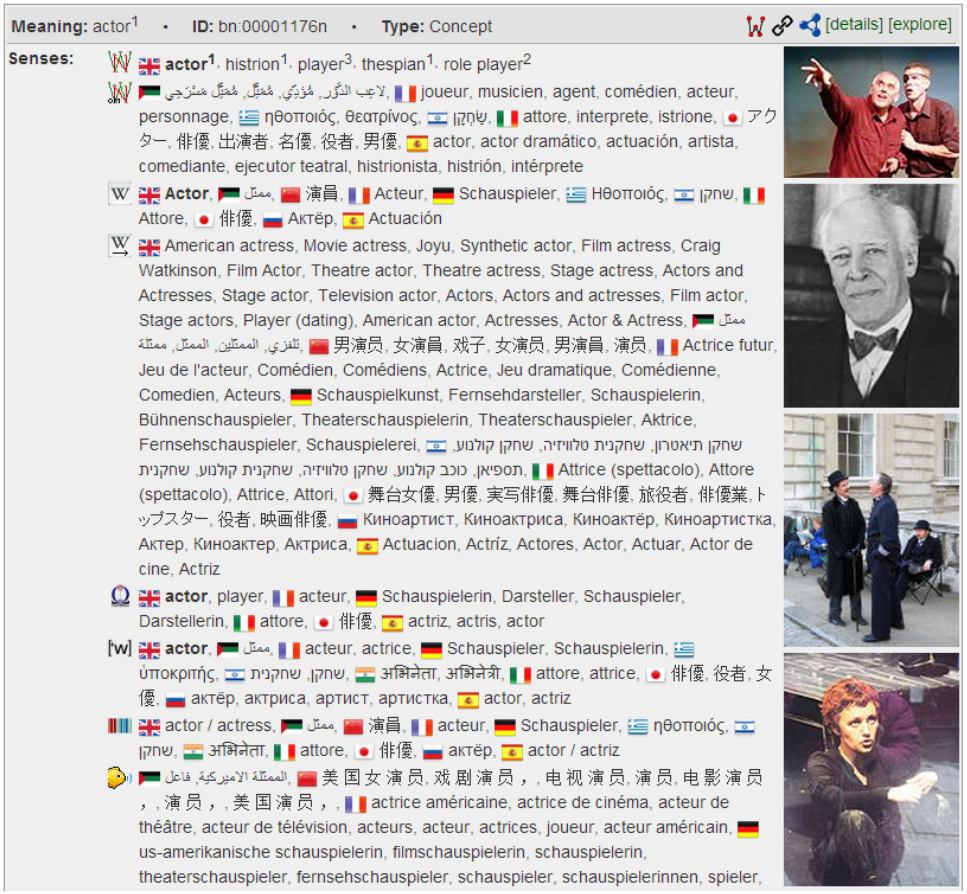
食人女	isA
Baidu百科	Film or TV?
Freebase	Film
豆瓣电影	Film

双语文字对齐:

movie:directed_by
"***Han Han***"@en,
"***韩寒***"@zh .
movie:genres
"***Comedy***"@en,
"***喜剧***"@zh .
movie:summary
"***The Continent is directed and written by Han Han…"@en,***
"***《后会无期》是一部由韩寒担任编剧及导演…***"@zh .

跨语言知识图谱

BabelNet 1.1.1
A very large multilingual ontology



演员/Actor

http://babelnet.org/

DBpedia

yago select knowledge

Freebase

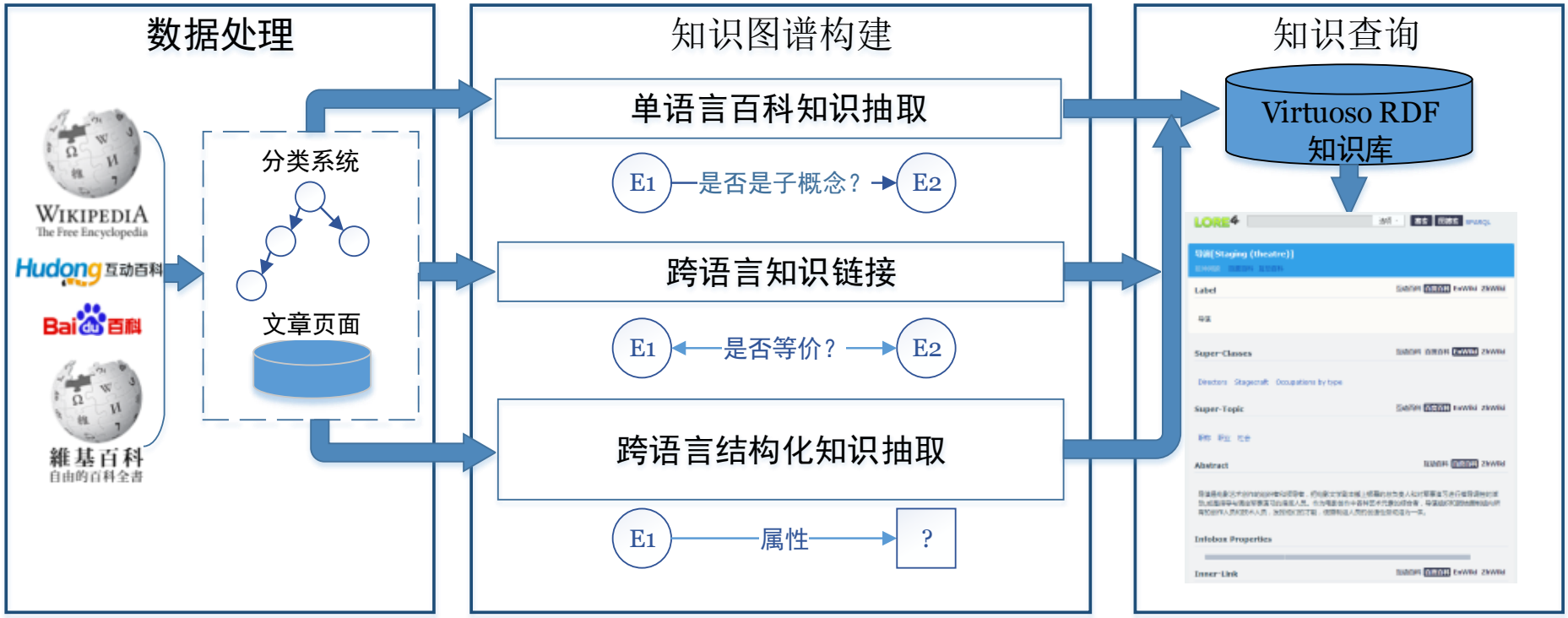
Google KG

EuroWordNet

XLORE

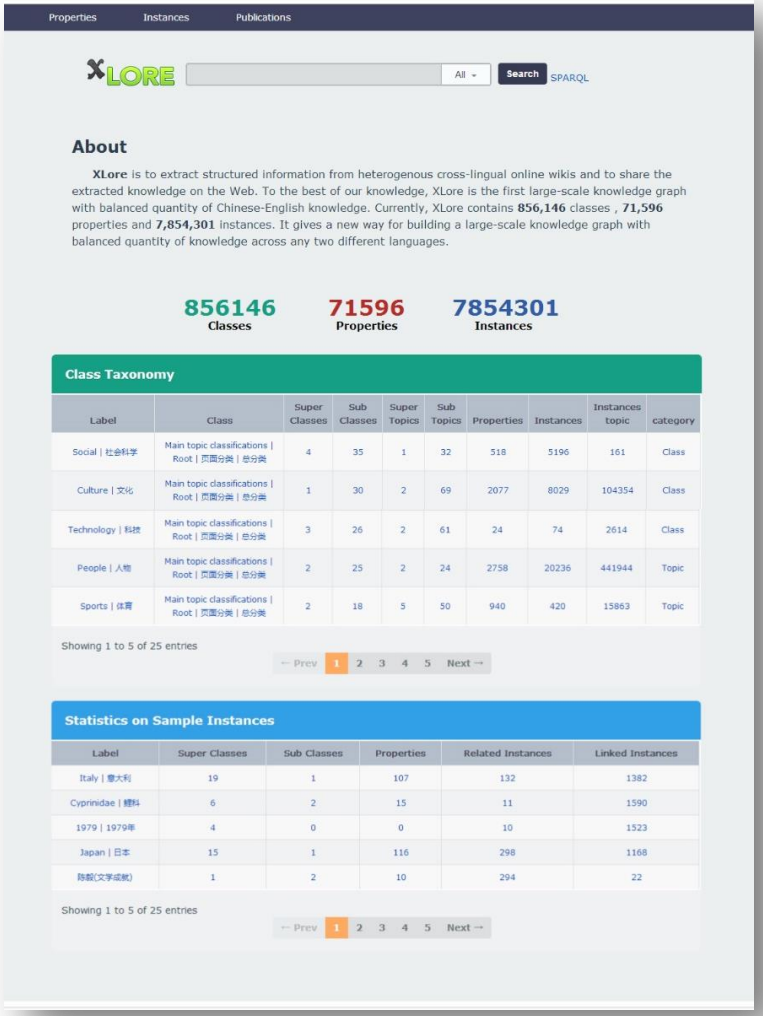
跨语言知识图谱XLORE

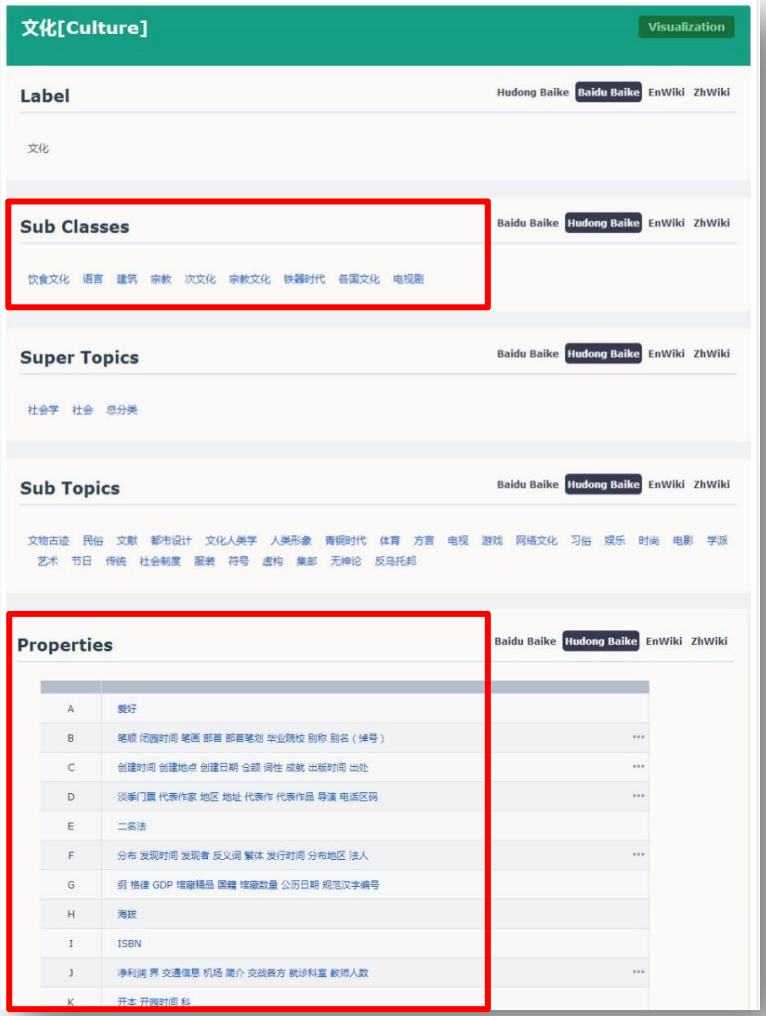
- 基于分布的在线异构百科资源，通过跨语言知识链接技术，构建一个中英文知识量比较平衡的大规模跨语言知识图谱



XLORE: 集成百度百科、互动百科、中文维基和英文维基，包含856,146个概念，71,596个属性，7,854,301个实例。

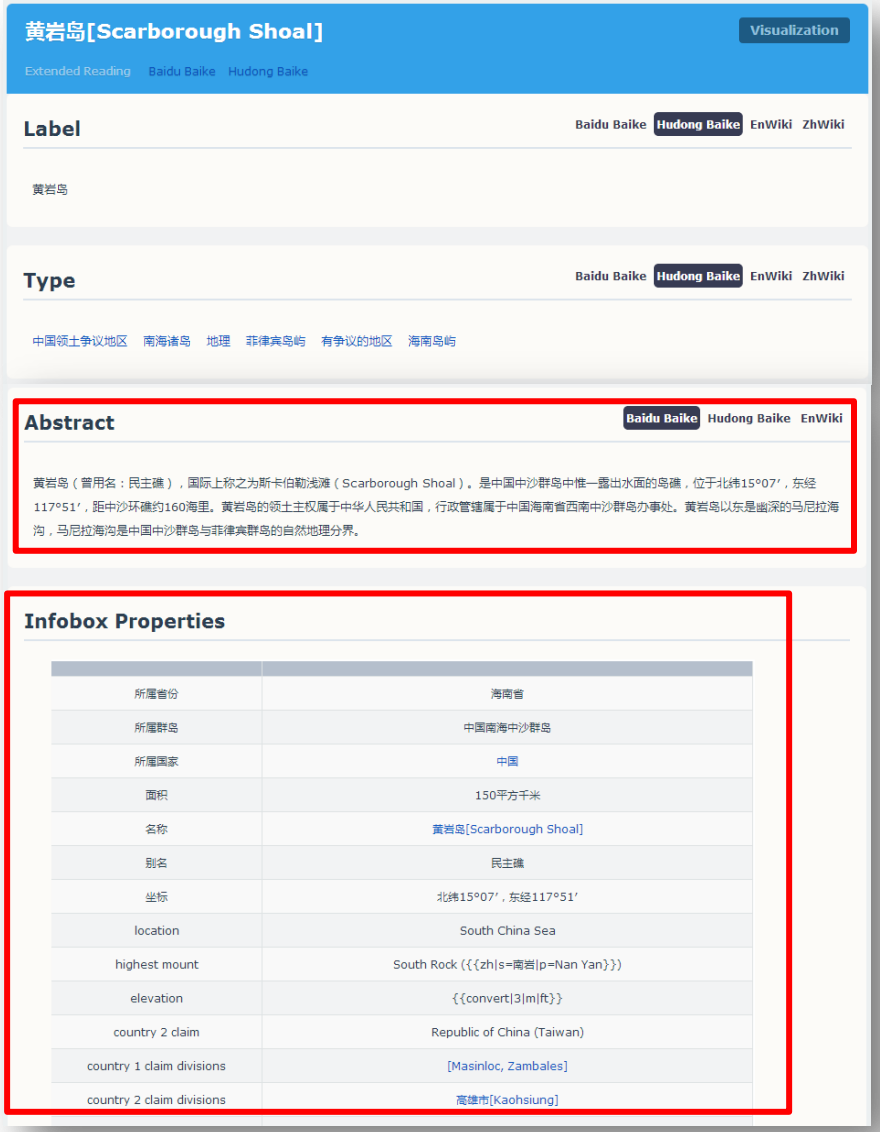
跨语言知识图谱XLORE

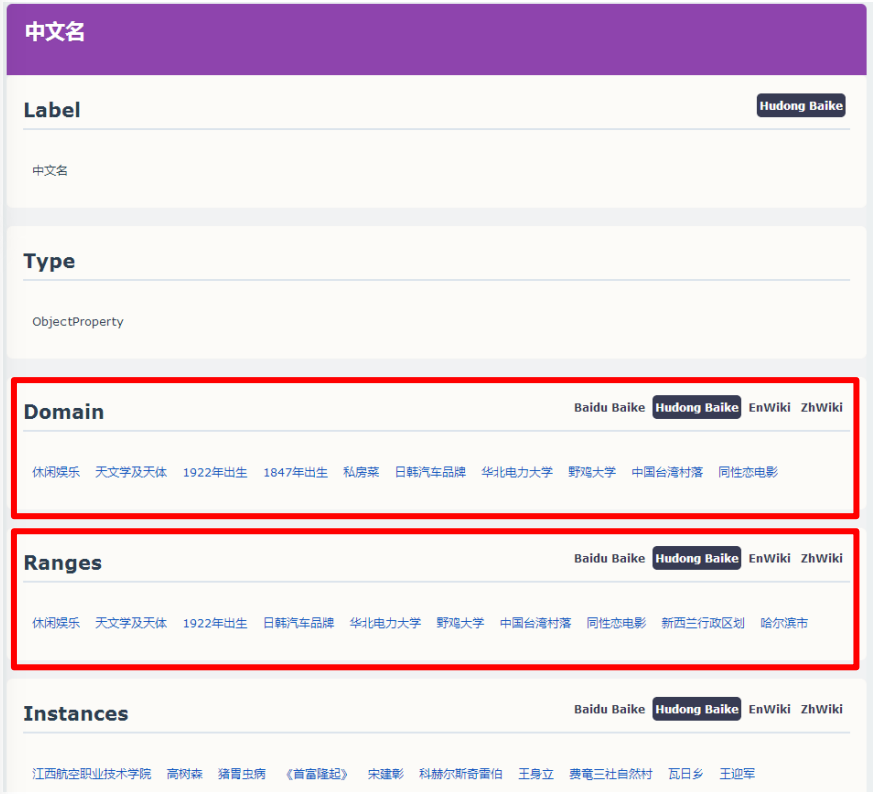




http://xlore.org/

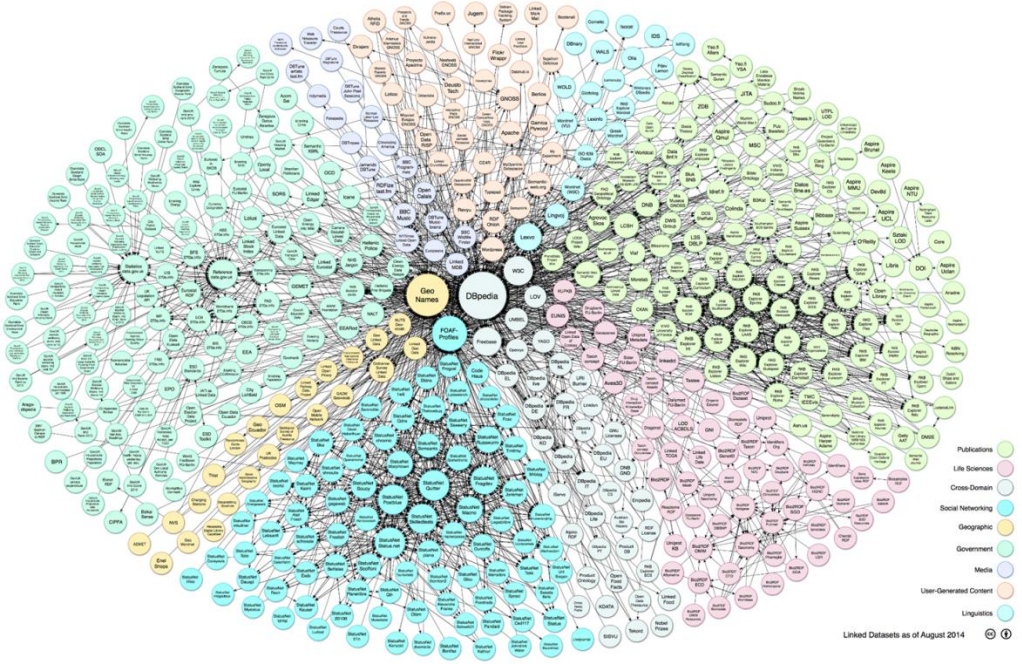
跨语言知识图谱XLORE





本节总结

- **知识图谱是对处理数据的结构化结果表示**
- **知识图谱可以表达：**
  - 实体及其关系知识
  - 事件知识
  - 计算知识
  - 限定领域知识
  - 面向特定任务的知识图谱
  - 跨语言知识
  - ……



  - 知识图谱是实现语义互操作的基础

知识图谱构建关键技术

一、基于wiki百科资源的知识图谱构建

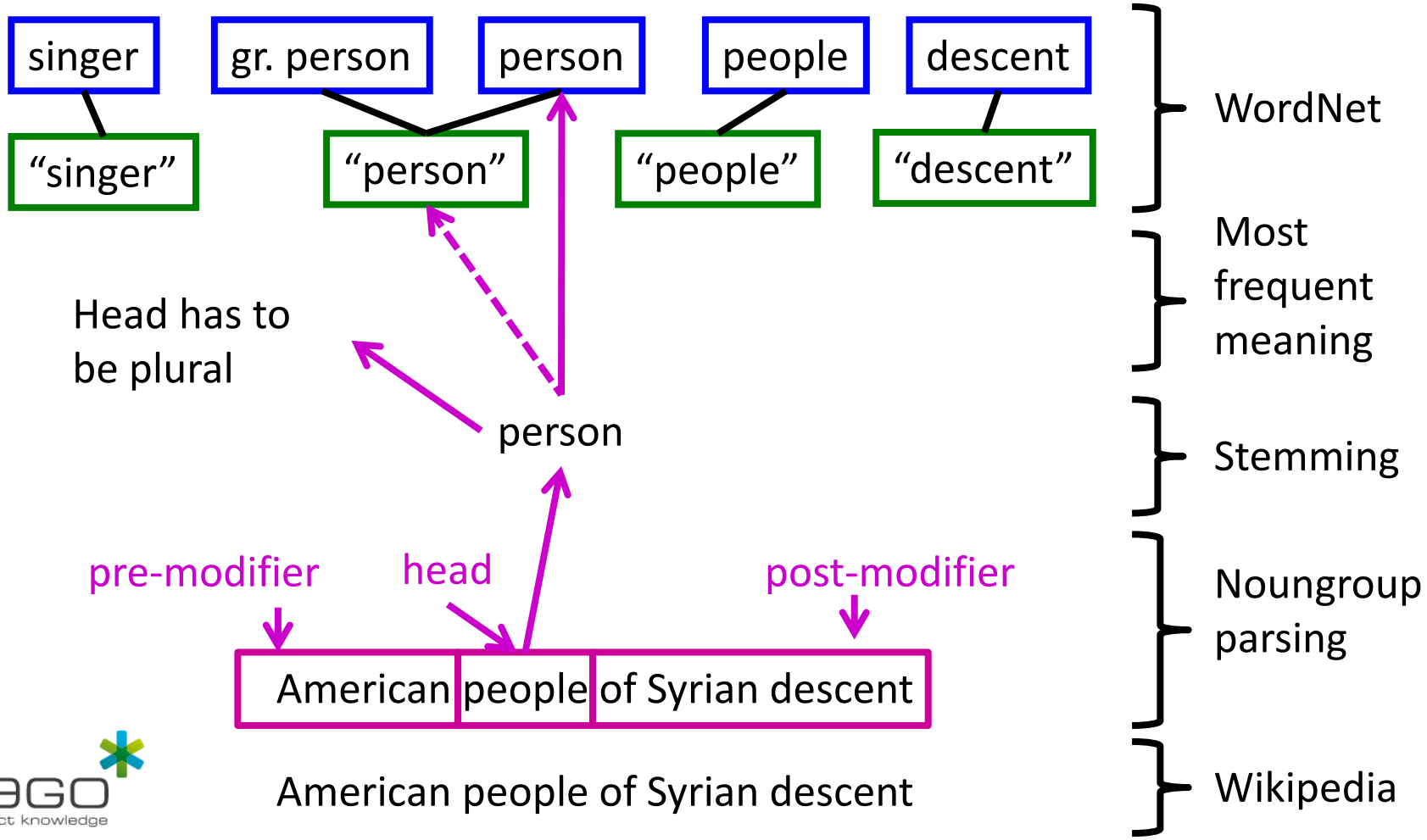
- Taxonomy知识抽取，AVP知识抽取

二、Beyond wiki百科资源的知识图谱构建

- 结构化数据，半结构化数据，非结构化数据

基于Wiki资源的Taxonomy知识抽取

- 维基百科分类映射到WordNet



Yago: a core of semantic knowledge. Suchanek et al. WWW 07.

基于Wiki资源的Taxonomy知识抽取

- 识别维基百科中正确的isA关系

中心词匹配

British Computer **Scientists** → Computer **Scientists** √

修饰成分分析

Crime Comics → Crime ×

Islamic Mysticism → Islam ×

中心词单复数形式

Albert Einstein → Naturalized **Citizens** of US √

WikiTaxonomy

Deriving a Large Scale Taxonomy from Wikipedia. Ponzetto et al. AAAI 07.

基于Wiki资源的Taxonomy知识抽取

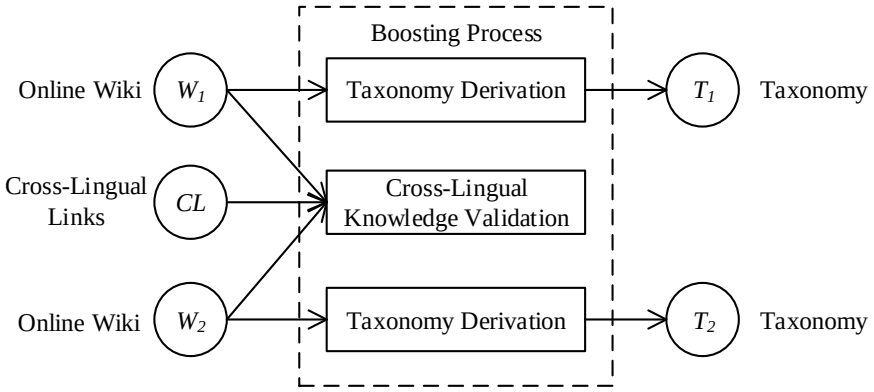
- 基于跨语言知识校验的isA关系识别

中心词单数

Educational People → **Education** ✗

前缀子串

教育人物 → 教育 ✗



- 文本特征
  - 中心词关系
  - 单复数形式
  - 前后缀关系
- 结构特征
  - Normalized Google Distance

XLORE

Cross-lingual Knowledge Validation Based Taxonomy Derivation from Heterogeneous Online Wikis. Wang et al. AAAI 14.

基于Wiki资源的AVP知识抽取

- 信息框(Infobox)抽取

WIKIPEDIA
The Free Encyclopedia

Barack Obama

Obama in front of the *Resolute* desk in the Oval Office of the White House on December 6, 2012

44th President of the United States	
Incumbent	
Assumed office	January 20, 2009
Vice President	Joe Biden
Preceded by	George W. Bush
United States Senator from Illinois	
In office	January 3, 2005 – November 16, 2008
Preceded by	Peter Fitzgerald
Succeeded by	Roland Burris
Member of the Illinois Senate from the 13th District	
In office	January 8, 1997 – November 4, 2004
Preceded by	Alice Palmer
Succeeded by	Kwame Raoul
Personal details	
Born	Barack Hussein Obama II August 4, 1961 (age 53) Honolulu, Hawaii, U.S.
Nationality	American
Political party	Democratic

dbpedia-owl:activeYearsEndDate	2004-11-04 (xsd:date) 2008-11-16 (xsd:date)
dbpedia-owl:activeYearsStartDate	1997-01-08 (xsd:date) 2005-01-03 (xsd:date) 2009-01-20 (xsd:date)
dbpedia-owl:almaMater	dbpedia:Columbia_University dbpedia:Occidental_College dbpedia:Harvard_Law_School
dbpedia-owl:birthDate	1961-08-04 (xsd:date)
dbpedia-owl:birthPlace	dbpedia:Hawaii dbpedia:Honolulu
dbpedia-owl:birthYear	1961-01-01 (xsd:date)
dbpedia-owl:bnfId	15591663c
dbpedia-owl:individualisedGnd	132522136
dbpedia-owl:lccnId	n/94/112934
dbpedia-owl:office	President of the United States Member of the Illinois Senate from the 13th District
dbpedia-owl:orderInOffice	44th
dbpedia-owl:party	dbpedia:Democratic_Party_(United_States)
dbpedia-owl:profession	dbpedia:Author dbpedia:Lawyer dbpedia:Community_organizing dbpedia:Professor
dbpedia-owl:region	dbpedia:Illinois
dbpedia-owl:relation	dbpedia:Maya_Soetoro-Ng dbpedia:Madelyn_Dunham dbpedia:Stanley_Armour_Dunham
dbpedia-owl:religion	dbpedia:Christianity
dbpedia-owl:residence	dbpedia:White_House dbpedia:Chicago
dbpedia-owl:selibrId	314463
dbpedia-owl:seniority	United States Senate
dbpedia-owl:successor	dbpedia:Kwame_Raoul dbpedia:Roland_Burris
dbpedia-owl:termPeriod	dbpedia:Barack_Obama__1 dbpedia:Barack_Obama__2

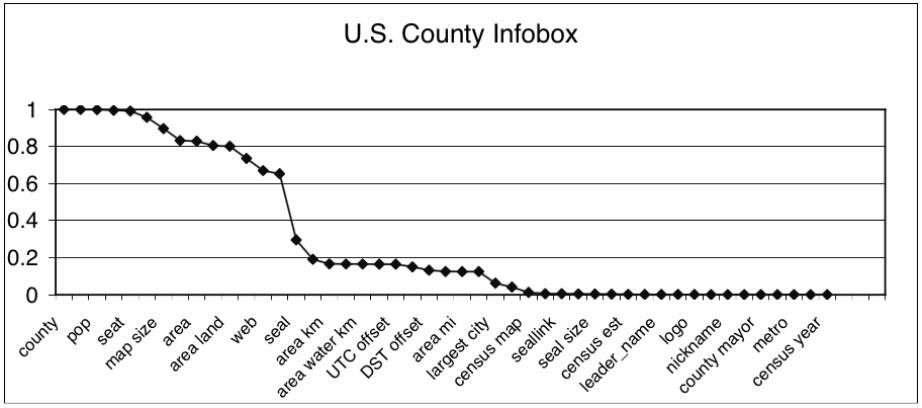
DBpedia

http://dbpedia.org/page/Barack_Obama

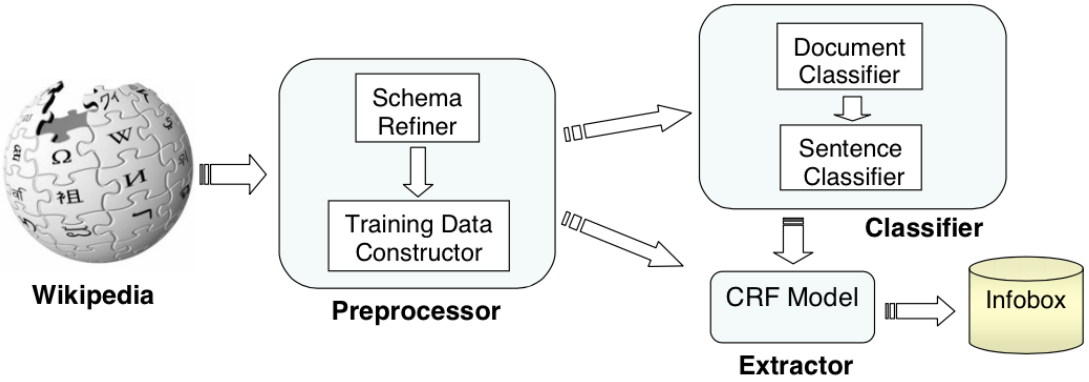
http://en.wikipedia.org/wiki/Barack_Obama

基于Wiki资源的AVP知识抽取

缺失信息框抽取



信息框“US County”属性使用比率



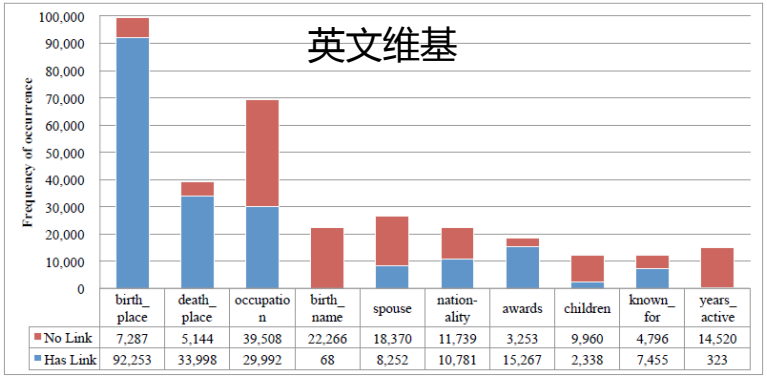
Feature Description	Example
First token of sentence	*Hello* world
In first half of sentence	*Hello* world
In second half of sentence	Hello *world*
Start with capital	Hawaii
Start with capital, end with period	Mr.
Single capital	A
All capital, end with period	CORP.
Contains at least one digit	AB3
Made up of two digits	99
Made up of four digits	1999
Contains a dollar sign	20$
Contains an underline symbol	km_square
Contains an percentage symbol	20%
Stop word	the; a; of
Purely numeric	1929
Number type	1932; 1,234; 5.6
Part of Speech tag	
Token itself	
NP chunking tag	
String normalization: capital to “A”, lowercase to “a”, digit to “1”, others to “0”	$TF-1 \Longrightarrow AA01$
Part of anchor text	Machine *Learning*
Beginning of anchor text	*Machine* Learning
Previous tokens (window size 5)	
Following tokens (window size 5)	
Previous token anchored	Machine *Learning*
Next token anchored	*Machine* Learning

CRF特征

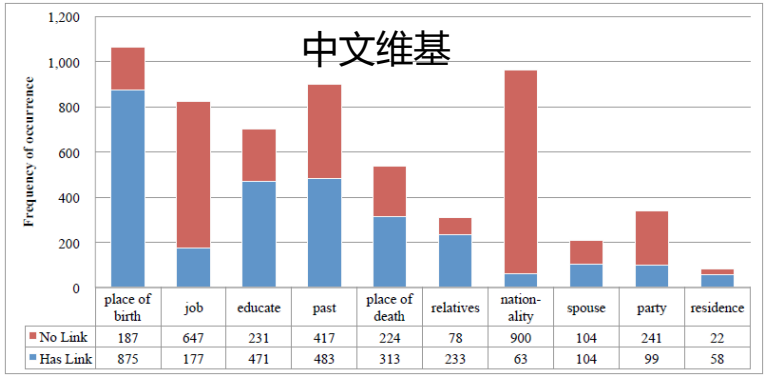
Autonomously Semantifying Wikipedia. Wu et al. CIKM 07.

基于Wiki资源的AVP知识抽取

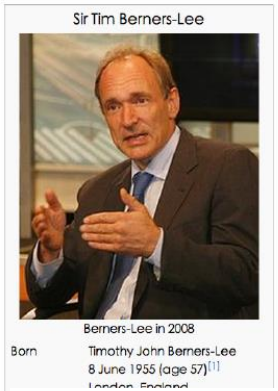
- 信息框值中链接缺失



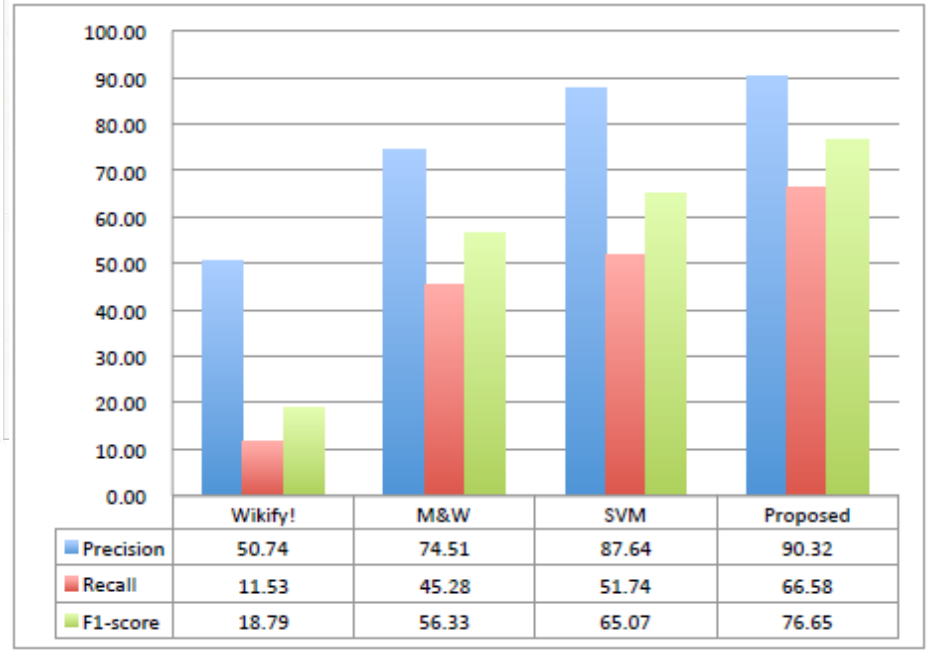
	birth_ place	death_ place	occupation	birth_ name	spouse	nation- ality	awards	children	known_ for	years_ active
No Link	7,287	5,144	39,508	22,266	18,370	11,739	3,253	9,960	4,796	14,520
Has Link	92,253	33,998	29,992	68	8,252	10,781	15,267	2,338	7,455	323



	place of birth	job	educate	past	place of death	relatives	nation- ality	spouse	party	residence
No Link	187	647	231	417	224	78	900	104	241	22
Has Link	875	177	471	483	313	233	63	104	99	58




```
{{Infobox person
| image          = Tim Berners-Lee-Knight-crop.jpg
| caption        = Berners-Lee in 2008
| birth_name     = Timothy John Berners-Lee
| birth_date     = {{birth date and age|1955|6|8|df=y}}
| birth_place    =London, England<br>United Kingdom
| nationality    = British
| residence      = United States and United Kingdom
| occupation     = [[Computer scientist]]
| employer       = {{Plainlist|
                  * [[World Wide Web Consortium]]
                  * [[University of Southampton]]
```



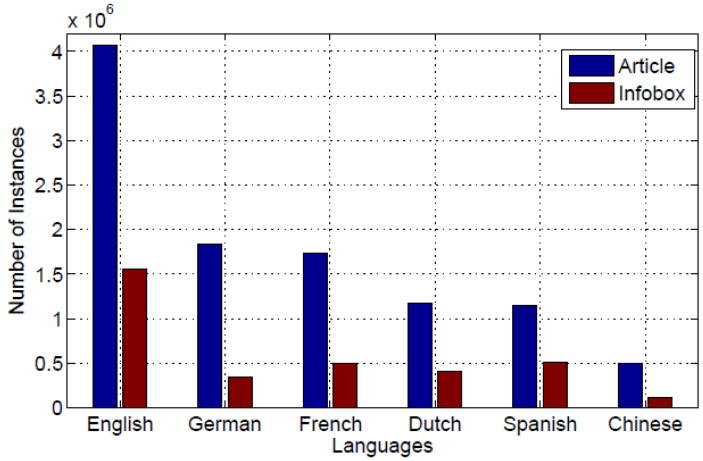
	Wikify!	M&W	SVM	Proposed
Precision	50.74	74.51	87.64	90.32
Recall	11.53	45.28	51.74	66.58
F1-score	18.79	56.33	65.07	76.65

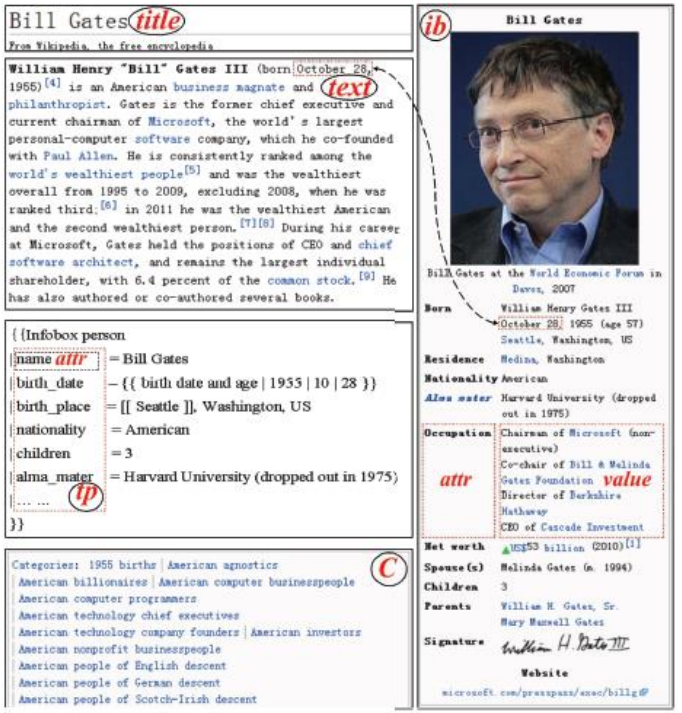
基于回归学习的实体链接方法

Discovering Missing Semantic Relations between Entities in Wikipedia. Xu et al. ISWC 13.

基于Wiki资源的AVP知识抽取

- **基于迁移学习的跨语言属性值抽取**
  - 信息框大量缺失
  - 不同语言下差异较大
- **现有方法**
  - 基于翻译的方法
  - 单语言信息抽取方法



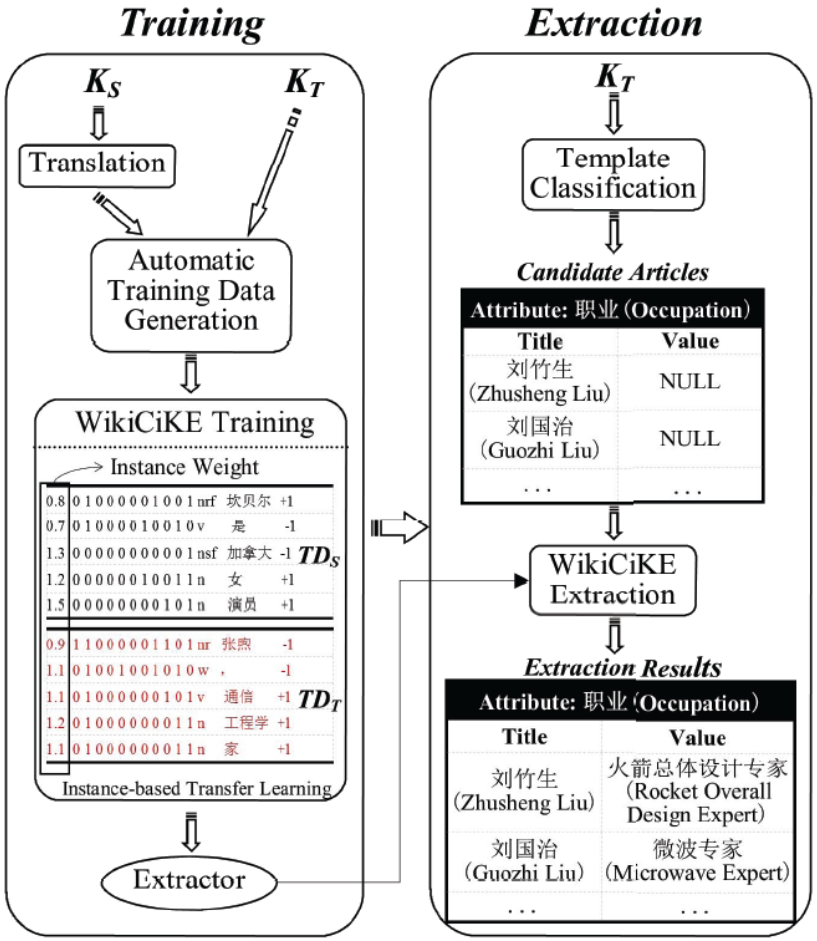


Attribute	*en*	*zh*	Attribute	*en*	*zh*
name	82,099	1,486	*awards*	2,310	38
birth date	77,850	1,481	*weight*	480	12
occupation	66,768	1,279	*influences*	450	6
nationality	20,048	730	*style*	127	1

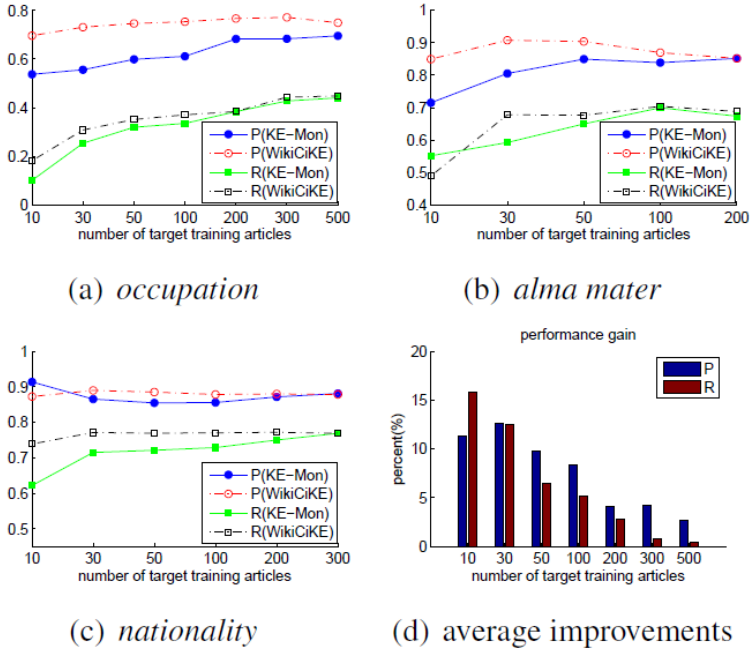
能否利用丰富的英文知识帮助自动化抽取缺失的中文知识？

基于Wiki资源的AVP知识抽取

基于迁移学习的跨语言属性值抽取



与单语言抽取方法对比



(a) *occupation* (b) *alma mater*

(c) *nationality* (d) average improvements

与翻译方法对比

Attribute	KE-Tr		WikiCiKE	
	P	*R*	*P*	*R*
occupation	27.4%	3.40%	64.8%	26.4%
nationality	66.3%	4.60%	70.0%	**55.0%**
alma mater	66.7%	0.70%	**76.3%**	8.20%

Transfer Learning Based Cross-lingual Knowledge Extraction for Wikipedia. Wang et al. ACL 13.

结构化数据转化为语义资源

- **结构化数据**
  - 大部分结构化数据都被存储在关系型数据库中。
- **将结构化数据转化为知识的RDF描述—D2R**
  - D2R是一种XML-based 语言，用来达到上面称述的映射目标。
  - D2R 映射步骤
    - 从关系型数据库中选取一个或者一组相似的类
    - 把选取中的记录按列分组
    - 为每个类下的实例进行URI或者blank node分配
    - 为每个instance 创建属性

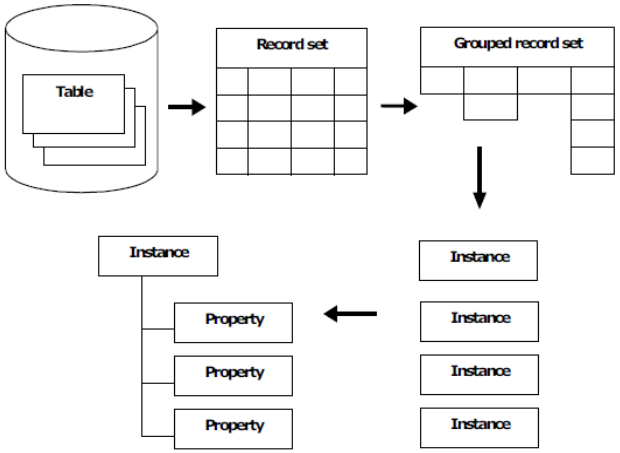


Figure 1. The D2R mapping process.

D2R映射过程

D2R MAP - A Database to RDF Mapping Language. WWW (Posters) 2003

半结构化知识抽取

Taxonomy知识抽取

从Web Table中提取实例与概念之间的上下位语义关系

半结构化知识抽取

Taxonomy知识抽取

1. 给出“种子（seeds）”作为搜索的起始。
   cities = {Paris, Shanghai, Brisbane}
2. 搜索包含一个或多个“种子”的表格

Paris	France
Shanghai	China
Berlin	Germany
London	UK

Paris	Iliad
Helena	Iliad
Odysseus	Odysee
Rama	Mahabaratha

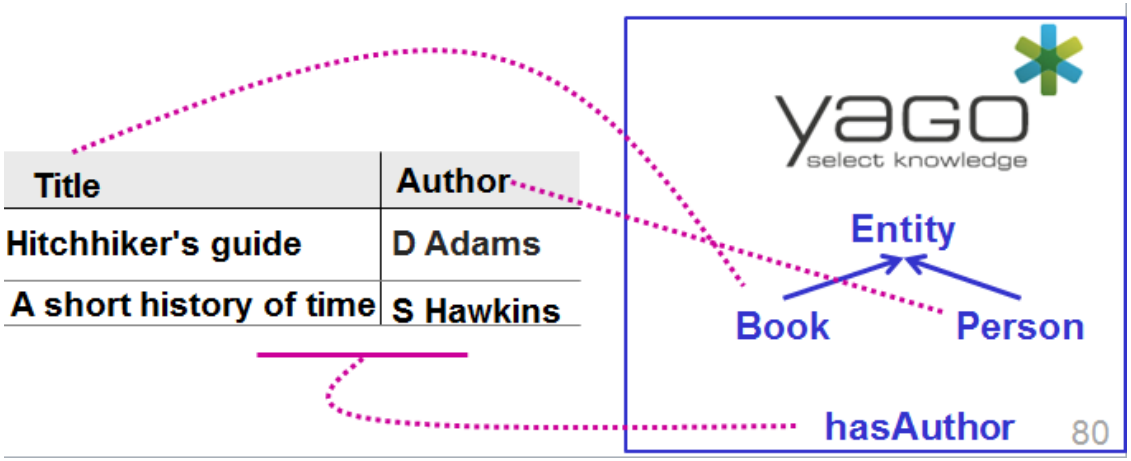
3. 从表格中抽取概念-实例关系

A semi-supervised method to learn and construct taxonomies using the web. Kozareva et al. EMNLP 10.

半结构化知识抽取

AVP知识抽取

- 目标：建立Web Table的实体关系链接，使在Web Table上进行语义搜索成为可能
- 思路：利用YAGO知识库对网页表格进行标注
  - 将列标题映射到YAGO类
  - 将单元格的值映射到YAGO实体
  - 利用因子图模型做AVP知识的联合计算



Annotating and Searching Web Tables Using Entities, Types and Relationships. Limaye et al. PVLDB 10.

非结构化资源的知识学习

- **Open Information Extraction**
  - 学习一般性模型来表示关系表示
  - 学习领域相关正则表达式

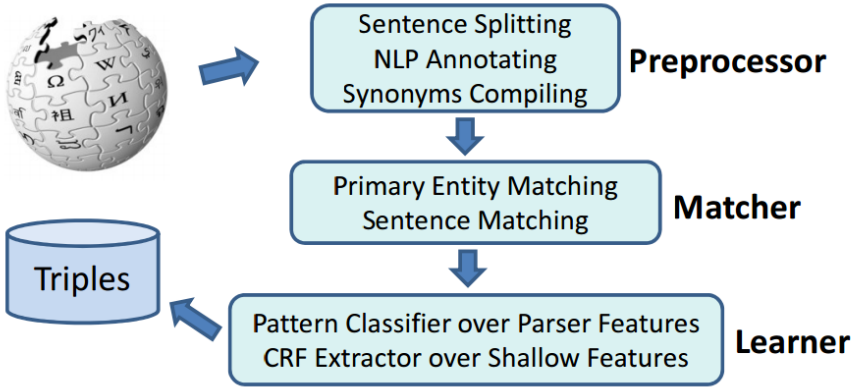
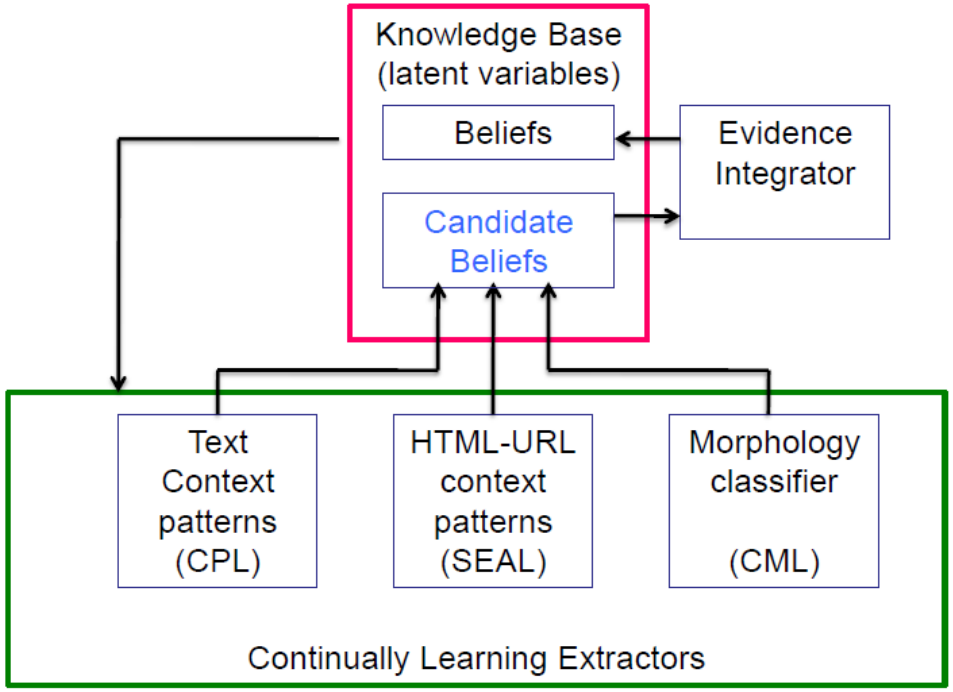


Figure 1: Architecture of WOE.

Open information extraction using Wikipedia . ACL 10

非结构化资源的知识学习

- **NELL (Never Ending Language Learning)**
  - 大规模信息抽取系统
  - 500-600个概念和关系
  - 3.2M的“低可信度”fact, 500K高可信度的fact



NELL系统图

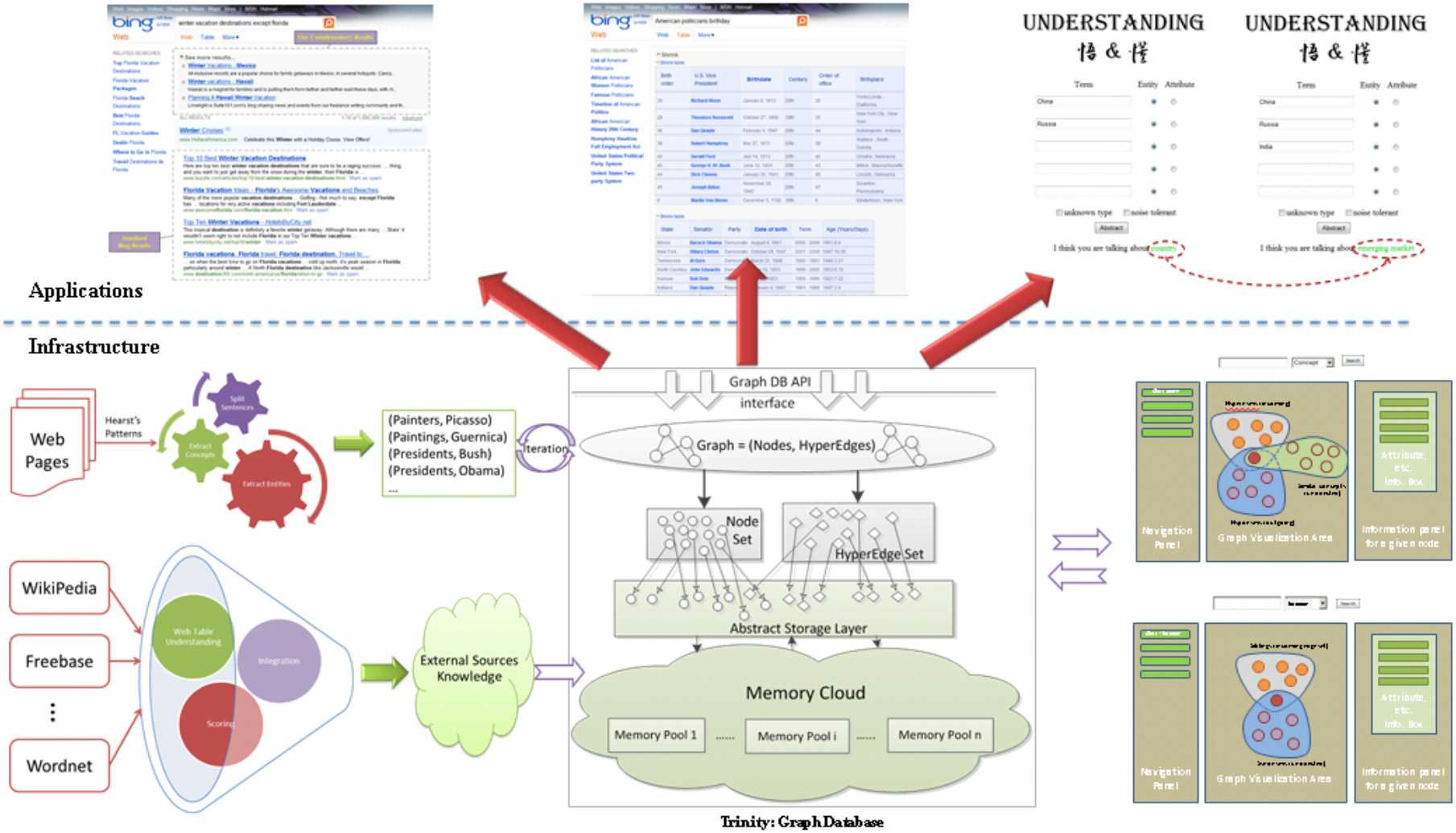
Toward an Architecture for Never-Ending Language Learning. AAAI 2010

非结构化资源的知识学习

- **Probase (A Probabilistic Knowledgebase)**
  - 目标
    - 通过注入“一般性知识”到计算机中，来更加了解人的交流，形成知识库
  - 建立Probase
    - 先用迭代的方法建立核心的taxonomy
    - 找到哪些属性和哪些类可以用来回答哪些问题，比如（中国，人口，14亿）可以用来回答“中国居住了多少人？”，虽然“人口”没有出现在问题中
    - 用一个非监督的bootstrapping 算法反复扫描网页文档集合来得到很多instance的关系
    - 用一种概率的数据集成机制来融合目前已有的结构化数据，例如Freebase, IMDB, Amazon

Probase: a probabilistic taxonomy for text understanding. SIGMOD 2012

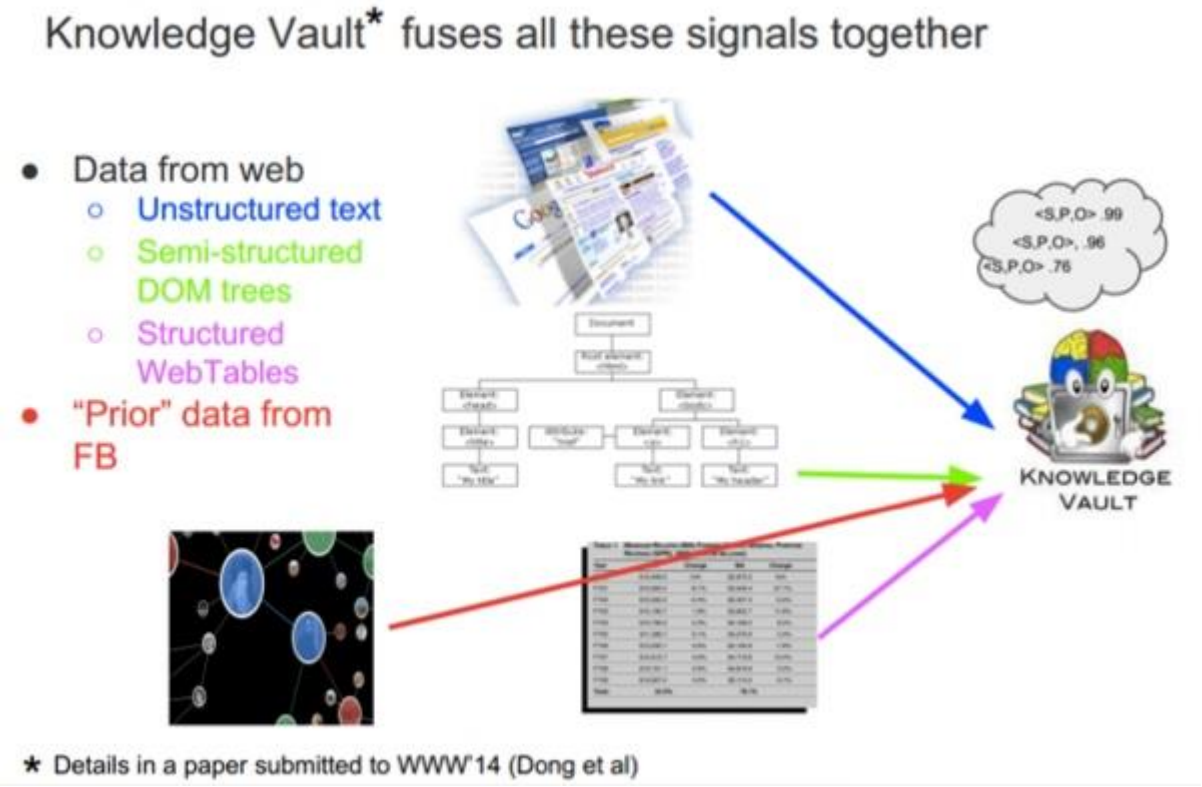
非结构化资源的知识学习



Probase系统图

多资源知识融合

Google Knowledge Vault



- Extractors
  - 三元组抽取
- Graph-based Priors
  - 三元组先验概率学习
- Knowledge Fusion
  - 三元组正确性预测

Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. Dong et al. KDD 14.

本节总结

- **互联网海量数据中富含各类知识**
- **对于不同类型的资源有不同的知识图谱获取方法**
- **知识获取仍然面临很多挑战性问题**
  - 增量式知识获取
  - 长尾知识获取
  - 知识的演化
  - 高精确度的知识获取
  - 开放taxonomy 知识的建立
  - 异构知识的处理
  - 利用群体智慧的知识获取
  - 推理规则的学习
  - ……

基于知识图谱的语义链接及其应用

- 基于知识图谱的语义标注

Barack Obama

44th U.S. President

Barack Hussein Obama II is the 44th and current President of the United States, and the first African American to hold the office. Wikipedia

Born: August 4, 1961 (age 53), Honolulu, Hawaii, United States

Spouse: Michelle Obama (m. 1992)

Office: President of the United States since 2009

Presidential term: January 20, 2009 –

Parents: Ann Dunham, Barack Obama, Sr.

Siblings: Maya Soetoro-Ng, Mark Okoth Obama Ndesandjo, more

Recent posts on Google+

Barack Obama
4,769,268 followers • Shared publicly

"It's long past time for us to raise the minimum wage." —President Obama Add your name if you agree: http://ofa.bo/b1Do #With1010 10 Oct 2014

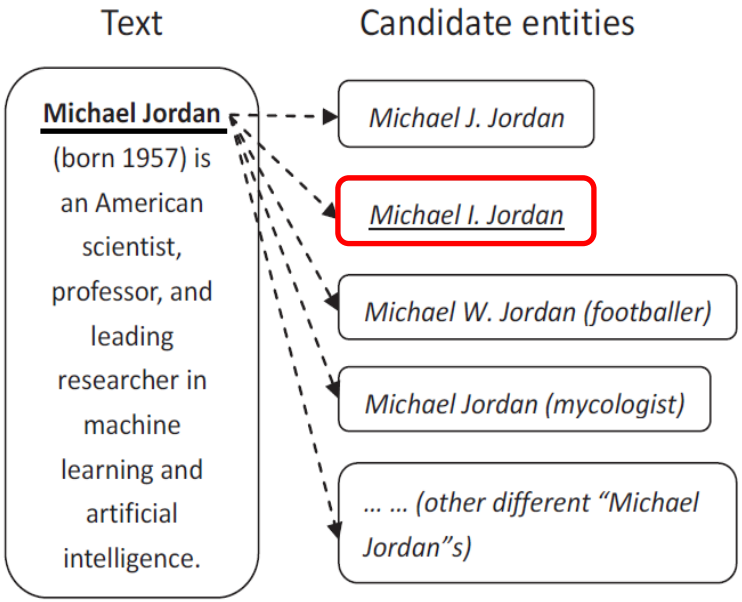
图像

社交主页

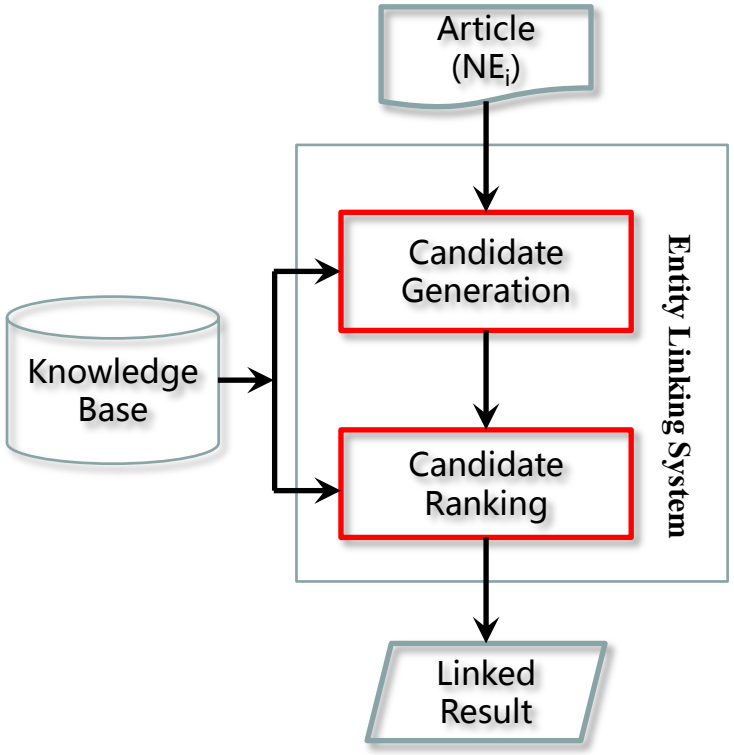
文本

实体链接

问题描述



系统架构



定义：为给定的一段文本（结构化，半结构化，长文本，短文本）中识别出的实体名字m，找到其在知识库中对应的实体em 的过程

实体链接

候选集主要构建方法

- 基于名称字典的构建方法[1]
  - 维基百科中实体页面，重定向页面，排歧页面，页面内超链接（锚文本）等
  - 查询记录和web文档寻找同义词
- 基于上下文环境的名称变体的构建方法[3]
  - 启发式匹配方法（临近括号，N-Gram，子串等）
  - 监督学习方法（通过训练数据集学习复杂缩写的实体全称）
- 基于搜索引擎的构建方法[4]
  - Google搜索返回结果中的wikipedia页面
  - Wikipedia搜索引擎

W. Shen, J. Wang, P. Luo, and M. Wang, "Linden: linking named entities with knowledge base via semantic knowledge," in *WWW, 2012, pp. 449–458.*

W. Zhang, Y. C. Sim, J. Su, and C. L. Tan, "Entity linking with effective acronym expansion, instance selection and topic modeling," in IJCAI, 2011, pp. 1909–1914.

X. Han and J. Zhao, "Nlpr kbp in TAC 2009 KBP track: A two stage method to entity linking," in TAC 2009 Workshop, 2009.

实体链接

特征选择

- 上下文无关的特征[1,2]
  - 字符串比较（edit distance, Dice coefficient score, skip bigram Dice）
  - 实体名称流行度（维基百科中命名实体出现频率）
  - 实体类别（类型一致）
- 上下文相关的特征[3]
  - 非结构化文本（Bag of words, Unigram language model）
  - 结构化信息（知识库中属性，概念，别名，上下位关系）

M. Dredze, P. McNamee, D. Rao, A. Gerber, and T. Finin, "Entity disambiguation for knowledge base population," in COLING, 2010, pp. 277–285.

J. Hoffart, M. A. Yosef, I. Bordino, H. F¨ urstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, "Robust disambiguation of named entities in text," in EMNLP, 2011, pp. 782–792.

W. Shen, J. Wang, P. Luo, and M. Wang, "Linden: linking named entities with knowledge base via semantic knowledge," in *WWW, 2012, pp. 449–458.*

实体链接

- **排歧主要方法（Ranking）**
  - 监督学习方法[5,6]
    - 二分类方法
      - 给定一对实体名称和候选实体，分类器决定实体名称是否指向该候选实体
    - 排序学习方法（SVM ranking）
      - 候选实体之间的偏序关系
    - 图模型
      - 多个实体名称与他们各自的候选集构建图模型进行联合推理（collective inference）
    - 模型组合
      - 将多个学习算法进行集成（投票）
  - 非监督学习方法[7]
    - 向量空间模型（VSM）
      - 将实体名称和候选实体向量化，进行向量相似度计算（如何向量化是关键）
    - 基于信息检索的方法
      - 基于排序的信息检索技术的统计语言模型（KL-divergence 抽取模型）

W. Zhang, Y. C. Sim, J. Su, and C. L. Tan, "Entity linking with effective acronym expansion, instance selection and topic modeling," in IJCAI, 2011, pp. 1909–1914.

S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti, "Collective annotation of Wikipedia entities in web text," in SIGKDD, 2009, pp. 457–466.

S. Gottipati and J. Jiang, "Linking entities to a knowledge base with query expansion," in EMNLP, 2011, pp. 804–813.

语义链接数据的应用

一、语义数据集成

二、互联网语义搜索

三、问答系统

四、基于知识的行业数据分析

语义数据集成

将知识图谱与图谱之外的数据源进行基于语义的集成。
搜狗、百度、谷歌等搜索引擎都实现了语义数据集成。

搜狗搜索

李娜 - 搜狗百科

李娜，湖北武汉人，中国著名女子网球运动员，毕业于华中科技大学新闻系。2008年，李娜获得北京奥运会第4名；2011年，获得法国网球公开赛女单冠军，成为中国乃至亚洲在网球四大满贯赛事上夺得单打冠军的第一人。截止2014年1月，李娜获得了8个WTA和19个ITF冠军...

李娜的微博

李娜的最新相关消息

49分钟前	退役后传递爱心 李娜美山慈善之行送温暖|李娜正...	新浪视频
2小时前	新加坡WTA年终总决赛 李娜作为宣传大使助阵	腾讯体育
3小时前	WTA年终总决赛新加坡启动 李娜变身"宣传大使"	搜狐体育
8小时前	李娜携美山投身公益 与小学生打球拔河开怀大笑	CRI国际在线
10小时前	新加坡年终总决赛将再为李娜举办退役仪式	腾讯体育

李娜的图片 共4625张>>

聊聊你对"李娜"的看法呗

娜姐，你是最棒的！你是中国的骄傲！
9天前 0 回复

SkUv发挥手机,平板,电脑的更大用处吧。全职妈妈、学生、无业者、上班族 赚外快啦！！！一天50-120以上,有意者请加企鹅712389860
4天前 0 回复

娜姐，你是我的骄傲，我永远支持你，加油娜姐，你是最棒的，谢谢
4天前 0 回复

加油
4天前 0 回复

评论 不超过100个字

匿名 发表评论

网球明星 展开


彭帅


玛利亚·莎拉波娃


郑洁


晏紫

互联网语义搜索

在网络搜索时，经常会出现多义的词条。如“李娜”可表示网球运动员李娜和歌手李娜。通常搜索结果会以结果列表的形式给出。

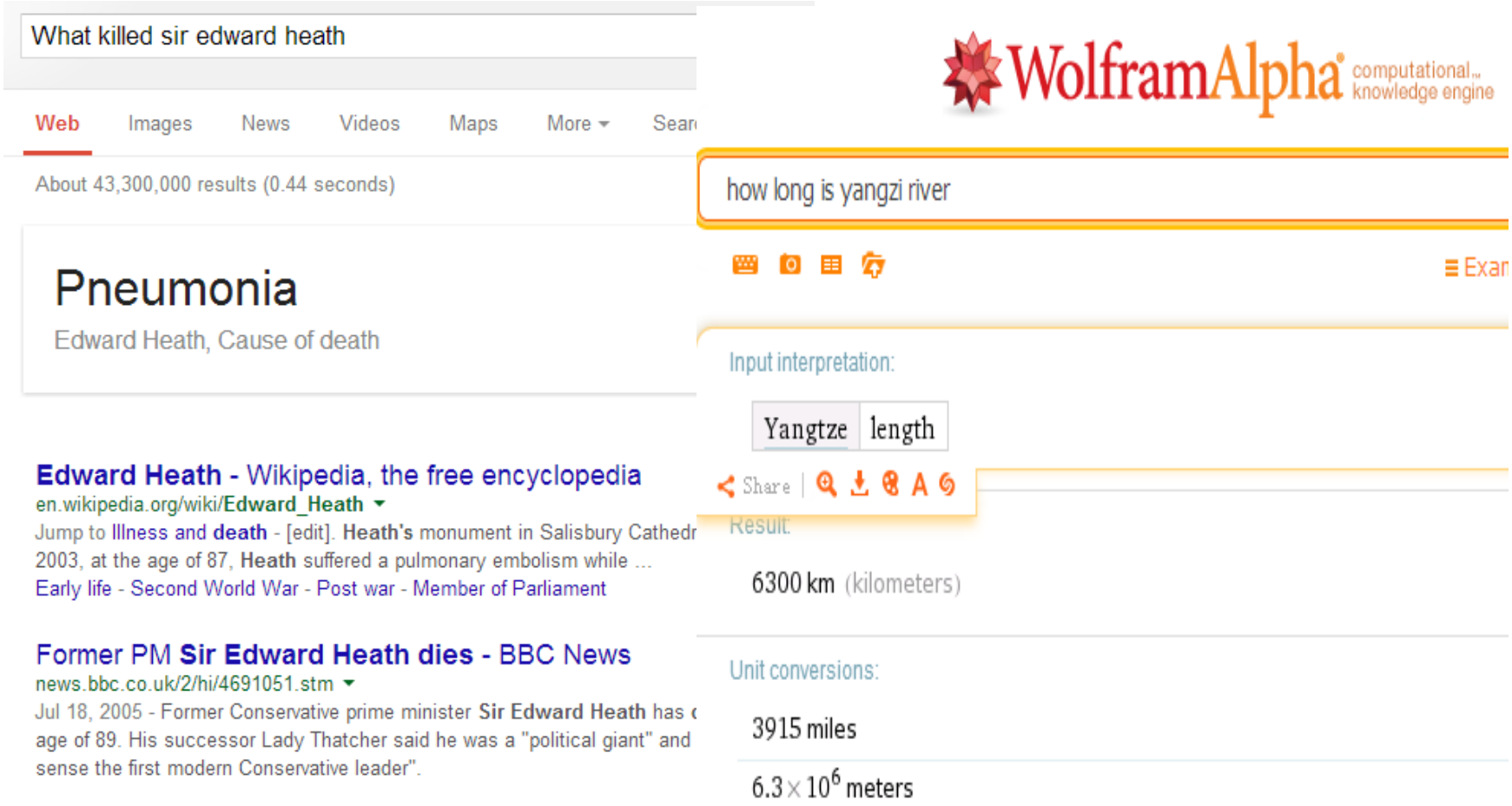
知识图谱的语义链接，使得搜索引擎可以用基于实体的搜索来代替基于字符串的搜索，从而实现搜索时的歧义消除。

Things, not Strings

互联网语义搜索


nd relation summarization
About 1,750,000 results (0.20 seconds)
纸牌屋- 维基百科，自由的百科全书 - Wikipedia
zh.wikipedia.org/zh/纸牌屋 Translate this page
關於香港歌手李克勤于2013年发行的迷你专辑，詳見「纸牌屋(EP)」。 ...
英语：House of Cards）是美国一部以政治为题材的电视连续剧。本剧由鲍
美国国务卿 - 戴维营 - 迈克尔·多布斯
纸牌屋第1季 - 搜狐视频
HOUSE of CARDS
tv.sohu.com/s2013/houseofcards1/
纸牌屋第1季电视剧,纸牌屋第1季在线观看,纸牌屋第
数轮激烈角逐,新一届美国总统加 ...
纸牌屋_百度百科
baike.baidu.com/view/5313136.htm Translate this page
《纸牌屋》是，基于迈克尔·多布斯同名小说创作，由大卫·芬奇执导，鲍尔
凯文·史派西、罗宾·怀特、迈克尔·吉尔等主演的一部政治为题材的Netflix的
凯文·史派西 - 罗宾·怀特 - 党鞭 - 克里斯汀·康诺利
纸牌屋第一季(豆瓣)
movie.douban.com/subject/6037429/ Translate this page
Rating: 9.1/10 - 78,912 votes
纸牌屋第一季电视剧简介和剧情介绍,纸牌屋第一季影评、图片、论坛.
House of Cards
American Television Series
9.1/10 - 豆瓣
House of Cards is an American political drama television series, developed and produced by Beau Willimon. It is an adaptation of BBC's mini-series of the same name and is based on the novel by Michael Dobbs. Wikipedia
First episode date: February 1, 2013
Network: Netflix
Awards: Golden Globe Award for Best Performance by an Actress In A Television Series - Drama, more
Writers: Beau Willimon, Michael Dobbs, Sarah Treem, Andrew Davies, Rick Cleveland, Kate Barnow, Keith Huff, Sam Forman
Nominations: Primetime Emmy Award for Outstanding Drama Series, more
Cast
View 15+ more

问答系统


What killed sir edward heath
Web
Images
News
Videos
Maps
More
About 43,300,000 results (0.44 seconds)
Pneumonia
Edward Heath, Cause of death
Edward Heath - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Edward_Heath
Jump to Illness and death - [edit]. Heath's monument in Salisbury Cathedr
2003, at the age of 87, Heath suffered a pulmonary embolism while ...
Early life - Second World War - Post war - Member of Parliament
Former PM Sir Edward Heath dies - BBC News
news.bbc.co.uk/2/hi/4691051.stm
Jul 18, 2005 - Former Conservative prime minister Sir Edward Heath has
age of 89. His successor Lady Thatcher said he was a "political giant" and
sense the first modern Conservative leader".
WolframAlpha computational knowledge engine
how long is yangzi river
Input interpretation:
Yangtze length
Share
Result:
6300 km (kilometers)
Unit conversions:
3915 miles
6.3 × 10^6 meters

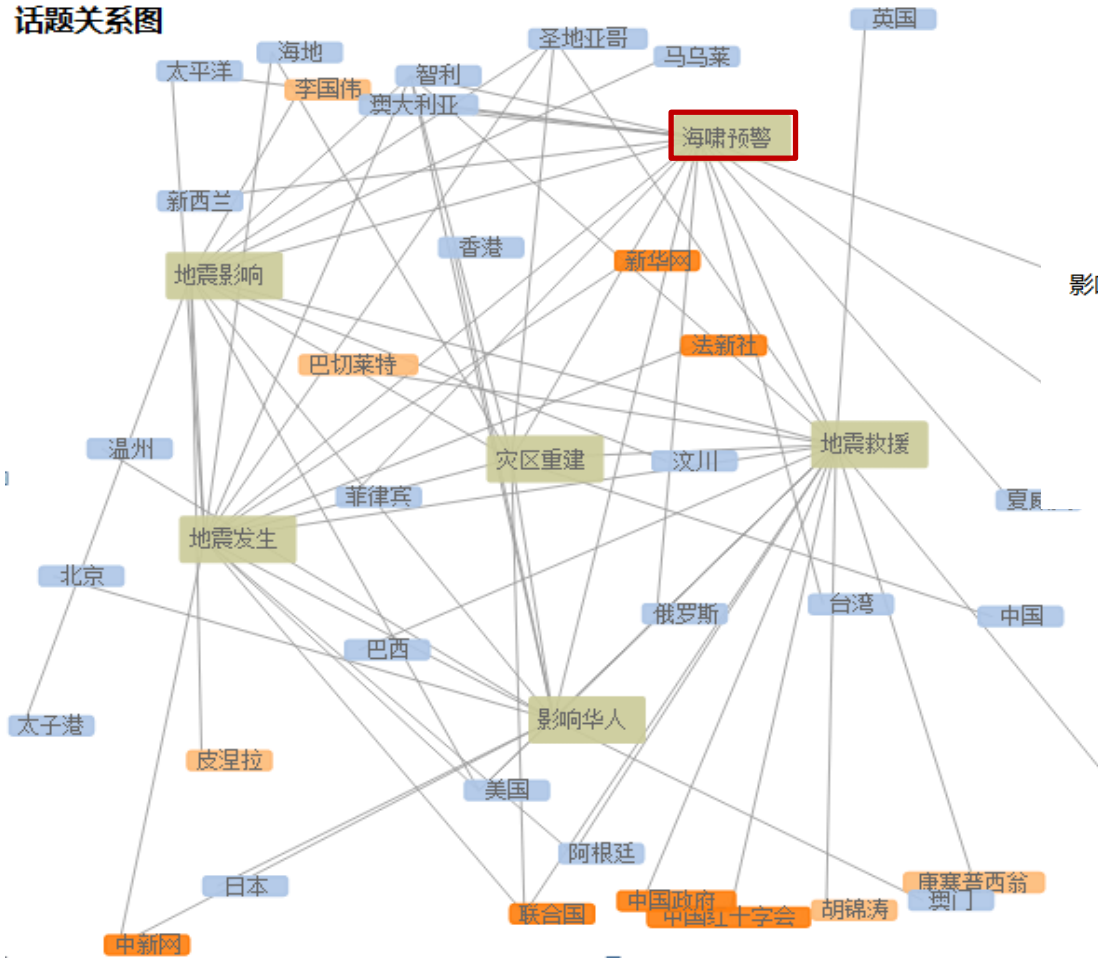
基于知识的行业大数据分析

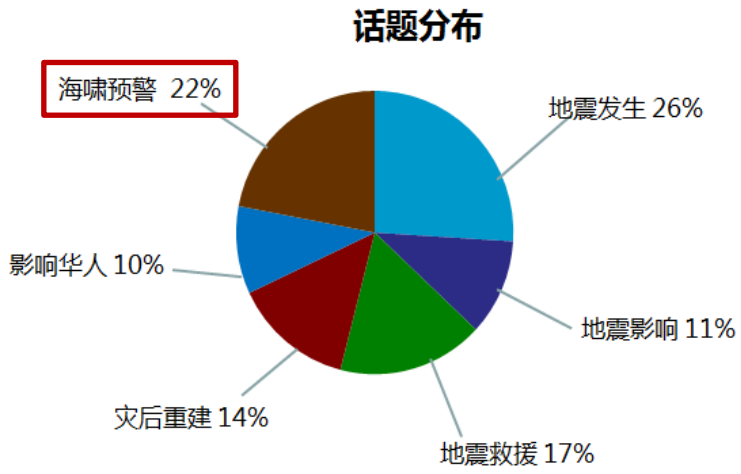
影视大数据分析

- 最具影响力和市场价值的主力受众：中年男性专业人士
- 受欢迎电视剧类型：政治惊悚剧
- 受欢迎导演：大卫•芬奇
- 受欢迎演员：凯文•史派西
- 观看偏好：一次观看多集
- 基于知识图谱的影视元素关系挖掘：预测出凯文.史派西、大卫.芬奇和“BBC出品”三种元素结合在一起的电视剧产品
- 相比传统文本的方式大大提高了影视数据分析的精准度和可行性

基于知识的行业大数据分析

新闻事件—2010 智利地震：话题关系图与分布图





总结

- 知识图谱使互联网从字符串描述到客观世界的具体事物描述
- 互联网为知识图谱构建提供了丰富的资源
- 知识图谱是大数据语义链接的基石
- 知识图谱互联网理解世界的基础设施

A little semantic, a long way to go.

We are on the way …

谢谢！
Q&A

李涓子
清华大学
lijuanzi@tsinghua.edu.cn