



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



网易新闻 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

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

易达旅行

工作时间：周一至周日 08:30-23:00

010-89678959



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

到达 9月29日 05:50

戴高乐机场

飞行 约12小时50分钟

成人退改签规则：^

退票规定：不得退票

改期规定：不得改期

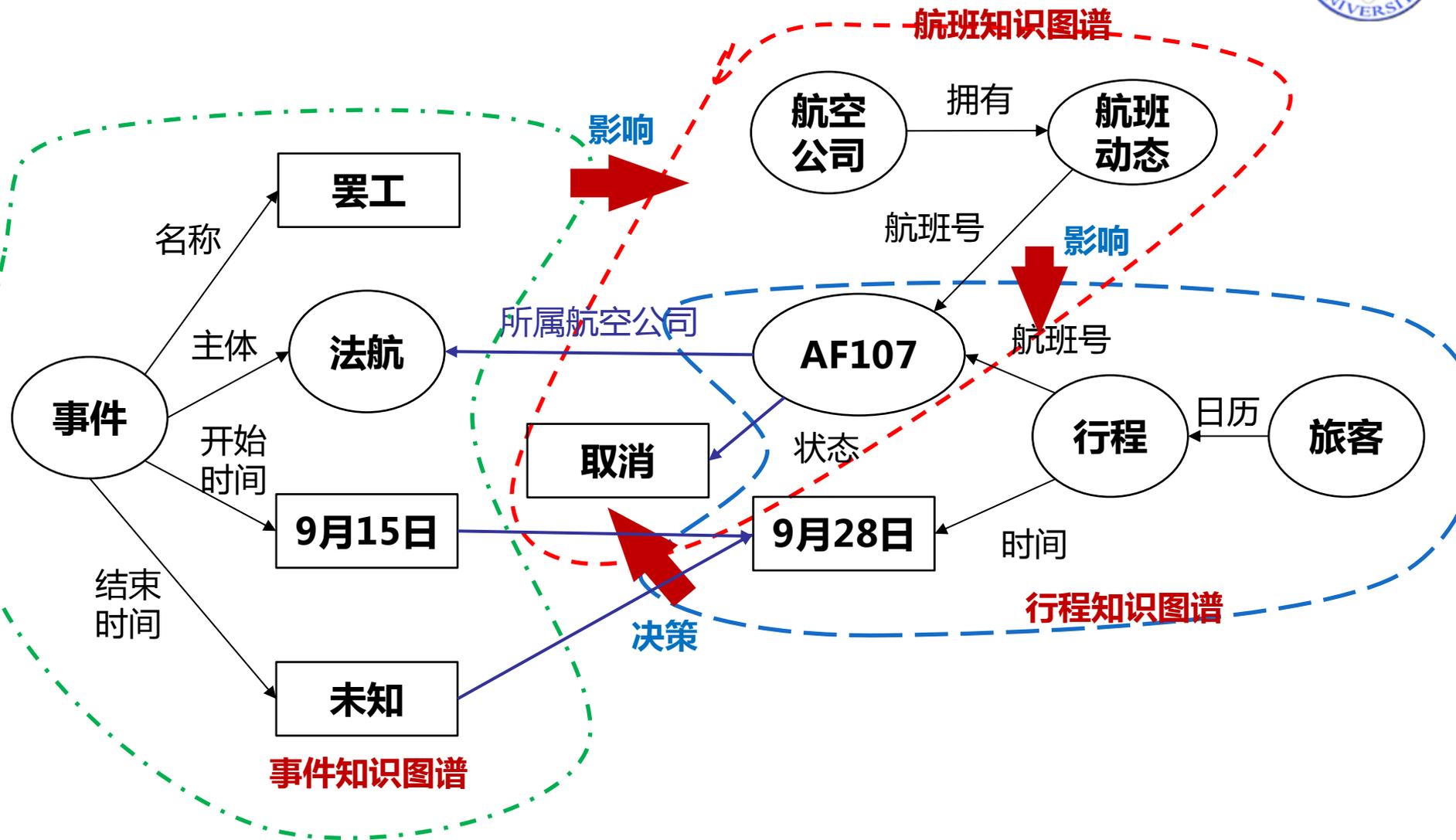
行李额规定：1pc

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

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



250概念
4M实例
6000属性
500M三元组
在线更新



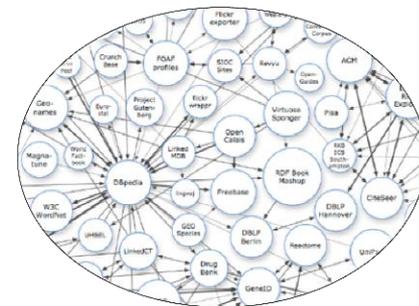
350K概念
10M实例
100属性
120M三元组



15K概念
40M实例
4000属性
1B三元组
Google KB核心



50M义项
50+种语言
262M三元组



850K概念
8M实例
70K属性



Google KG

15K概念
600M实例
20B三元组

NELL

OpenIE
(Reverb, OLLIE)

WordNet
7种欧洲语言
跨语言链接



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

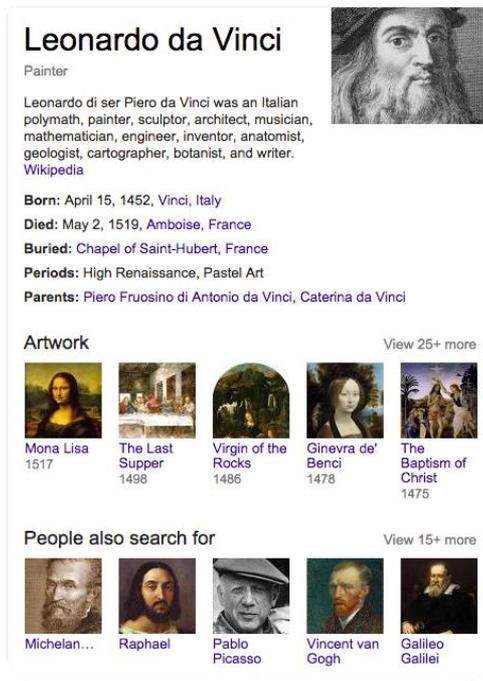
实体及其之间的关系图。

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？



Leonardo da Vinci
Painter

Leonardo di ser Piero da Vinci was an Italian polymath, painter, sculptor, architect, musician, mathematician, engineer, inventor, anatomist, geologist, cartographer, botanist, and writer.
Wikipedia

Born: April 15, 1452, Vinci, Italy
Died: May 2, 1519, Amboise, France
Buried: Chapel of Saint-Hubert, France
Periods: High Renaissance, Pastel Art
Parents: Piero Fruosino di Antonio da Vinci, Caterina da Vinci

Artwork View 25+ more

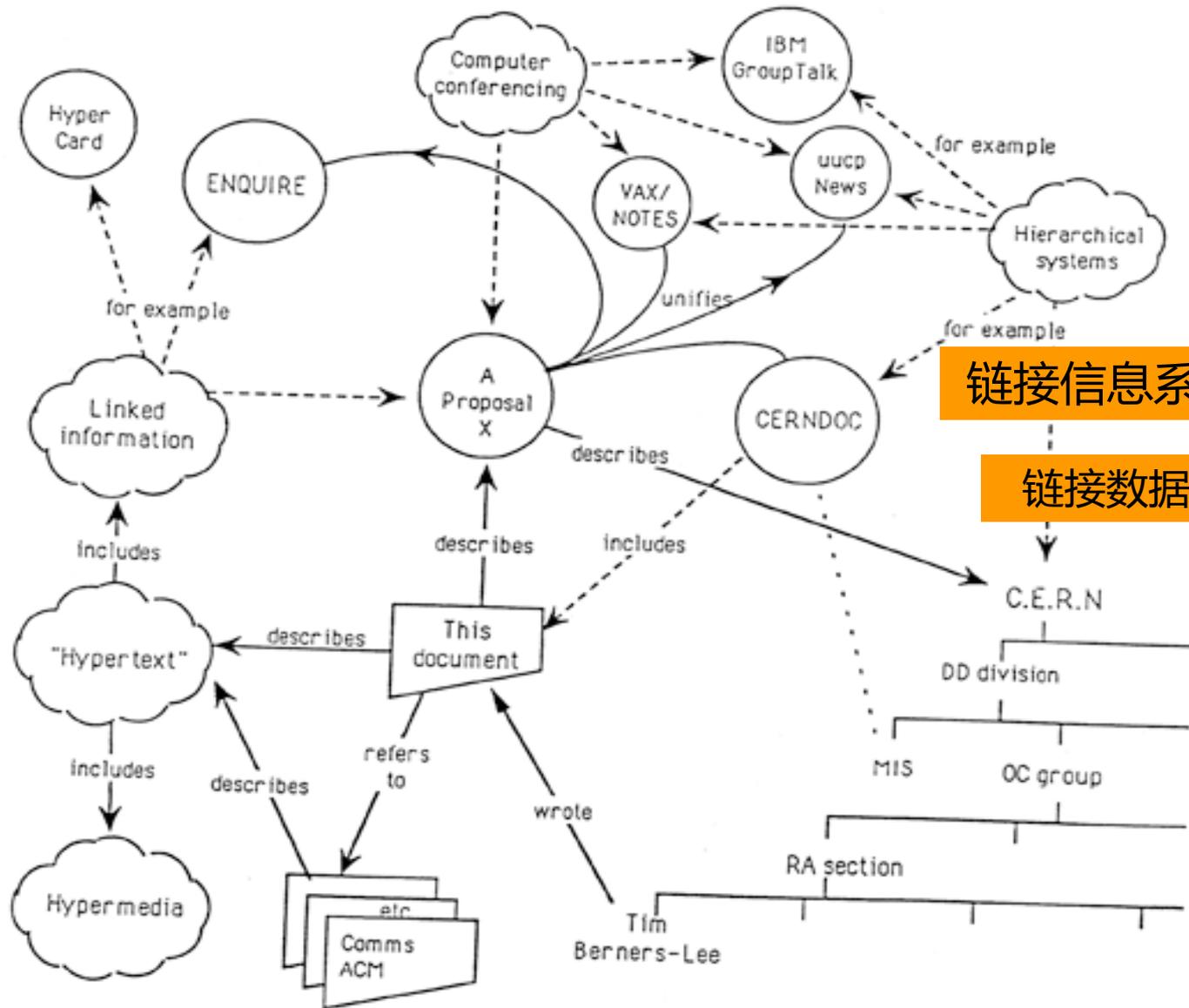
 Mona Lisa 1517	 The Last Supper 1498	 Virgin of the Rocks 1486	 Ginevra de' Benci 1478	 The Baptism of Christ 1475
--	--	--	--	--

People also search for View 15+ more

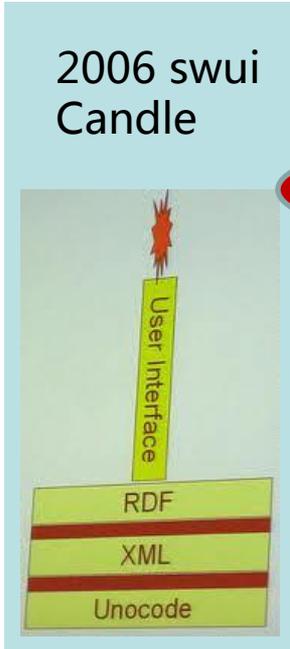
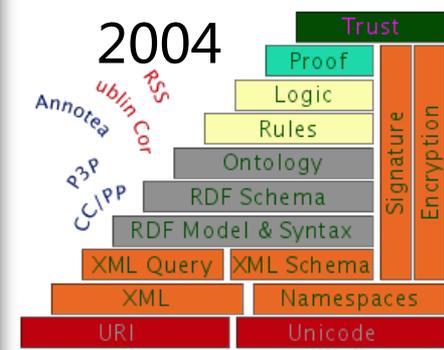
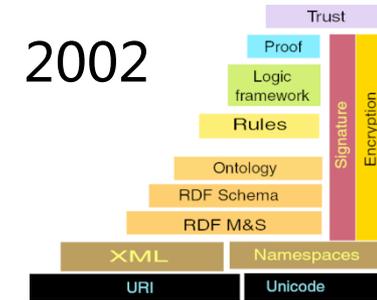
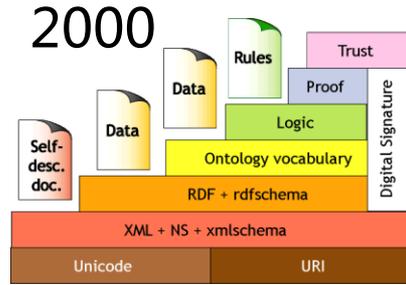
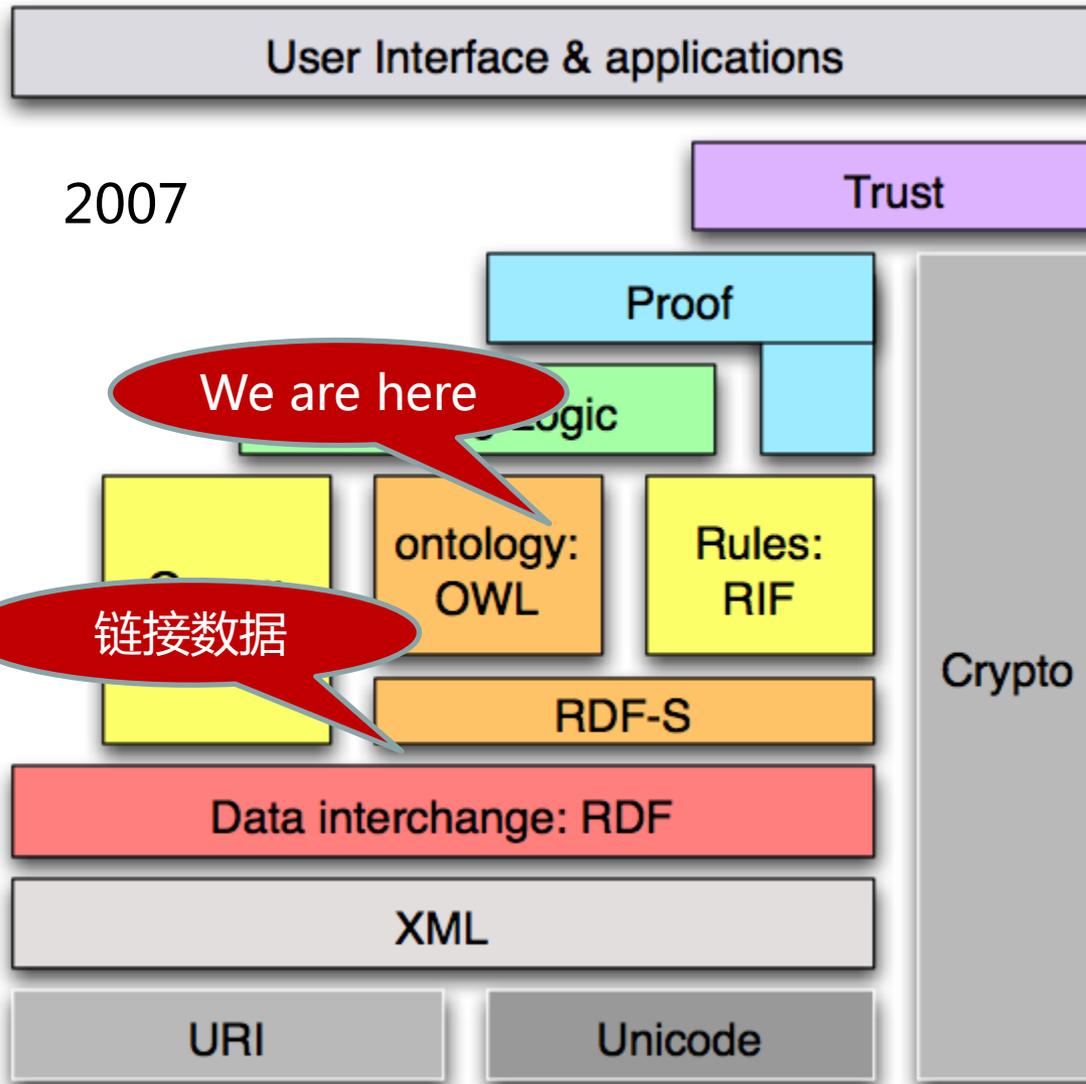
 Michelangelo	 Raphael	 Pablo Picasso	 Vincent van Gogh	 Galileo Galilei
---	--	--	---	--



Tim Berners-Lee's Proposal 1989



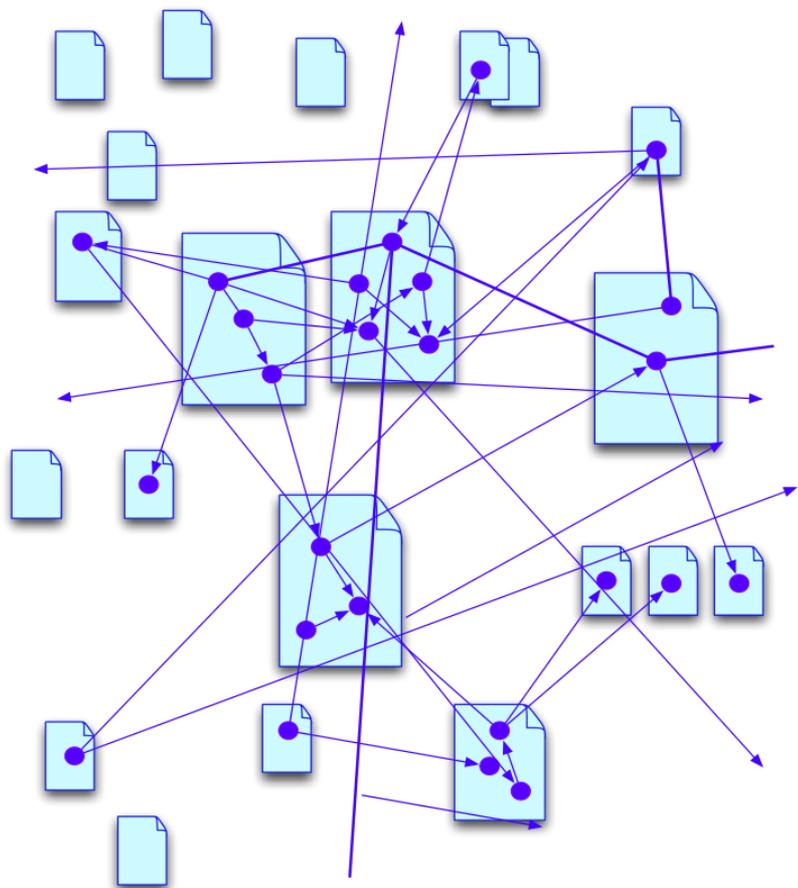
万维网信息描述语言塔



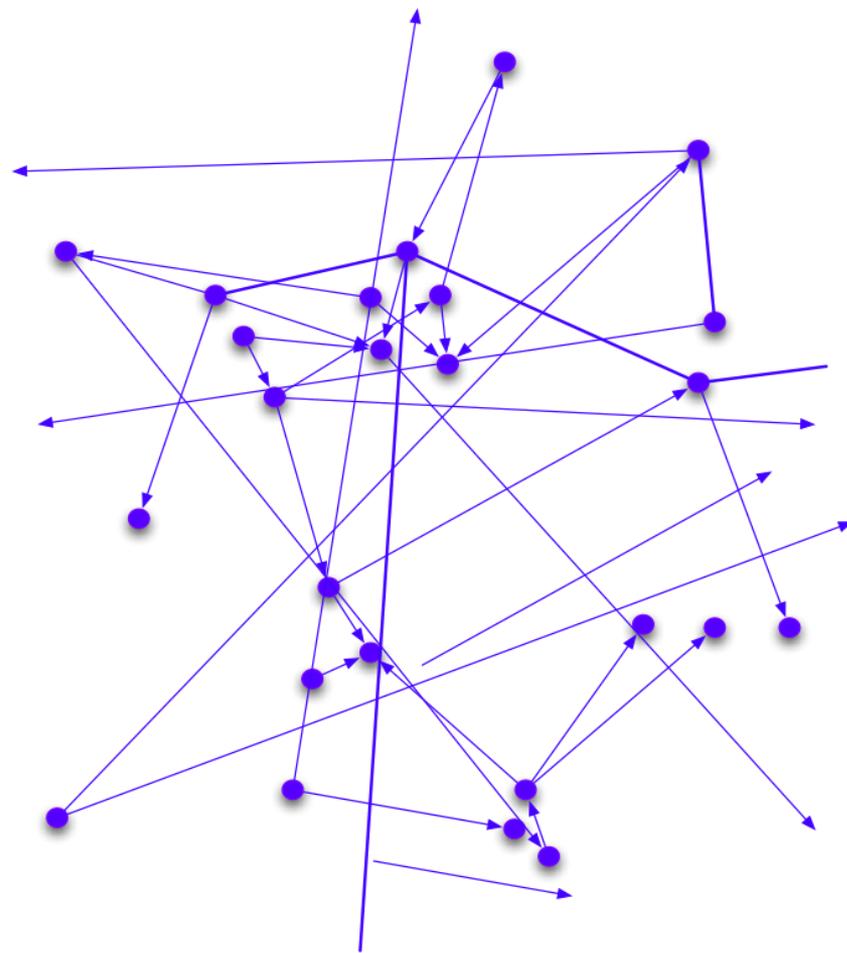
<http://bbs.w3china.org/dispbbs.asp?boardID=2&ID=86430>

从文档万维网到数据万维网

文档万维网



数据万维网



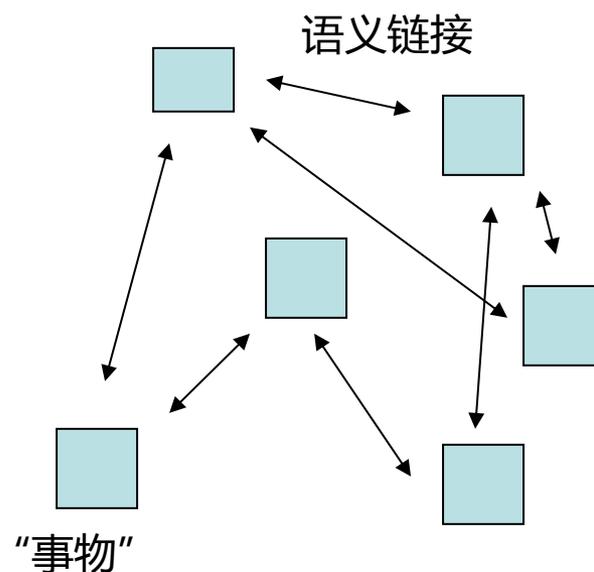
<http://www.w3.org/2007/Talks/1211-whit-tbl/#%2828%29>

数据万维网

• 特征：

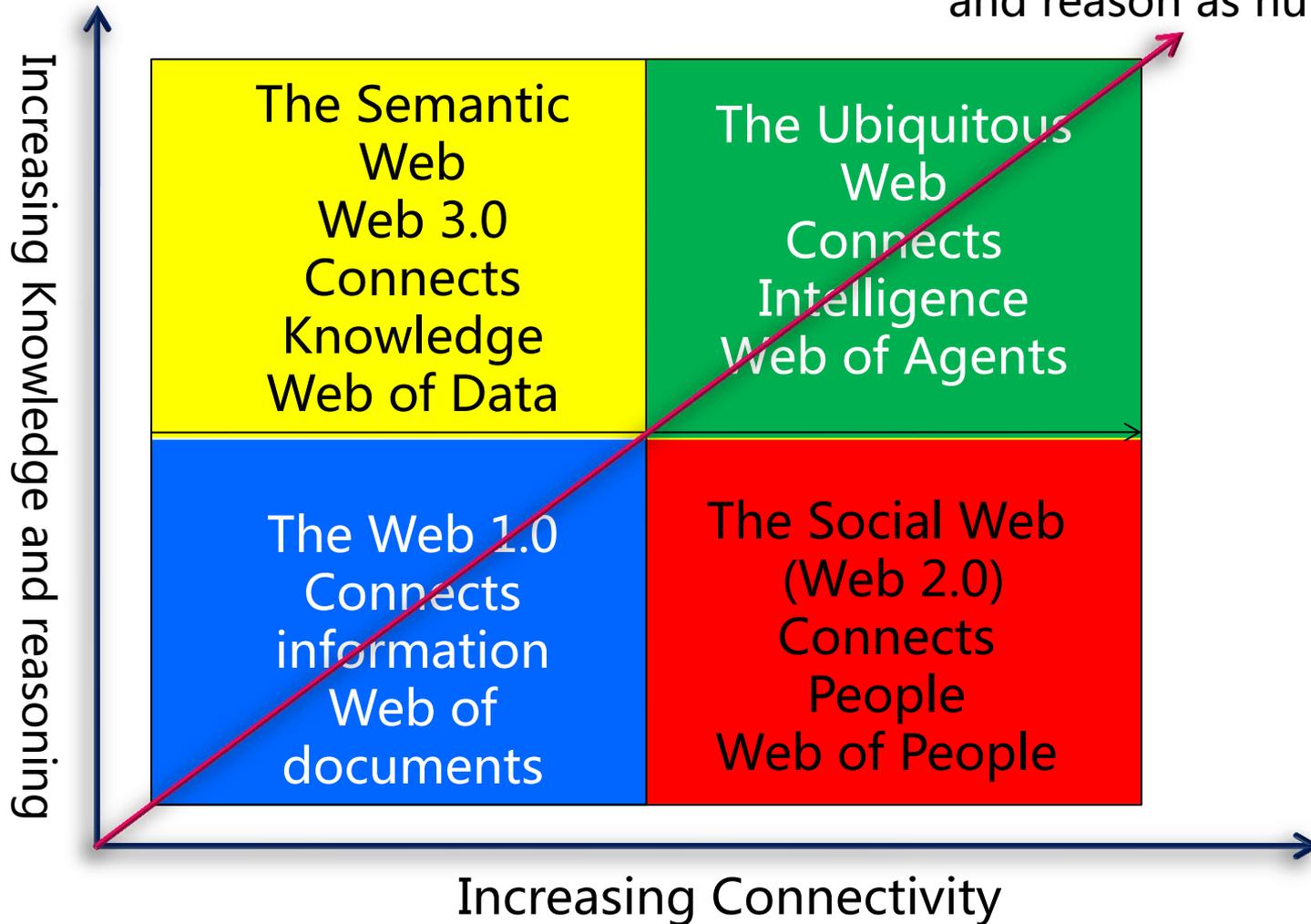
- Web上的事物拥有唯一的URI
- 事物之间由链接关联（如人物、地点、事件、建筑物）
- 事物之间链接显式存在并拥有类型
- Web上数据的结构显式存在

- 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



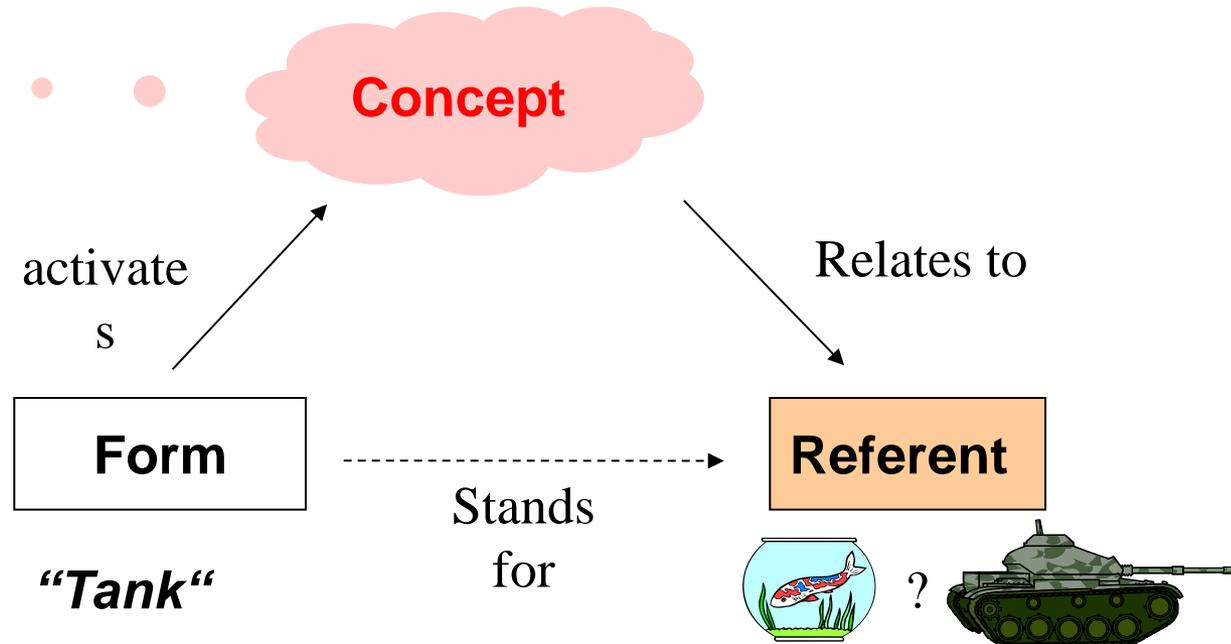
Bring structure to the meaningful content of Web pages



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

□ 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

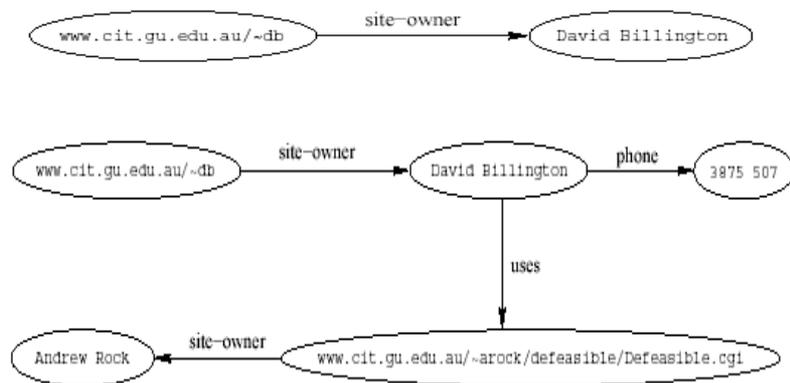
- 使用**URI**唯一标示一个资源
- 一个资源通常表示一个事物(Thing)

■ 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**

■ *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)

} Taxonomy知识

} 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

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.hits.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



<i>name</i>	<i># of concepts</i>	<i># of isA pairs</i>
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



维基百科
自由的百科全书

- 首页
- 分类索引
- 特色内容
- 新闻动态
- 最近更新
- 随机条目

帮助

- 帮助
- 社区主页
- 方针与指引
- 互助客栈
- 知识问答
- 字词转换
- IRC即时聊天
- 联系我们
- 关于维基百科
- 资助维基百科

工具

- 链入页面
- 相关更改
- 上传文件
- 特殊页面
- 打印版本
- 固定链接
- 永久链接

页面 讨论 简体 汉 藏 阅读 编辑

2014年鲁甸地震 [编辑]



本文记述一项中国大陆新闻。随着事件发展，内容可能会快速更新。

除特别注明外，本文所有时间均以东八区时间（UTC+8）为准。

2014年鲁甸地震发生于北京时间（UTC+8）2014年8月3日16时30分10秒，地震规模为里氏6.5级^[1]。震中位于中国云南省昭通市鲁甸县，震源距离地表12千米。震中在昭通市鲁甸县龙头山镇，距离昭通市区约49公里，昆明、重庆、成都、西安等地震感明显。^[8]

目录 [隐藏]

- 背景
- 伤亡及损害
- 救援
- 分析
- 余震
- 各界反应
 - 6.1 两岸三地
 - 6.2 国际
- 捐赠
 - 7.1 政府及社会组织
 - 7.2 企业
 - 7.3 个人
- 纪念
- 相关事件
- 参考资料
- 外部链接

主题

属性

infobox

2014年鲁甸地震



主题

日期 2014年8月3日 16:30:10 (UTC+8) ^[1]

震中 27.1°N 103.3°E﻿ / ﻿27.1°N 103.3°E﻿ / ﻿

规模 6.5 M_L^[1]6.1 M_w^[2]

最大烈度 IX

深度 12千米^[1]

次数 783次余震^[3]
(截止8月6日 20:00)

最大规模 4.2 M_L余震^[3]
(截止8月6日 20:00)

破坏

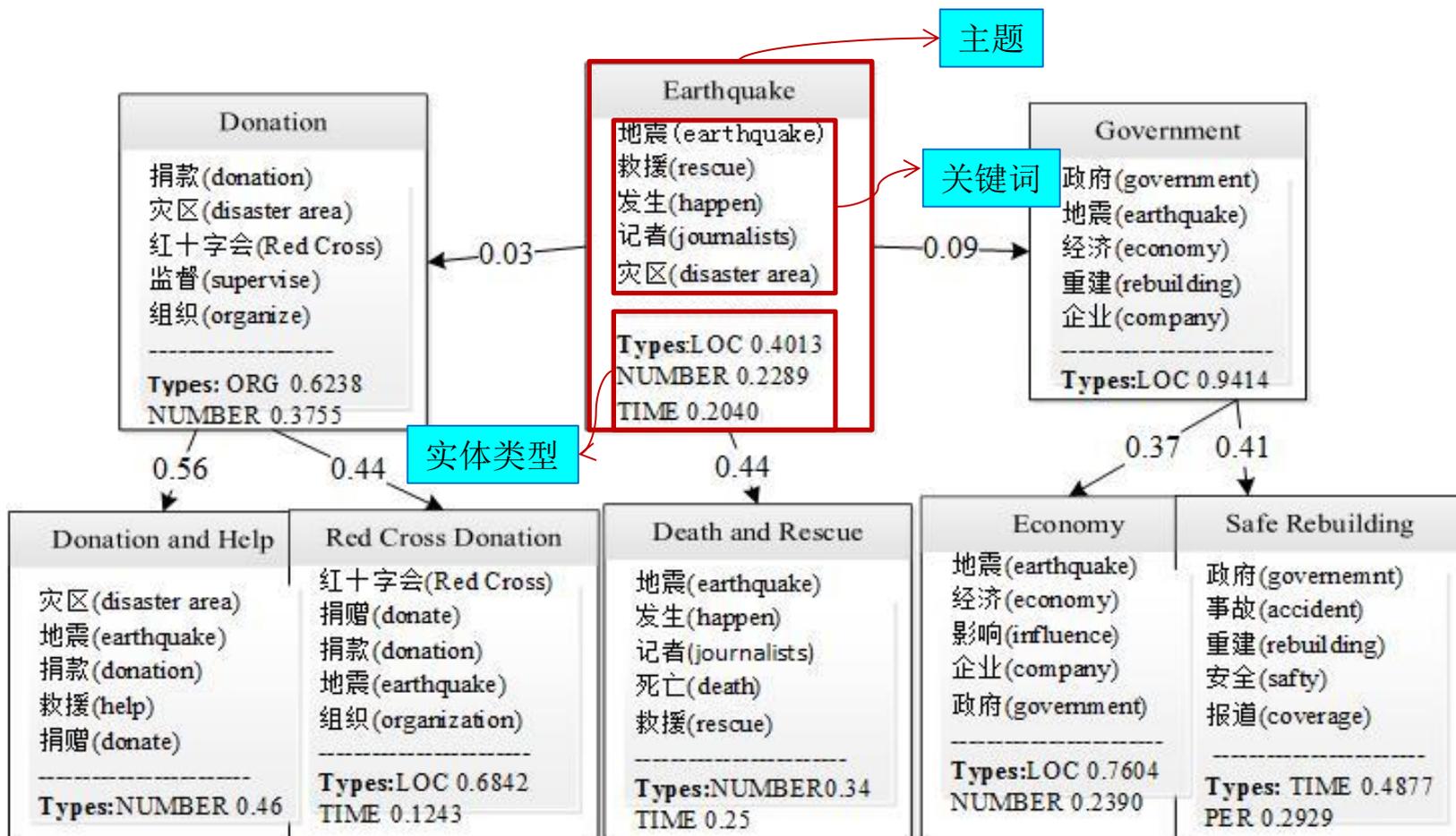
受影响地区 中国云南省昭通市鲁甸县、巧家县、曲靖市会泽县^[1]

伤亡人数 死亡：617
(截止8月8日 15:00) ^[4]
受伤：3,143
(截止8月8日 15:00) ^[4]
重伤：？
(截止8月8日 15:00) ^[4]
失踪：112
(截止8月8日 15:00) ^[4]
转移安置：22.97万
(截止8月8日 15:00) ^[4]
受灾：108.84万
(截止8月8日 15:00) ^[4]

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

WolframAlpha computational knowledge engine



计算知识引擎WolframAlpha

The image shows a grid of 12 category tiles from the WolframAlpha interface. Each tile has an orange header and a white body with icons and text. The categories are:

- MATHEMATICS**: Includes icons for a knot, a graph, and a fraction. Text: $2x^2 + 4x^2 - 2x - \dots$, $x^2 + \frac{1}{x^3} - x - \dots$. Links: Elementary Math, Numbers, Plotting, Algebra, Matrices, Calculus, Geometry, Trigonometry, Discrete Math, Number Theory, Applied Math, Logic, Functions, Definitions, ...
- WORDS & LINGUISTICS**: Includes a keyboard icon. Text: Word Properties, Dictionary, Lookup, Word Puzzles, Anagrams, Languages, Document Length, Morse Code, Soundex, Number Names, Character Encodings, ...
- UNITS & MEASURES**: Includes icons for a scale and a thermometer. Text: 30 093 meters, 3.009×10^6 cm (cent), 98 732 feet, 16.25 mm (nauclos). Links: Conversions, Calculations, Comparisons, Dimensional Analysis, Industrial & Construction, Batteries, Bulk Materials, Paint, Display Formats, Ring Sizes, Shoe Sizes, ...
- STEP-BY-STEP SOLUTIONS PRO**: Includes a pencil icon. Text: Move terms with x to the left hand side. Subtract 2x from both sides: $(4x - (2x)) - 6 = (2x - (2x)) + 8$. Links: Chemistry, Arithmetic, Number Theory, Algebra, Trigonometry, Calculus, Linear Algebra, Statistics, ...
- STATISTICS & DATA ANALYSIS**: Includes a bell curve icon. Text: Descriptive Statistics, Statistical Inference, Regression, Statistical Distributions, Random Variables, Probability, ...
- PEOPLE & HISTORY**: Includes an image of Albert Einstein. Text: People, Genealogy, Names, Occupations, Political Leaders, Historical Events, Historical Periods, Historical Countries, Historical Numerals, Historical US Money, Inventions, ...
- DATES & TIMES**: Includes clock icons. Text: Tokyo, Champaign, Illinois. 4:26:37 am JT, 2:26:37 pm CDT. Links: Date Computations, Time Zones, Calendars, Holidays, Geological Time, Birthstones, Birth Flowers, Wedding Anniversaries, ...
- DATA INPUT PRO**: Includes a table of city prices. Text: Automatic Analysis, Statistical Analysis, Time Series Analysis, Geographic Data, Data Visualization, ...
- CHEMISTRY**: Includes a ball-and-stick molecular model. Text: Elements, Compounds, Ions, ...
- CULTURE & MEDIA**: Includes an image of people. Text: Doreenall Analytine, Nntahla Terte, ...
- MONEY & FINANCE**: Includes a stock market chart. Text: Stock Data, Indices, Mutual Funds, Futures, Mortgages, Present Value, Currency, Time, ...
- IMAGE INPUT PRO**: Includes an image of a camera lens. Text: Image Analysis, Image Filtering, Feature Detection, Color Processing, Image Effects, ...

Input: $x^3 - 6x^2 + 4x + 12$

Input interpretation: plot $x^3 - 6x^2 + 4x + 12$

Plots:

Enable interactivity

Input: 5" by 7" photo vs 8" by 10" photo

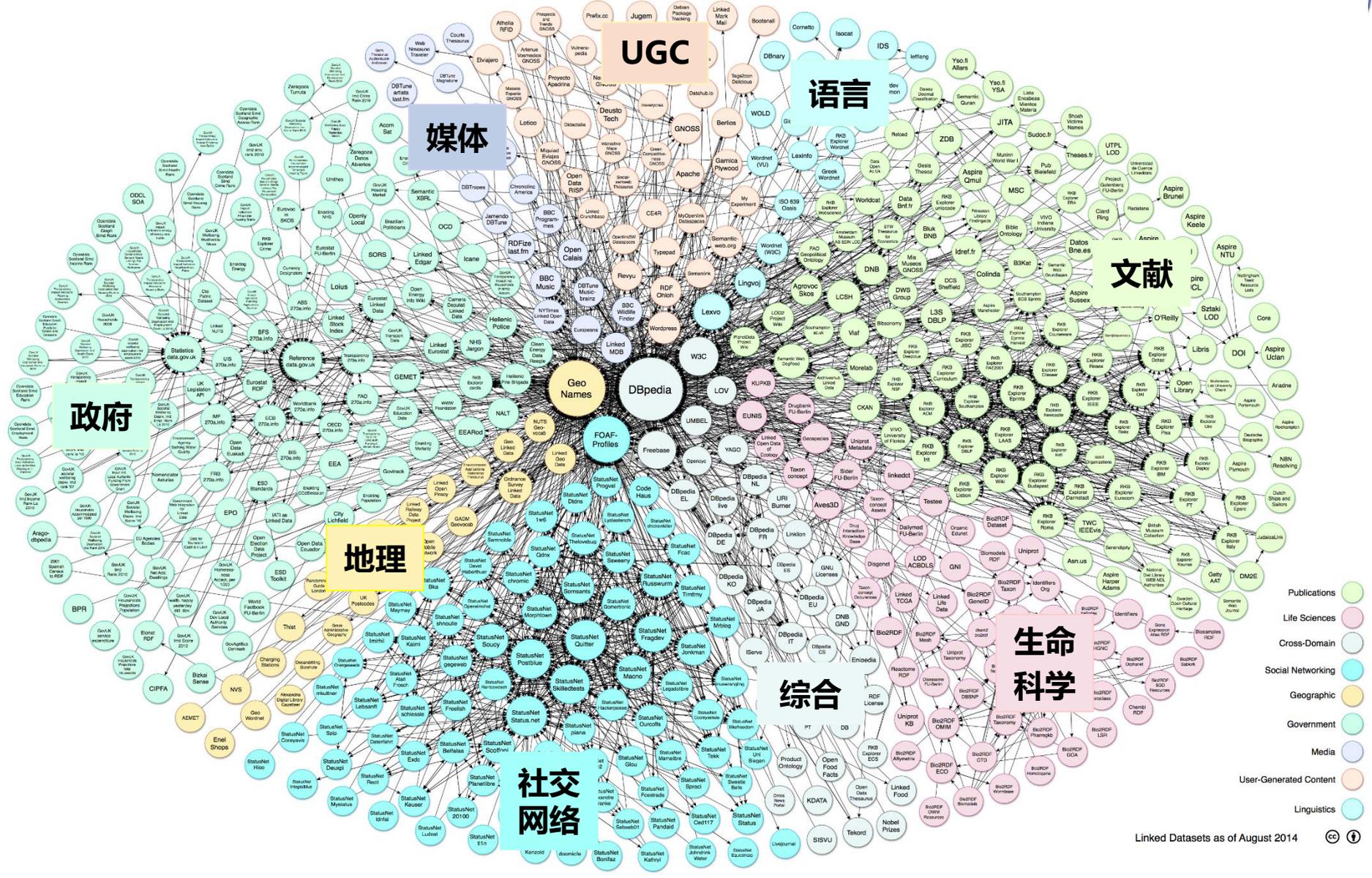
Input interpretation: 5 in x 7 in (display size) | 8 in x 10 in (display size)

Basic properties:

	5 in x 7 in	8 in x 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



- 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 <i>The Continent</i>	
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
配乐作曲	小林武忠
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



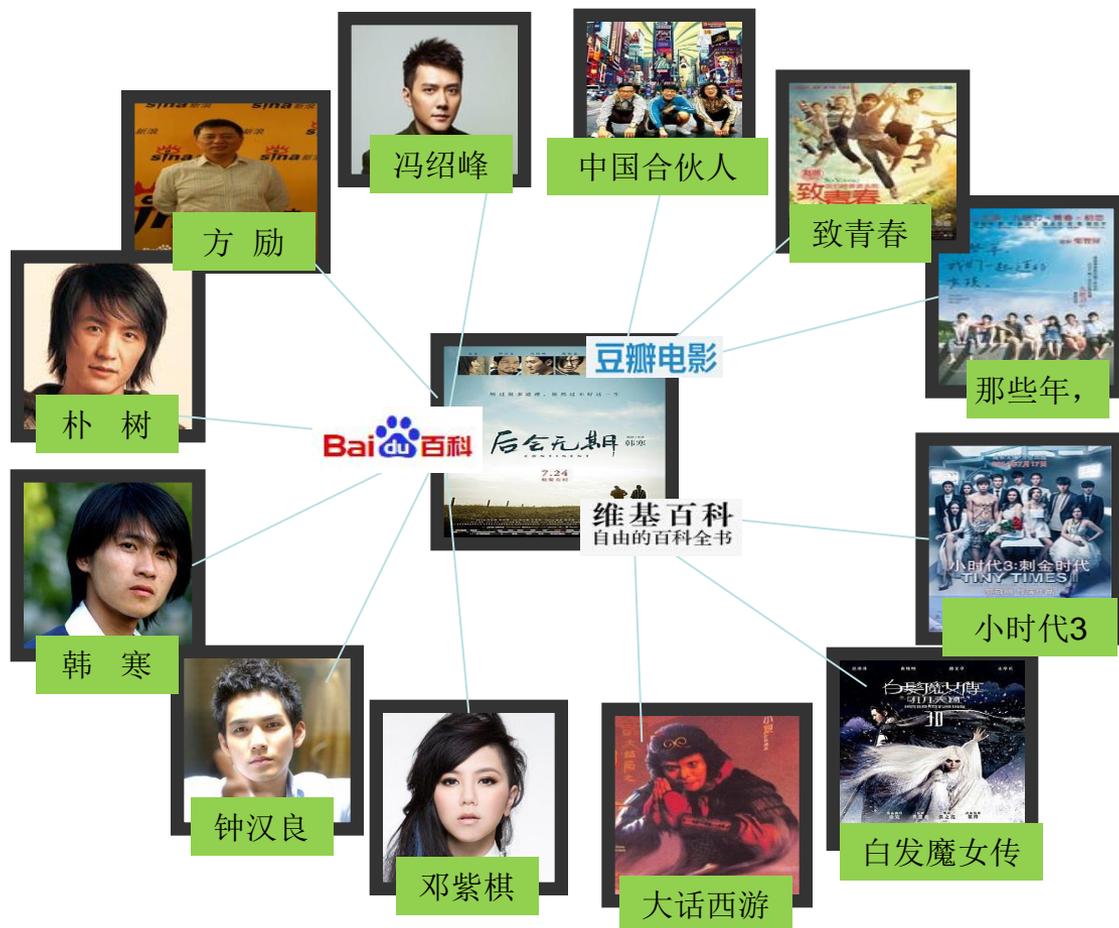
链接补充

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

影视知识融合

相关实体:



isA关系验证:

	食人女	isA
	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 .

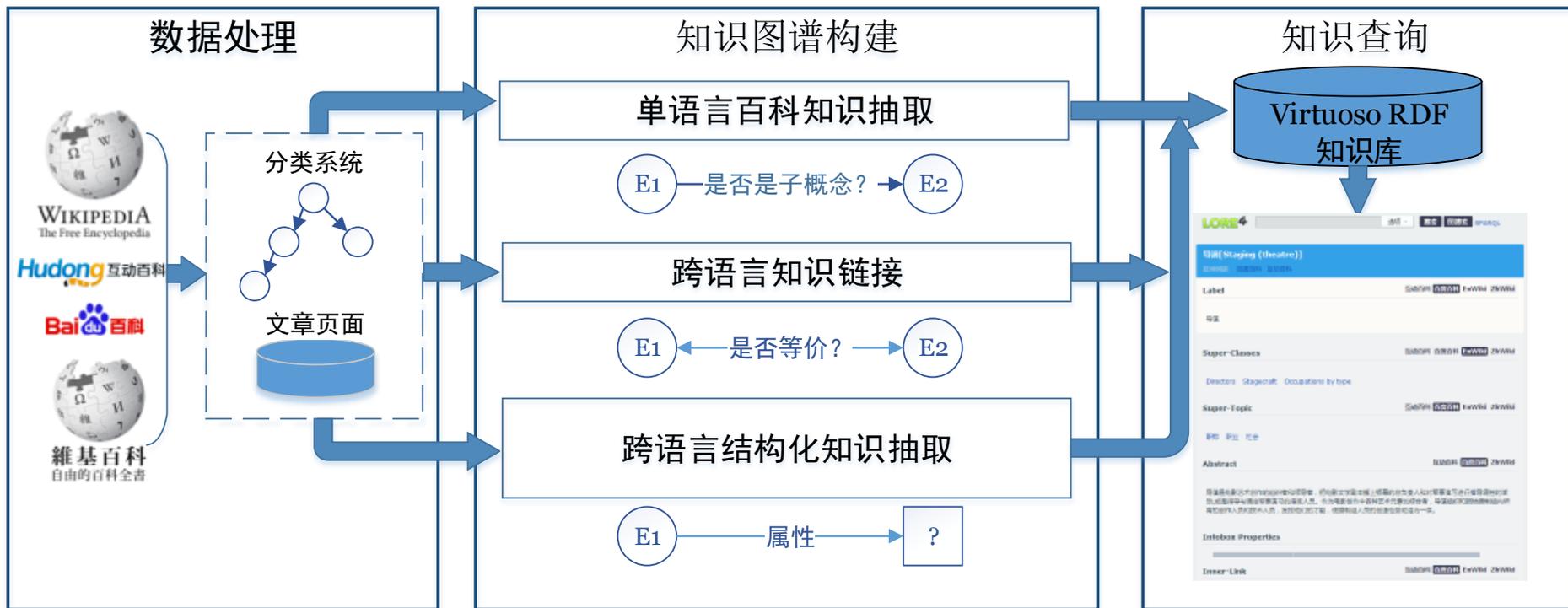
movie:genres
 “Comedy”@en,
 “喜剧”@zh .

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



Properties Instances Publications

XLORE All Search SPARQL

About

Xlore is to extract structured information from heterogenous cross-lingual online wikis and to share the extracted knowledge on the Web. To the best of our knowledge, Xlore is the first large-scale knowledge graph with balanced quantity of Chinese-English knowledge. Currently, Xlore contains **856,146** classes, **71,596** properties and **7,854,301** instances. It gives a new way for building a large-scale knowledge graph with balanced quantity of knowledge across any two different languages.

856146 Classes **71596** Properties **7854301** Instances

Class Taxonomy

Label	Class	Super Classes	Sub Classes	Super Topics	Sub Topics	Properties	Instances	Instances topic	category
Social 社会科学	Main topic classifications Root 页面分类 总分类	4	35	1	32	518	5196	161	Class
Culture 文化	Main topic classifications Root 页面分类 总分类	1	30	2	69	2077	8029	104354	Class
Technology 科技	Main topic classifications Root 页面分类 总分类	3	26	2	61	24	74	2614	Class
People 人物	Main topic classifications Root 页面分类 总分类	2	25	2	24	2758	20236	441944	Topic
Sports 体育	Main topic classifications Root 页面分类 总分类	2	18	5	50	940	420	15863	Topic

Showing 1 to 5 of 25 entries

Statistics on Sample Instances

Label	Super Classes	Sub Classes	Properties	Related Instances	Linked Instances
Italy 意大利	19	1	107	132	1382
Cyprinidae 鲤科	6	2	15	11	1590
1979 1979年	4	0	0	10	1523
Japan 日本	15	1	116	298	1168
陈毅(文字或标)	1	2	10	294	22

Showing 1 to 5 of 25 entries

文化[Culture]

Visualization

Label Hudong Baike Baidu Baike EnWiki ZhWiki

文化

Sub Classes Baidu Baike Hudong Baike EnWiki ZhWiki

饮食文化 语言 建筑 宗教 次文化 宗教文化 铁路时代 各国文化 电视剧

Super Topics Baidu Baike Hudong Baike EnWiki ZhWiki

社会学 社会 总分类

Sub Topics Baidu Baike Hudong Baike EnWiki ZhWiki

文物古迹 民俗 文献 都市设计 文化人类学 人类形象 青铜时代 体育 方言 电视 游戏 网络文化 习俗 娱乐 时尚 电影 学派 艺术 节日 传统 社会制度 服装 符号 虚构 集邮 无神论 反马托邦

Properties Baidu Baike Hudong Baike EnWiki ZhWiki

A	爱好	
B	笔顺 阅读时间 笔顺 部首 部首笔划 毕业院校 别称 别名 (绰号)	***
C	创建时间 创建地点 创建日期 全部 词性 成就 出版时间 出处	***
D	淡新门匾 代表作家 地区 地址 代表作 代表作品 导演 电话区码	***
E	二条法	
F	分布 发现时间 发现者 反义词 繁体 发行时间 分布地区 法人	***
G	病 标准 GDP 馆藏精品 国籍 馆藏数量 公历日期 规范汉字编号	
H	海拔	
I	ISBN	
J	净利润 界 交通信息 机场 简介 交战各方 就诊科室 教师人数	***
K	开本 开网时间 科	

<http://xlore.org/>

跨语言知识图谱XLORE



黄岩岛[Scarborough Shoal]

Extended Reading Baidu Baïke Hudong Baïke Visualization

Label Baidu Baïke **Hudong Baïke** EnWiki ZhWiki

黄岩岛

Type Baidu Baïke **Hudong Baïke** EnWiki ZhWiki

中国领土争议地区 南海诸岛 地理 菲律宾岛屿 有争议的地区 海南岛屿

Abstract Baidu Baïke **Hudong Baïke** EnWiki

黄岩岛（曾用名：民主礁），国际上称之为斯卡伯勒浅滩（Scarborough Shoal），是中国中沙群岛中惟一露出水面的岛礁，位于北纬15°07'，东经117°51'，距中沙环礁约160海里。黄岩岛的领土主权属于中华人民共和国，行政管理属于中国海南省西南中沙群岛办事处。黄岩岛以东是幽深的马尼拉海沟，马尼拉海沟是中国中沙群岛与菲律宾群岛的自然地理分界。

Infobox Properties

所属省份	海南省
所属群岛	中国南海中沙群岛
所属国家	中国
面积	150平方千米
名称	黄岩岛[Scarborough Shoal]
别名	民主礁
坐标	北纬15°07'，东经117°51'
location	South China Sea
highest mount	South Rock ({{zh s=南岩 p=Nan Yan}})
elevation	{{convert 3 m ft}}
country 2 claim	Republic of China (Taiwan)
country 1 claim divisions	[Masinloc, Zambales]
country 2 claim divisions	高雄市[Kaohsiung]

中文名

Label **Hudong Baïke**

中文名

Type

ObjectProperty

Domain Baidu Baïke **Hudong Baïke** EnWiki ZhWiki

休闲娱乐 天文学及天体 1922年出生 1847年出生 私房菜 日韩汽车品牌 华北电力大学 野鸡大学 中国台湾村落 同性恋电影

Ranges Baidu Baïke **Hudong Baïke** EnWiki ZhWiki

休闲娱乐 天文学及天体 1922年出生 日韩汽车品牌 华北电力大学 野鸡大学 中国台湾村落 同性恋电影 新西兰行政区划 哈尔滨市

Instances Baidu Baïke **Hudong Baïke** EnWiki ZhWiki

江西航空职业技术学院 高树森 猪胃虫病 《首富隆起》 宋建彰 科赫尔斯奇霍伯 王身立 费电三社自然村 瓦日乡 王迎军

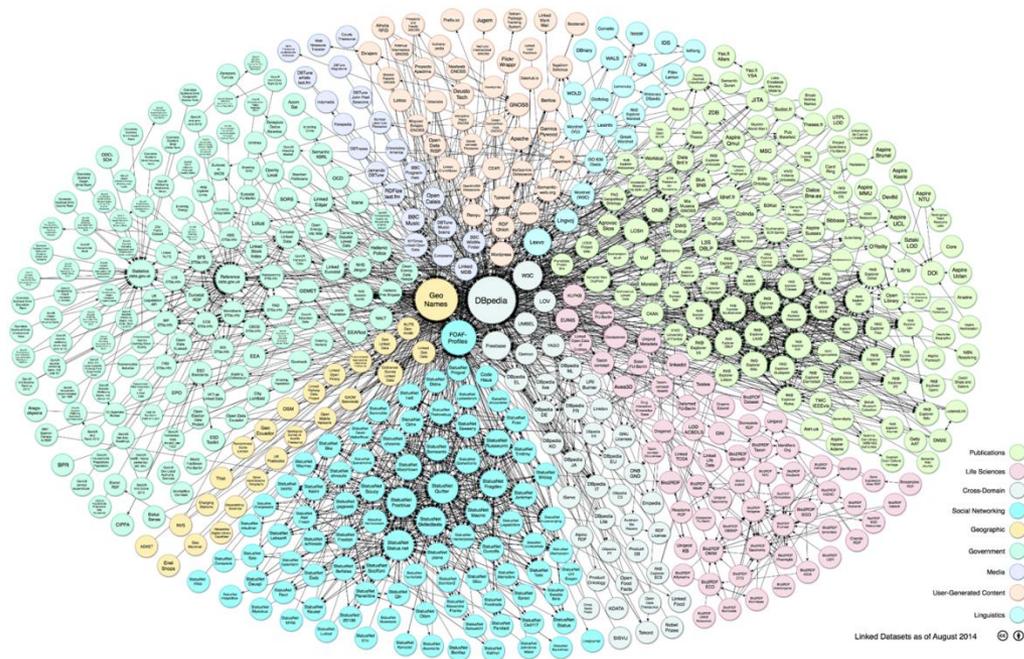
本节总结



□ 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

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

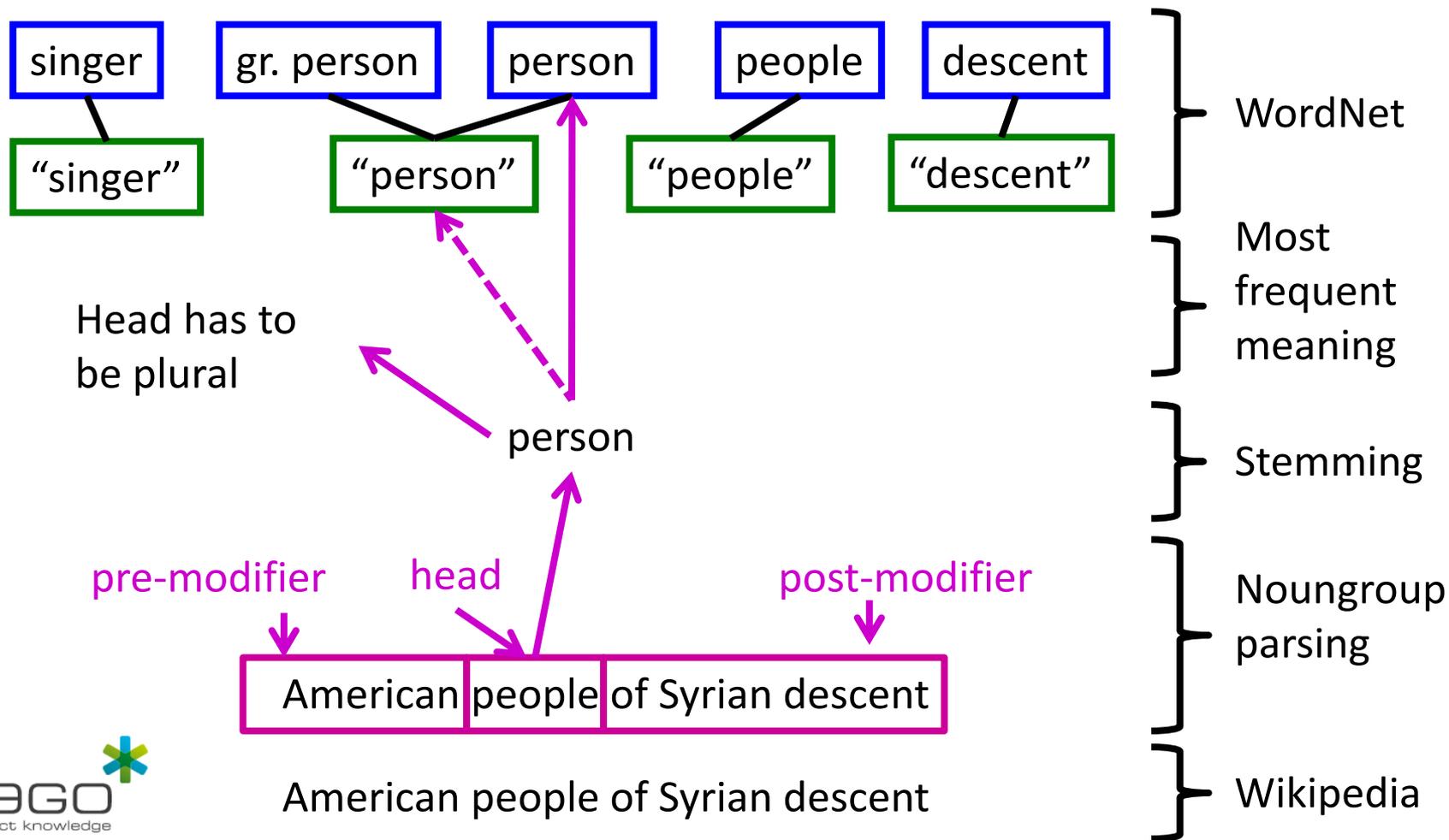
二、Beyond wiki百科资源的知识图谱构建

- 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

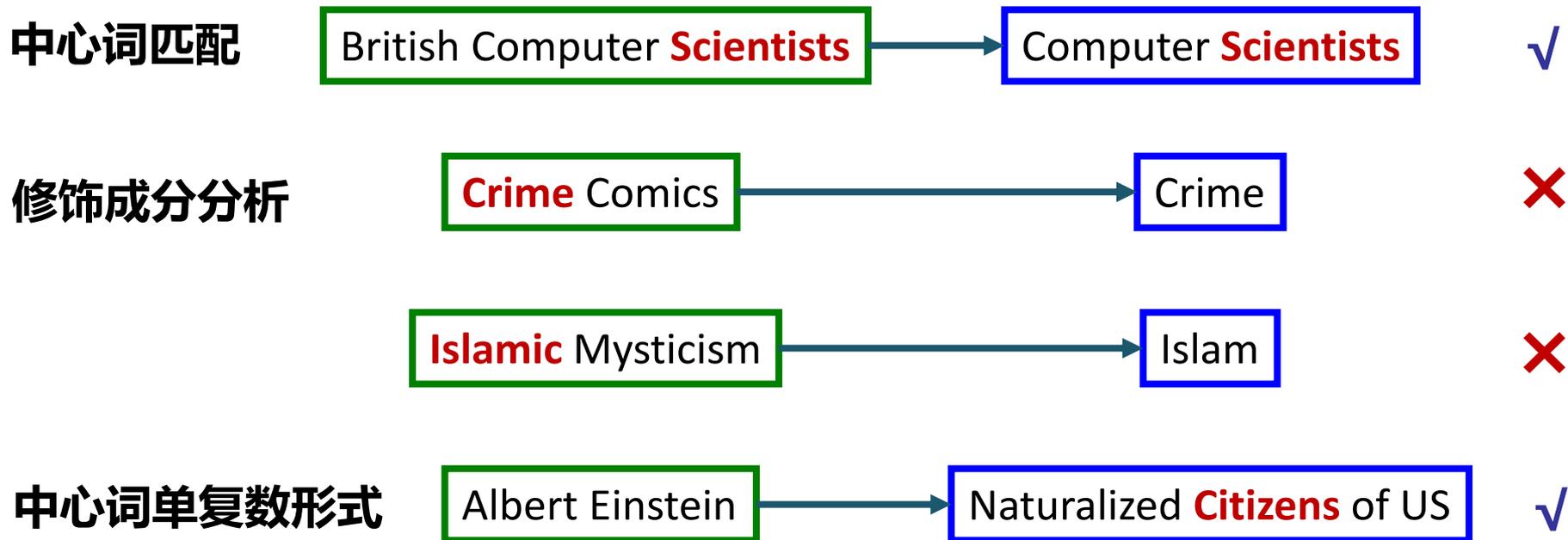


Yago: a core of semantic knowledge. Suchanek et al. WWW 07.

基于Wiki资源的Taxonomy知识抽取



识别维基百科中正确的isA关系



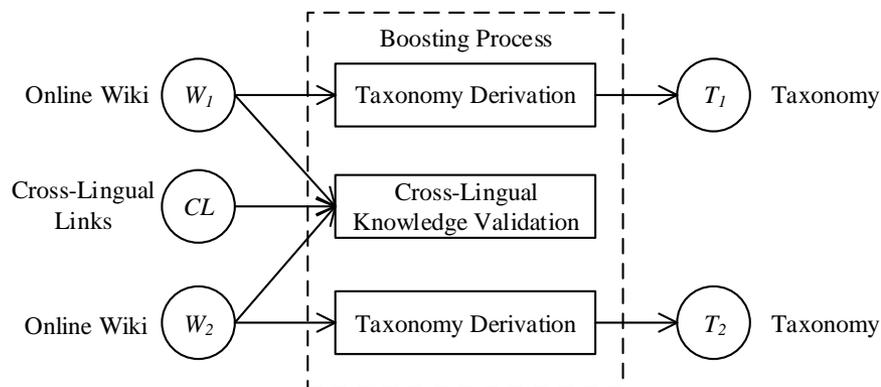
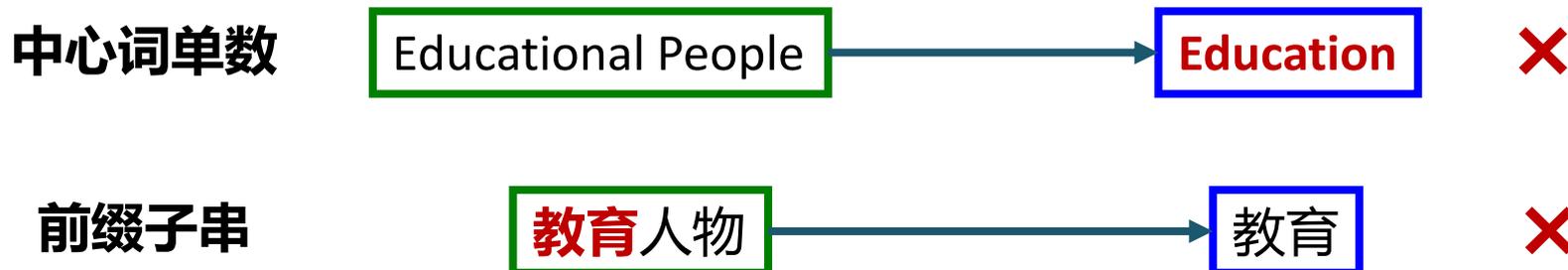
WikiTaxonomy

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



- 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



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

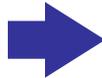
Succeeded by Kwame Raoul

Personal details

Born Barack Hussein Obama II
August 4, 1961 (age 53)
Honolulu, Hawaii, U.S.

Nationality American

Political party Democratic



dbpedia-owl:activeYearsEndDate	<ul style="list-style-type: none">2004-11-04 (xsd:date)2008-11-16 (xsd:date)
dbpedia-owl:activeYearsStartDate	<ul style="list-style-type: none">1997-01-08 (xsd:date)2005-01-03 (xsd:date)2009-01-20 (xsd:date)
dbpedia-owl:almaMater	<ul style="list-style-type: none">dbpedia:Columbia_Universitydbpedia:Occidental_Collegedbpedia:Harvard_Law_School
dbpedia-owl:birthDate	<ul style="list-style-type: none">1961-08-04 (xsd:date)
dbpedia-owl:birthPlace	<ul style="list-style-type: none">dbpedia:Hawaiidbpedia:Honolulu
dbpedia-owl:birthYear	<ul style="list-style-type: none">1961-01-01 (xsd:date)
dbpedia-owl:bnfid	<ul style="list-style-type: none">15591663c
dbpedia-owl:individualisedGnd	<ul style="list-style-type: none">132522136
dbpedia-owl:lccnId	<ul style="list-style-type: none">n/94/112934
dbpedia-owl:office	<ul style="list-style-type: none">President of the United StatesMember of the Illinois Senatefrom the 13th District
dbpedia-owl:orderInOffice	<ul style="list-style-type: none">44th
dbpedia-owl:party	<ul style="list-style-type: none">dbpedia:Democratic_Party_(United_States)
dbpedia-owl:profession	<ul style="list-style-type: none">dbpedia:Authordbpedia:Lawyerdbpedia:Community_organizingdbpedia:Professor
dbpedia-owl:region	<ul style="list-style-type: none">dbpedia:Illinois
dbpedia-owl:relation	<ul style="list-style-type: none">dbpedia:Maya_Soetoro-Ngdbpedia:Madelyn_Dunhamdbpedia:Stanley_Armour_Dunham
dbpedia-owl:religion	<ul style="list-style-type: none">dbpedia:Christianity
dbpedia-owl:residence	<ul style="list-style-type: none">dbpedia:White_Housedbpedia:Chicago
dbpedia-owl:selibrid	<ul style="list-style-type: none">314463
dbpedia-owl:seniority	<ul style="list-style-type: none">United States Senate
dbpedia-owl:successor	<ul style="list-style-type: none">dbpedia:Kwame_Raouldbpedia:Roland_Burris
dbpedia-owl:termPeriod	<ul style="list-style-type: none">dbpedia:Barack_Obama__1dbpedia:Barack_Obama__2



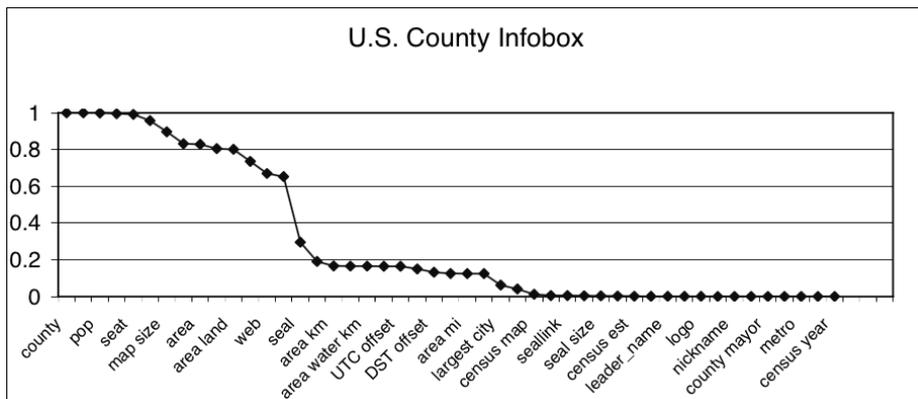
http://dbpedia.org/page/Barack_Obama

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

基于Wiki资源的AVP知识抽取



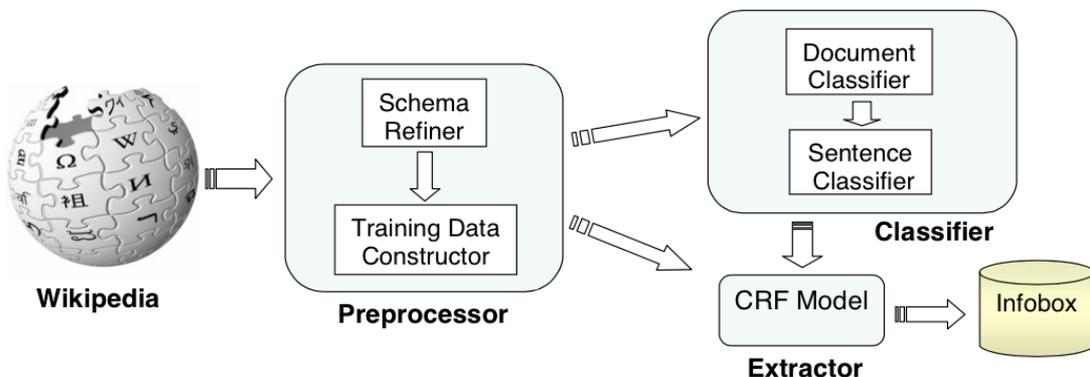
缺失信息框抽取



信息框 “US County” 属性使用比率

Feature Description	Example
First token of sentence	<i>Hello world</i>
In first half of sentence	<i>Hello world</i>
In second half of sentence	<i>Hello world</i>
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"	$TF - 1 \implies AA01$
Part of anchor text	<u>Machine Learning</u>
Beginning of anchor text	<u>Machine Learning</u>
Previous tokens (window size 5)	
Following tokens (window size 5)	
Previous token anchored	<u>Machine Learning</u>
Next token anchored	<u>Machine Learning</u>

CRF特征

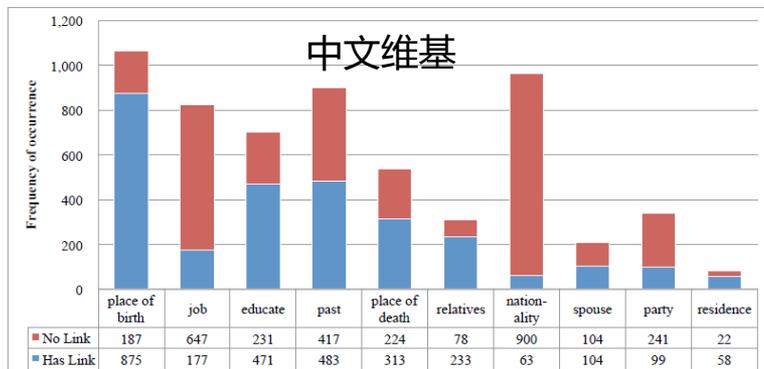
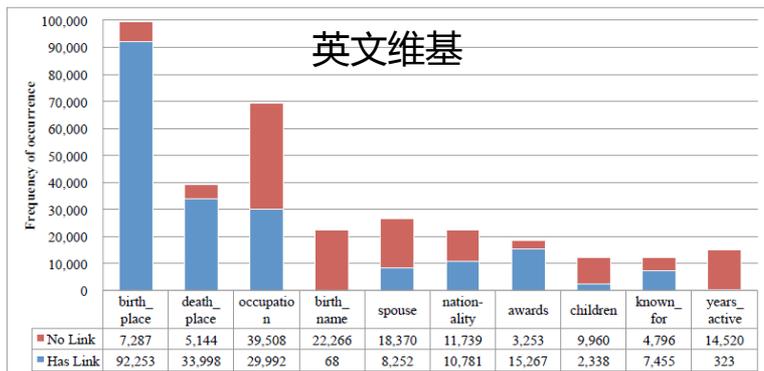


Autonomously Semantifying Wikipedia. Wu et al. CIKM 07.

基于Wiki资源的AVP知识抽取



信息框值中链接缺失



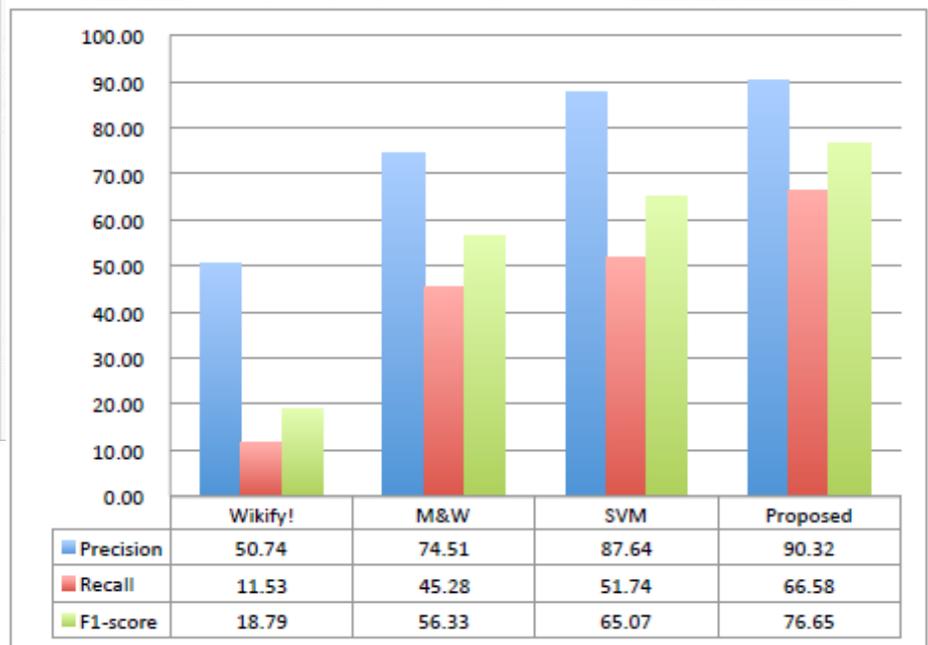
Sir Tim Berners-Lee

Berners-Lee in 2008
Born Timothy John Berners-Lee 8 June 1955 (age 57)^[1] London, England

```

{{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]]

```



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

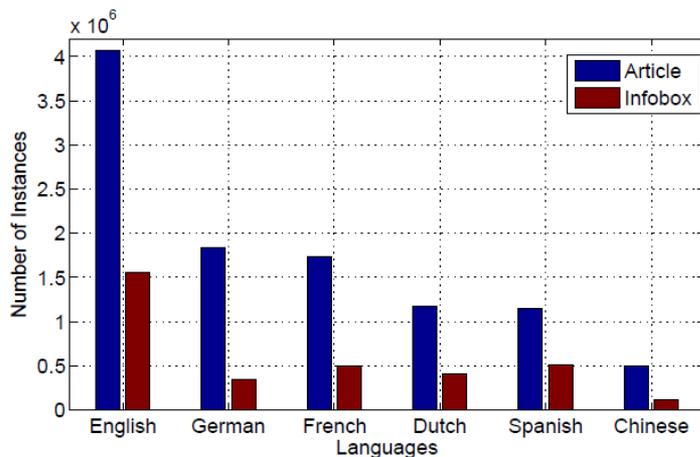


基于迁移学习的跨语言属性值抽取

- 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



Bill Gates title

From Wikipedia, the free encyclopedia

William Henry "Bill" Gates III (born October 28, 1955)^[4] is an American business magnate and philanthropist. Gates is the former chief executive and current chairman of Microsoft, the world's largest personal-computer software company, which he co-founded with Paul Allen. He is consistently ranked among the world's wealthiest people^[5] and was the wealthiest overall from 1995 to 2009, excluding 2008, when he was ranked third;^[6] in 2011 he was the wealthiest American and the second wealthiest person.^{[7][8]} During his career at Microsoft, Gates held the positions of CEO and chief software architect, and remains the largest individual shareholder, with 6.4 percent of the common stock.^[9] He has also authored or co-authored several books.

```
{ {Infobox person
|name      = Bill Gates
|birth_date   = {{ {birth date and age} 1955 | 10 | 28 }}
|birth_place = [[ Seattle ]], Washington, US
|nationality = American
|children    = 3
|alma_mater  = Harvard University (dropped out in 1975)
|...
}}
```

Categories: 1955 births | American agnostics | American billionaires | American computer businesspeople | American computer programmers | American nonprofit businesspeople | American technology chief executives | American technology company founders | American investors | American people of English descent | American people of German descent | American people of Scotch-Irish descent

ib Bill Gates



Bill Gates at the World Economic Forum in Davos, 2007

Born William Henry Gates III
October 28, 1955 (age 57)
Seattle, Washington, US

Residence Medina, Washington

Nationality American

Alma mater Harvard University (dropped out in 1975)

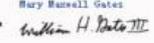
Occupation Chairman of Microsoft (non-executive)
Co-chair of Bill & Melinda Gates Foundation value
Director of Berkshire Hathaway
CEO of Cascade Investment

Net worth US\$53 billion (2010)^[1]

Spouse(s) Melinda Gates (m. 1994)

Children 3

Parents William H. Gates, Sr.
Mary Maxwell Gates

Signature 

Website microsoft.com/prnspsaz/assoc/billg

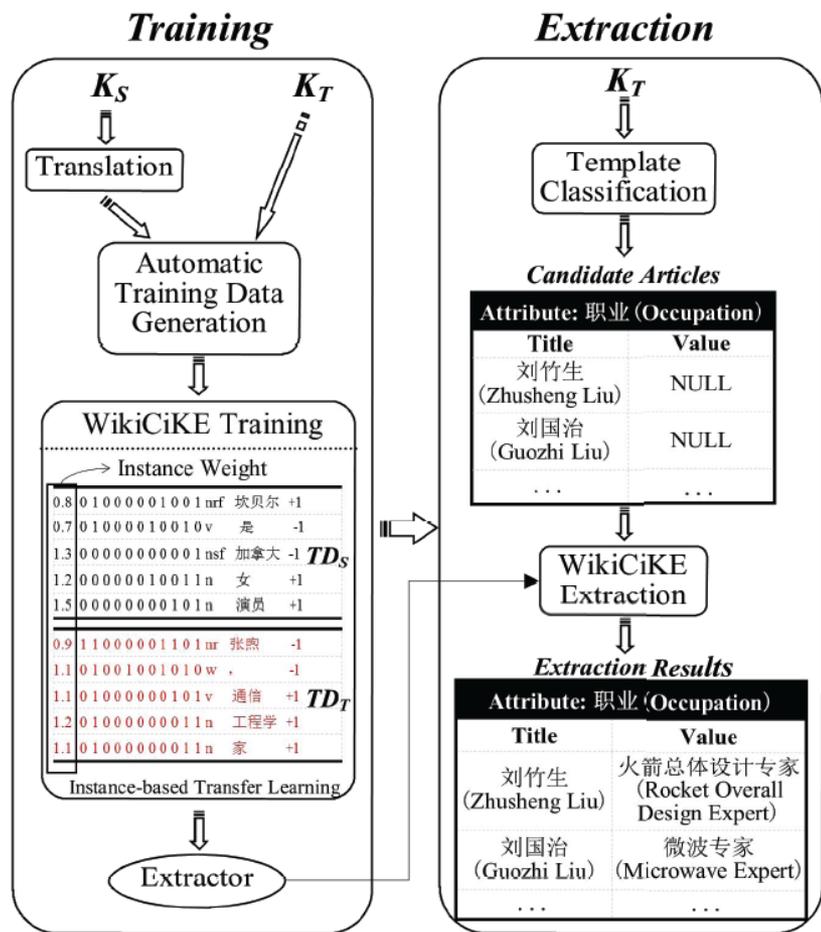
Attribute	en	zh	Attribute	en	zh
<i>name</i>	82,099	1,486	<i>awards</i>	2,310	38
<i>birth date</i>	77,850	1,481	<i>weight</i>	480	12
<i>occupation</i>	66,768	1,279	<i>influences</i>	450	6
<i>nationality</i>	20,048	730	<i>style</i>	127	1

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？

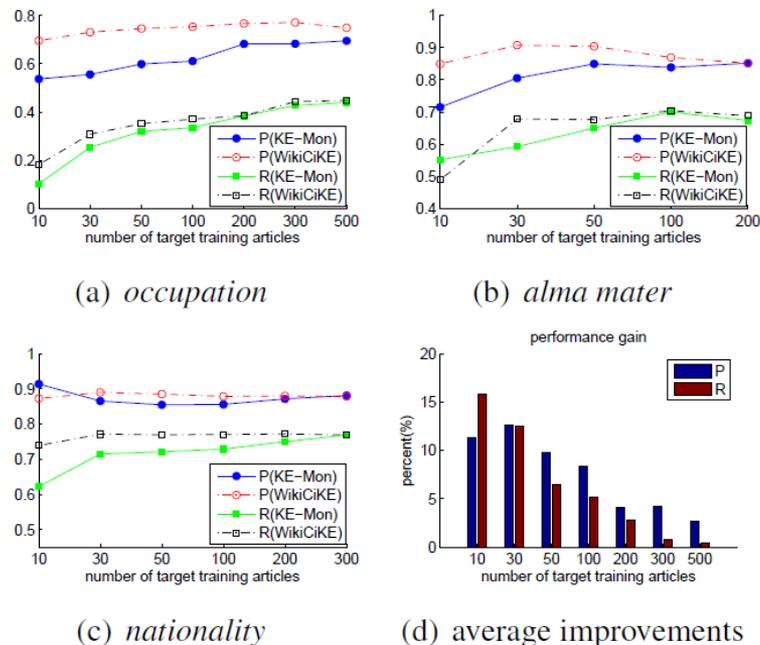
基于Wiki资源的AVP知识抽取



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



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

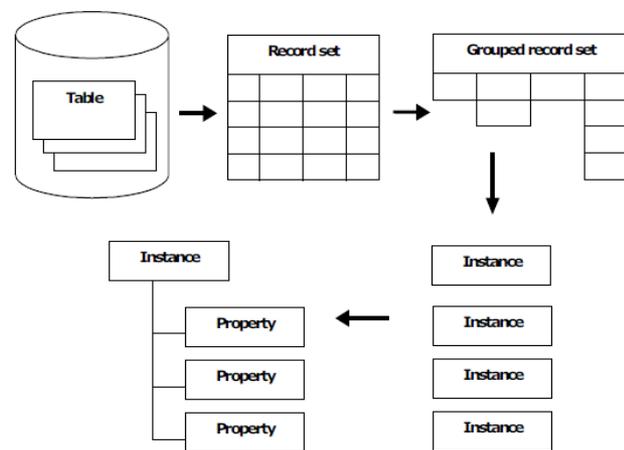


Figure 1. The D2R mapping process.

D2R映射过程

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

半结构化知识抽取

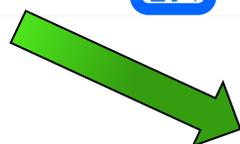
□ Taxonomy知识抽取

从Web Table中提取实例与概念之间的上下位语义关系

演职员表 编辑

演员表

角色	演员	备注
耿浩	黄渤 ^[5]	二手音响店老板, 过气歌手
郝义	徐峥	耿浩的发小, 知名制片人; 乐天派, 花花公子, 泡妞达人
康小雨	袁泉	文艺小清新的女汉子
周丽娟	周冬雨	90后小镇杀马特姑娘
莎莎	马苏	东北老妹儿 泼辣火爆的特殊服务人员
东东	陶慧	阿凡达女郎 草台班子的舞蹈演员
旅店老板	岳小军	低调的艺术大叔



□ 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

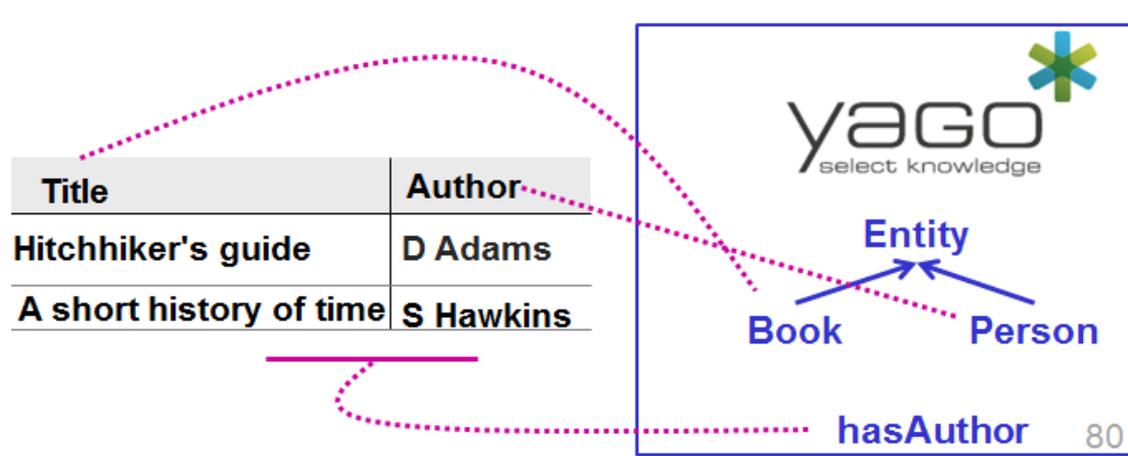


A semi-supervised method to learn and construct taxonomies using the web. Kozareva et al. EMNLP 10.

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.

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

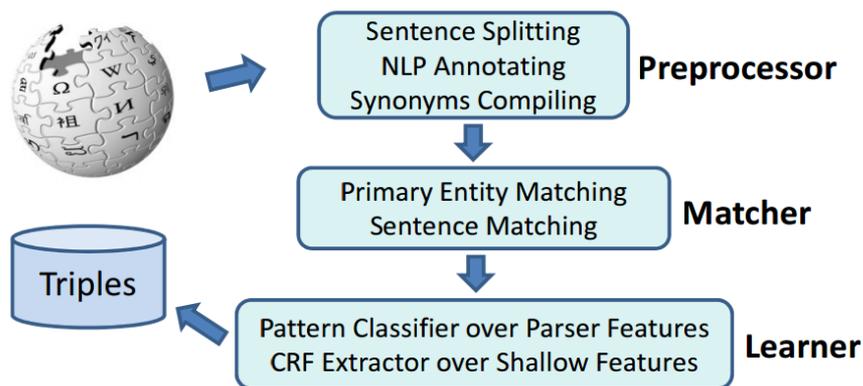


Figure 1: Architecture of WOE.

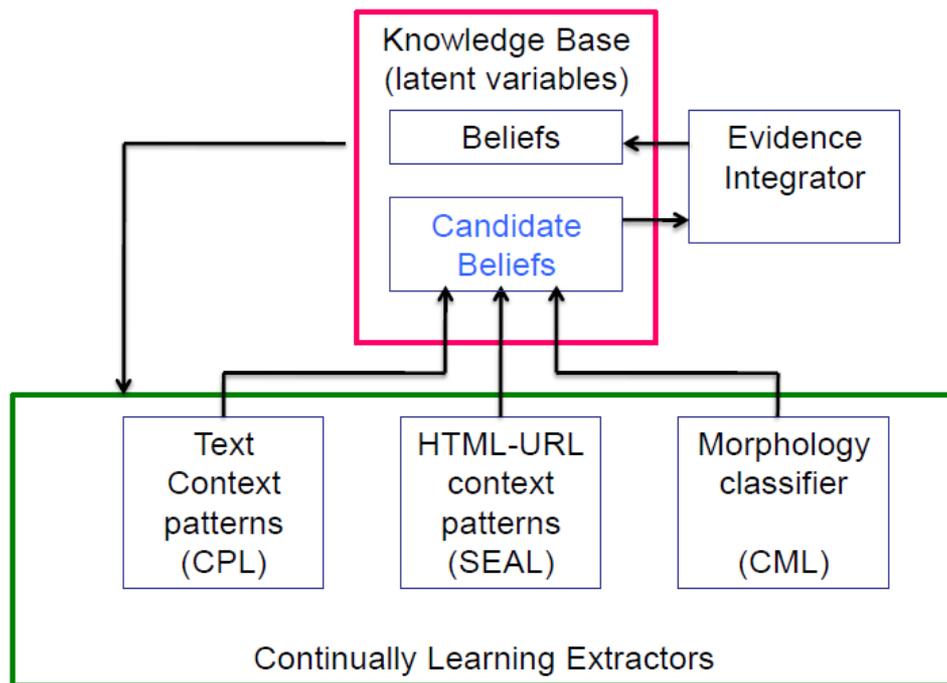
Open information extraction using Wikipedia . ACL 10

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. AAI 2010

□ Probase (A Probabilistic Knowledgebase)

■ 目标

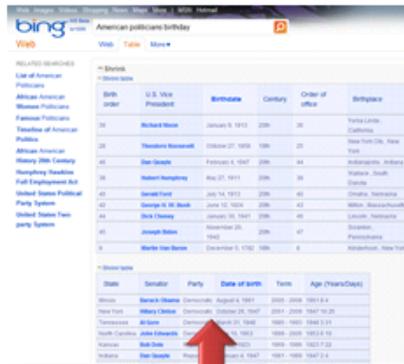
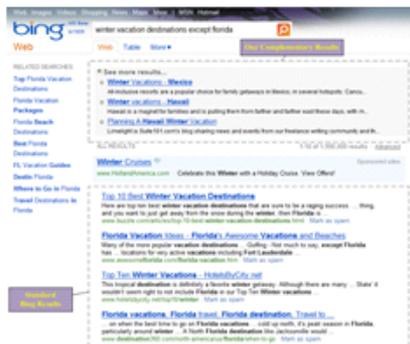
- 通过注入“一般性知识”到计算机中，来更加了解人的交流，形成知识库

■ 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



UNDERSTANDING

悟 & 懂

Terms	Entity	Attribute
China	<input checked="" type="radio"/>	<input type="radio"/>
Russia	<input checked="" type="radio"/>	<input type="radio"/>
India	<input checked="" type="radio"/>	<input type="radio"/>
...	<input checked="" type="radio"/>	<input type="radio"/>

unknown type noise tolerant

Abstract

I think you are talking about country

UNDERSTANDING

悟 & 懂

Terms	Entity	Attribute
China	<input checked="" type="radio"/>	<input type="radio"/>
Russia	<input checked="" type="radio"/>	<input type="radio"/>
India	<input checked="" type="radio"/>	<input type="radio"/>
...	<input checked="" type="radio"/>	<input type="radio"/>

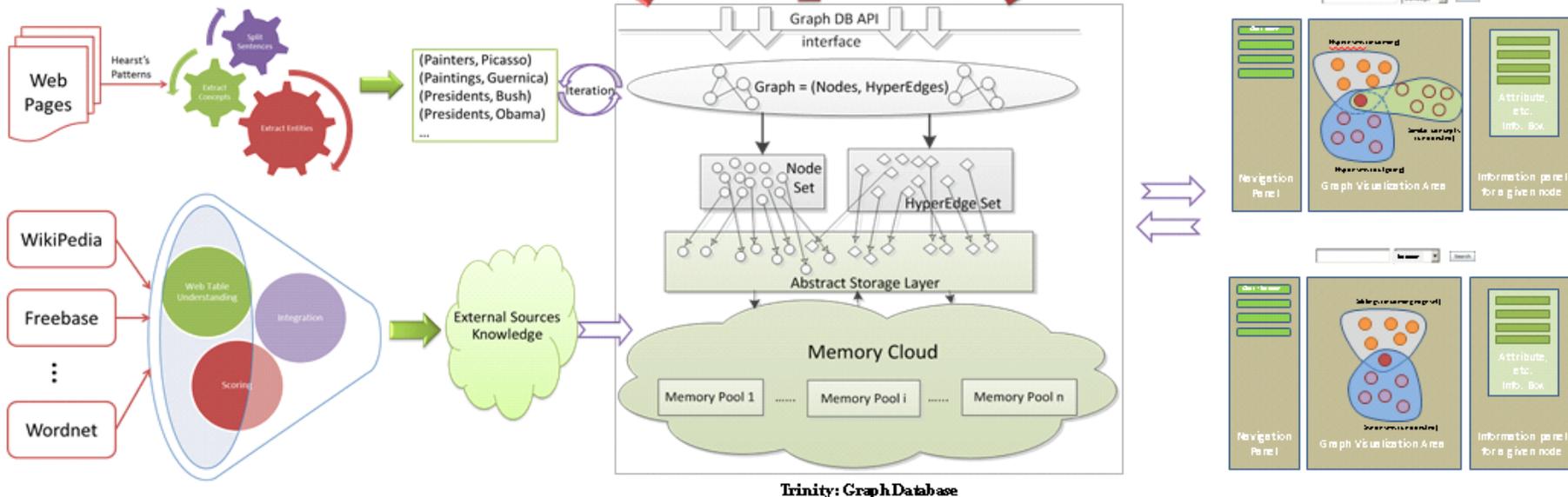
unknown type noise tolerant

Abstract

I think you are talking about country

Applications

Infrastructure

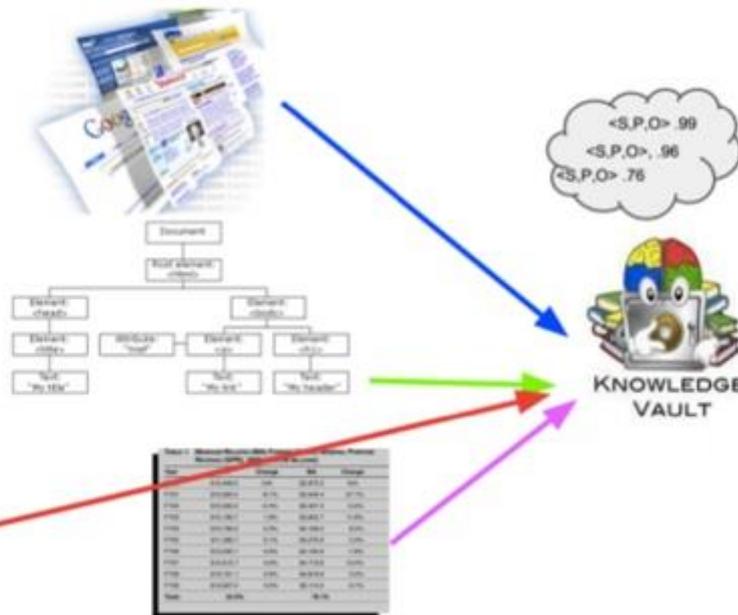
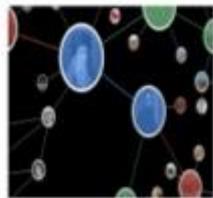


Probases系统图

□ Google Knowledge Vault

Knowledge Vault* fuses all these signals together

- Data from web
 - Unstructured text
 - Semi-structured DOM trees
 - Structured WebTables
- "Prior" data from FB



- 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

* Details in a paper submitted to WWW'14 (Dong et al)

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



图像

文本

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



Barack Obama
www.barackobama.com

[Follow](#)

4,723,528 followers | 103,921,572 views

社交主页

Barack Obama
Shared publicly - Oct 11, 2014 #With1010

"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



A higher minimum wage would give a raise to

28 MILLION AMERICANS

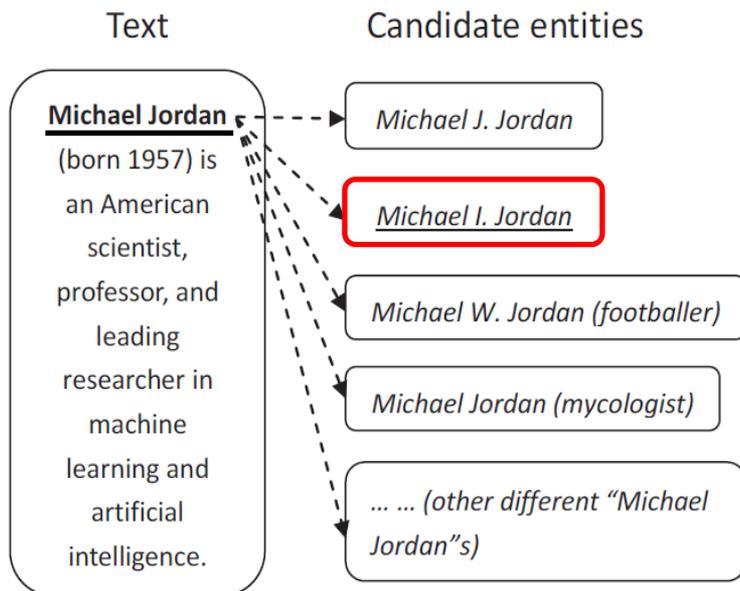
What would \$10.10 mean for you?

#With1010

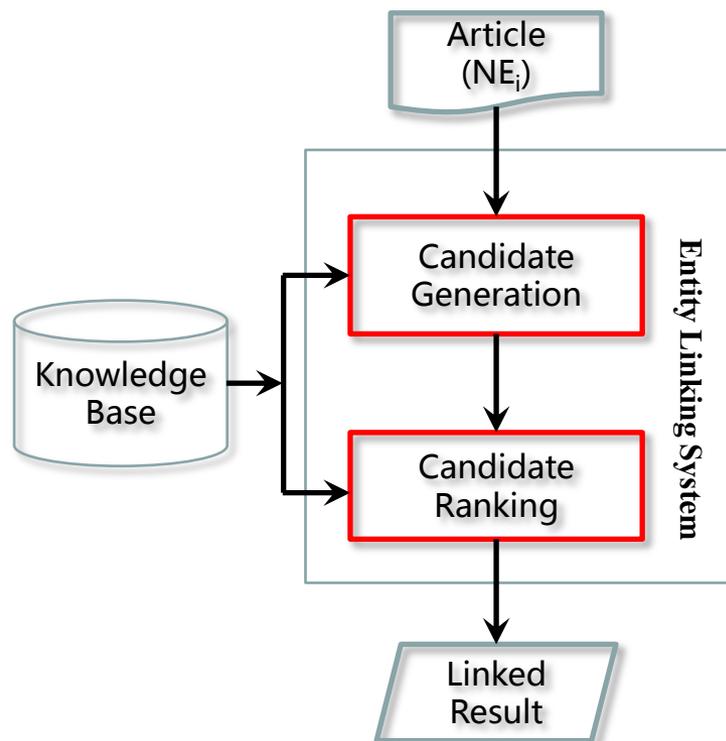
+1473 112

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 e_m 的过程

□ 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*.

□ 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.

□ 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.

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](#)

共4625张

李娜的最新相关消息

- 49分钟前 ● [退役后传递爱心 李娜姜山慈善之行送温暖|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 ● [李娜携姜山投身公益 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



李娜

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。



The screenshot shows the Baidu search interface. At the top, there are navigation links for various content types: 新闻, 网页, 贴吧, 知道, 音乐, 图片, 视频, 地图, 百科, 文库. The search bar contains the text '李娜'. Below the search bar, there are buttons for '进入词条', '搜索词条', and '帮助'. The main navigation bar includes '首页', '分类频道', '特色百科', '玩转百科', '百科用户', '百科校园', '百科合作', '手机百科', and '个人中心'. The search results section displays a message: '李娜是一个多义词, 请在下列义项中选择浏览 (共25个义项)'. Below this message, there is a grid of nine disambiguation options:

- 著名中国网球运动员
- 中国舞台导演, 舞蹈编导
- 水木年华演唱歌曲
- 流行歌手、佛门女弟子
- 中国女子跳水运动员
- 广西艺术学院教授
- 青年歌唱家
- 中国女子击剑运动员
- 河北医科大学第三医院教授

At the bottom right of the results section, there is a link for '全部展开'.

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

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。本剧由鮑
美国国务卿 - 戴维营 - 迈克尔·多布斯

纸牌屋第1季 - 搜狐视频



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.

and relation summarization

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 Search

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 Cathedral
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 died at the age of 89. His successor Lady Thatcher said he was a "political giant" and "sense the first modern Conservative leader".



how long is yangzi river



Exam

Input interpretation:

Yangtze length

Share | Search | Download | Print | A | G

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

纸牌屋



《纸牌屋》是美国一部以政治为题材的电视连续剧。本剧由鲍尔·威利蒙创作并改编自安德鲁·戴维斯主创的BBC同名电视剧。两部电视剧都是基于迈克尔·多布斯同名小说创作的。第一季的全部13集于2013年2月1日在流媒体服务商Netflix首播。第二季全部13集则于2014年2月14日同时放出。
[维基百科](#)

开播时间：2013年2月1日

网络：Netflix

语言：英语

所获奖项：皮博迪奖

编剧：迈克尔·多布斯，安德鲁·戴维斯，里克·克里夫兰，萨姆·福曼

演员表

还有5+项



凯文·史派西



罗宾·怀特



凯特·玛拉



科里·施托尔



克莉丝汀·康诺利

用户还搜索了

还有15+项



政坛野兽
2012年



丑闻(电视剧)
自2012年



铁杉树丛
自2013年



三军统帅
2005年 - 2006年



白宫群英
1999年 - 2006年

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