

中文知识图谱

—十亿节点知识图谱和google下一代搜索

复旦大学
肖仰华



Outline





What is knowledge graph?

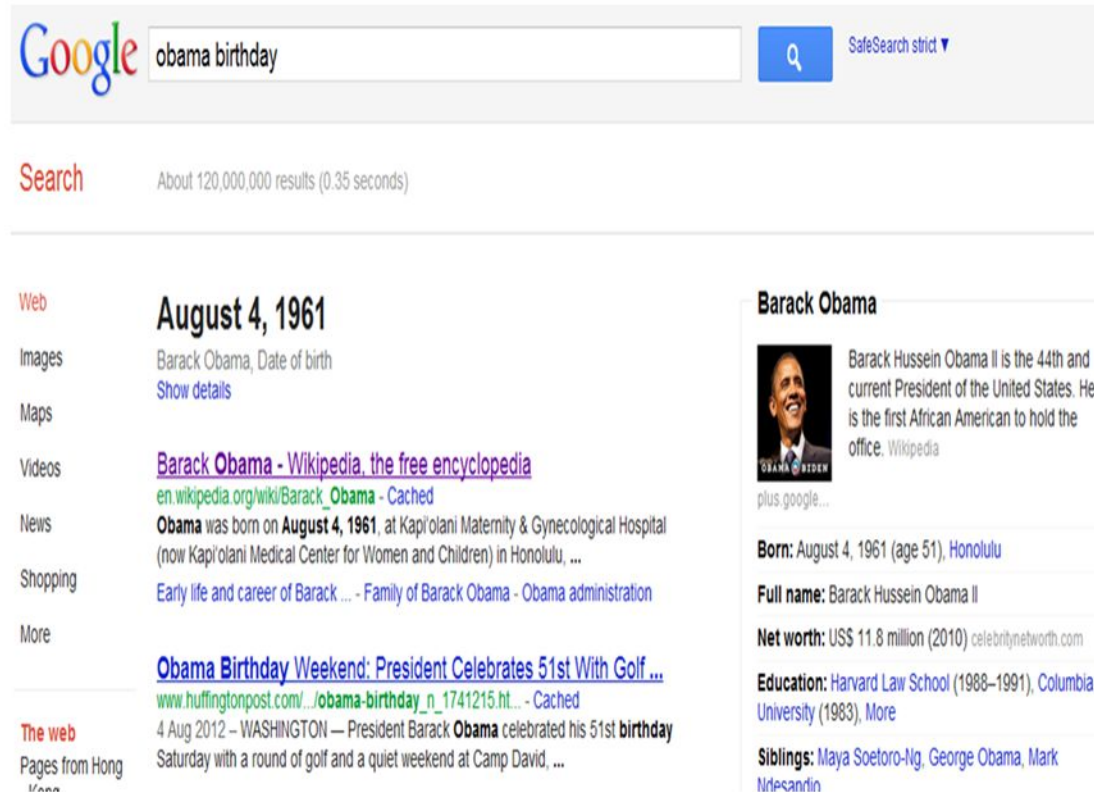
knowledge graph contains entities/concepts as vertices
and semantic relationships as edges

What makes knowledge graph different ?

- Ontology
 - Domain dependent
 - Small scale
- Semantic network
 - Focus on concepts instead of entities
- Knowledge graph
 - Large scale
 - Cover entities and concepts
 - Cover different semantic relationships

Google Knowledge Graph

- Source
 - CIA Factbook
 - Freebase
 - Wiki
- Current status
 - 500 million entities and more than 3.5 billion facts



Google search results for "obama birthday". The search bar shows "obama birthday" and the results indicate "About 120,000,000 results (0.35 seconds)".

Search About 120,000,000 results (0.35 seconds)

Web **August 4, 1961**
Barack Obama, Date of birth
[Show details](#)

Images

Maps

Videos [Barack Obama - Wikipedia, the free encyclopedia](#)
[en.wikipedia.org/wiki/Barack_Obama - Cached](#)


News **Obama** was born on **August 4, 1961**, at Kapi'olani Maternity & Gynecological Hospital (now Kapi'olani Medical Center for Women and Children) in Honolulu, ...
[Early life and career of Barack ... - Family of Barack Obama - Obama administration](#)

Shopping

More

The web [Obama Birthday Weekend: President Celebrates 51st With Golf ...](#)
[www.huffingtonpost.com/.../obama-birthday_n_1741215.ht... - Cached](#)
4 Aug 2012 – WASHINGTON — President Barack **Obama** celebrated his 51st birthday Saturday with a round of golf and a quiet weekend at Camp David, ...

Barack Obama

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

[plus.google...](#)

Born: August 4, 1961 (age 51), Honolulu

Full name: Barack Hussein Obama II

Net worth: US\$ 11.8 million (2010) [celebritynetworth.com](#)

Education: Harvard Law School (1988–1991), Columbia University (1983), [More](#)

Siblings: Maya Soetoro-Ng, George Obama, Mark Nilesandlin



1

GDM@FUDAN <http://gdm.fudan.edu.cn>

Graph Data Management Lab, School of Computer Science

More than 2.7 million concepts automatically
harnessed from 1.68 billion documents

ProBase

2

Computation/Reasoning enabled
by scoring:

Consensus:

e.g., is there a company called Apple?

Typicality:

e.g. how likely you think of Apple when
you think about companies?

Ambiguity:

e.g., does the word *Apple*, sans any
context, represent *Apple the company*?

Similarity:

e.g., how likely is an actor also a celebrity?

Freshness:

e.g., *Pluto as a dwarf planet* is a claim more
fresh than *Pluto as a planet*.

4

A little knowledge
goes a long way after
machines
acquire a
human
touch

Machines
have better
understanding
of human
world

Capture
concepts
in human
mind

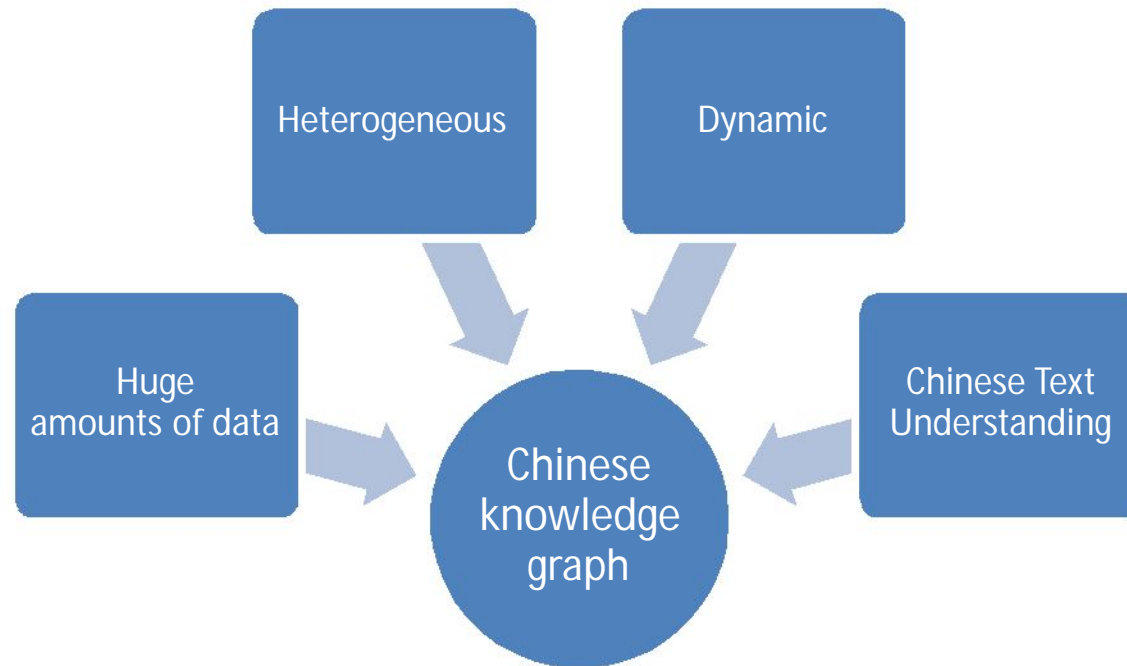
Represent
them in a
computable
form

Transform
them to
machines

3

Give machines a new CPU
(Commonsense Processing Unit)
powered by a distributed graph engine called Trinity.

Goals and Challenge



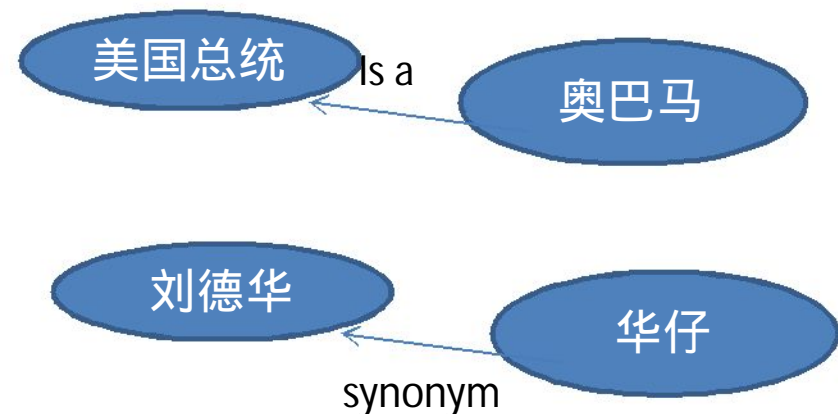
- Current status
 - 20Million concepts and entities
 - 50Million facts
- Next milestone
 - 100Million concepts/entities
 - 1Billion facts

Application 1: Semantic Search



What if we have the following information?

Entity	Attribute	Value
奥巴马	生日	****
克林顿	生日	****



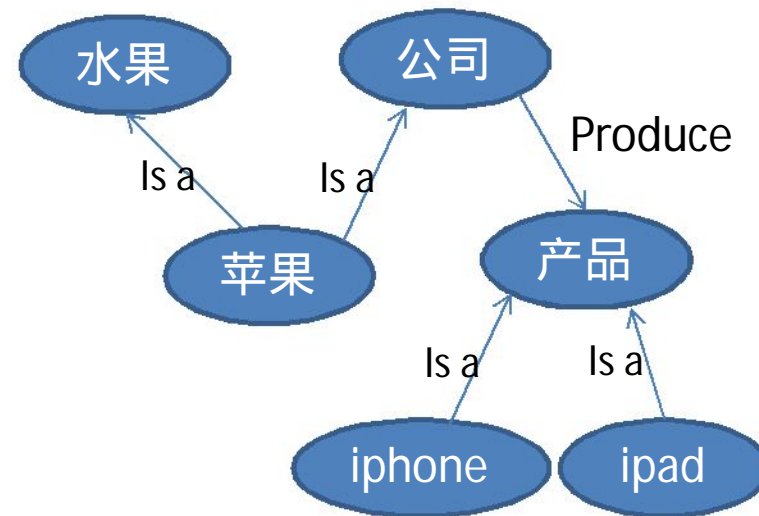
We need **semantic** search instead of **key word search** !!

Application 2: QA

- QA robot

你好，我是 机器人，有什么问题吗？
我 17:01:24 苹果公司有哪些产品
17:01:25 设计并创造了 iPod 和 iTunes、Mac 便携式和台式电脑、OSX 操作系统以及革命性的 iPhone 和 iPad。
我 17:01:33 苹果有哪些产品
17:01:34 哦~~~

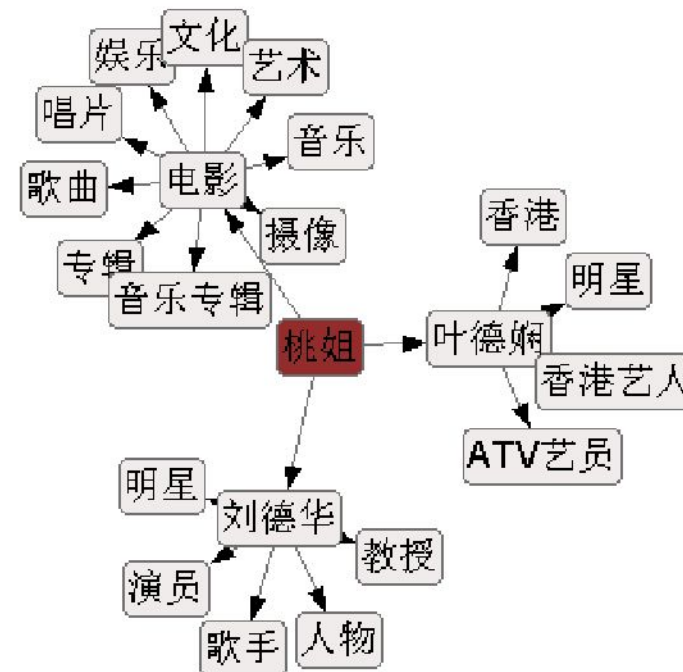
- What if we have the following information?





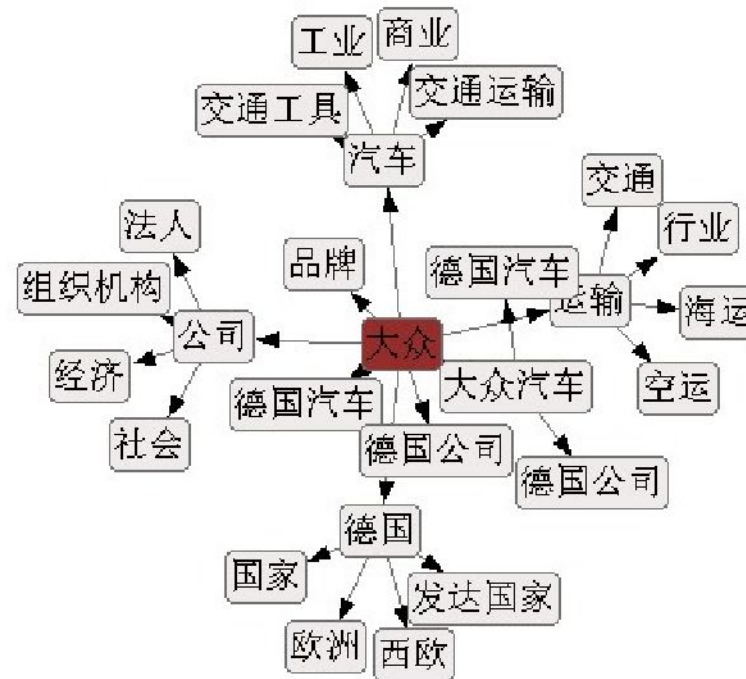
Application 3: Chinese Word Segmentation

- Traditional Chinese word segmentation tools don't know the new words
 - 《桃姐》讲述的是一个感人的故事
 - 《/wkz 桃/n 姐/n 》/wky 讲述/v 的/ude1 是/vshi 一个/mq 感人/a 的/ude1 故事/n
- Our results
 - 《/wkz 桃姐/n 》/wky 讲述/v 的/ude1 是/vshi 一个/mq 感人/a 的/ude1 故事/n
- Our latest evaluation shows our accuracy is more than 90% on open domain



Application 4. vertical search

- What key words should you propose for vertical Search?
 - Social CRM
 - 輿情监控



Application 5. E-book Reading

奥巴马2009年就职演说

拷贝

高亮

分享

笔记

搜索

我今天站在这里，因为面前的任务而感到谦卑，因为你们的信任而心存感激，同时铭记先辈们做所出的巨大牺牲。感谢布什总统为这个国家做出的贡献，同时也谢谢他在整个政权交接期间表现出的慷慨与合作。

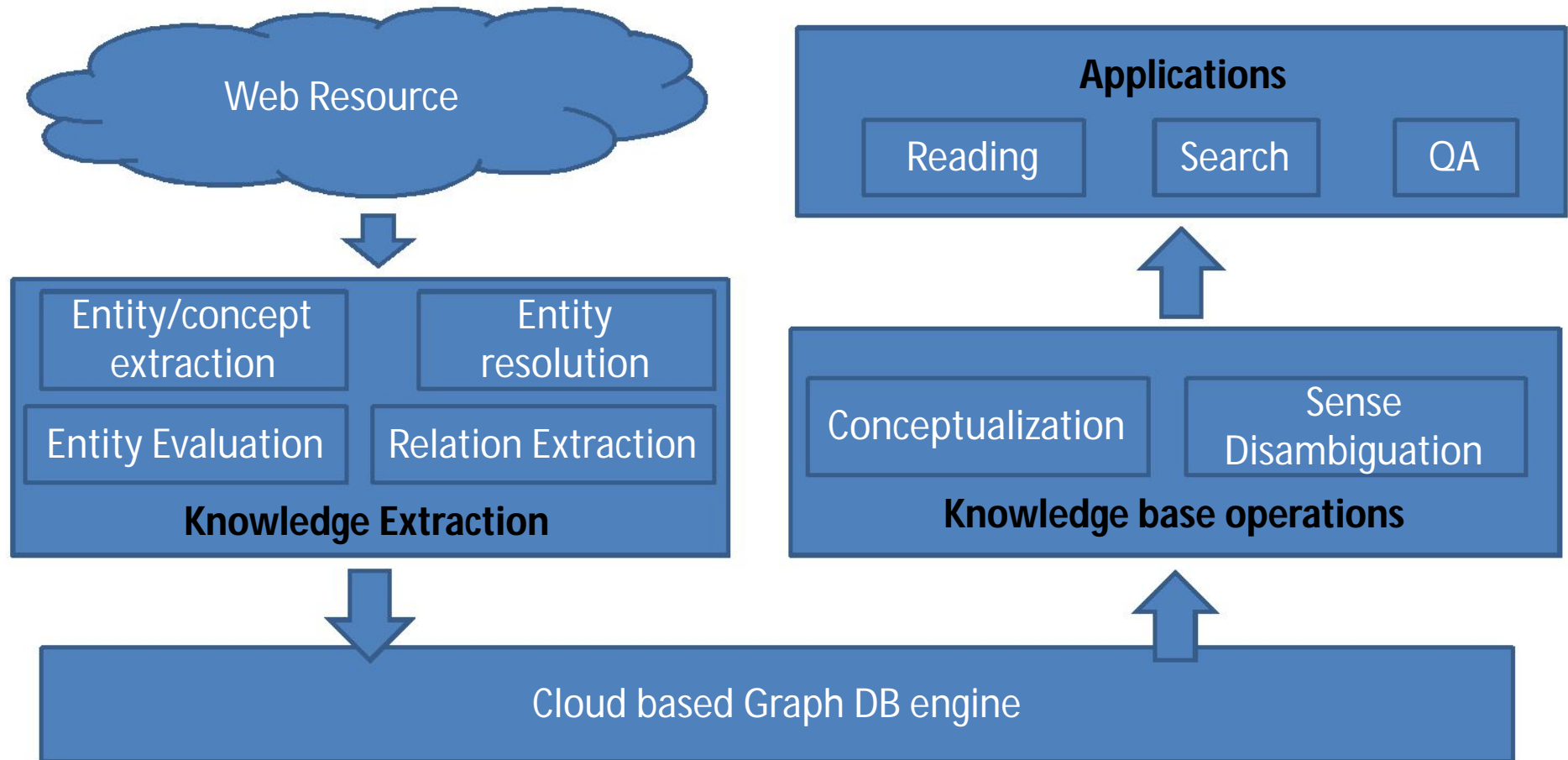
迄今已经有44名美国人宣誓就任总统。这些誓词曾出现在繁荣的上升趋势和如水般平静的和平中，当然，也经常会出现乌云密布和狂风暴雨之时。在这各种时刻，美国一直在继续前行，这不仅仅是因为执政的技巧或者有先见之明，而是因为我们的人民一直在坚守先辈们的理想，忠实履行我们的建国宣言。过去是这样，这一代的美国人仍将会坚持这样做。

页码 1 / 13 - 章节 4 / 34 - 中文



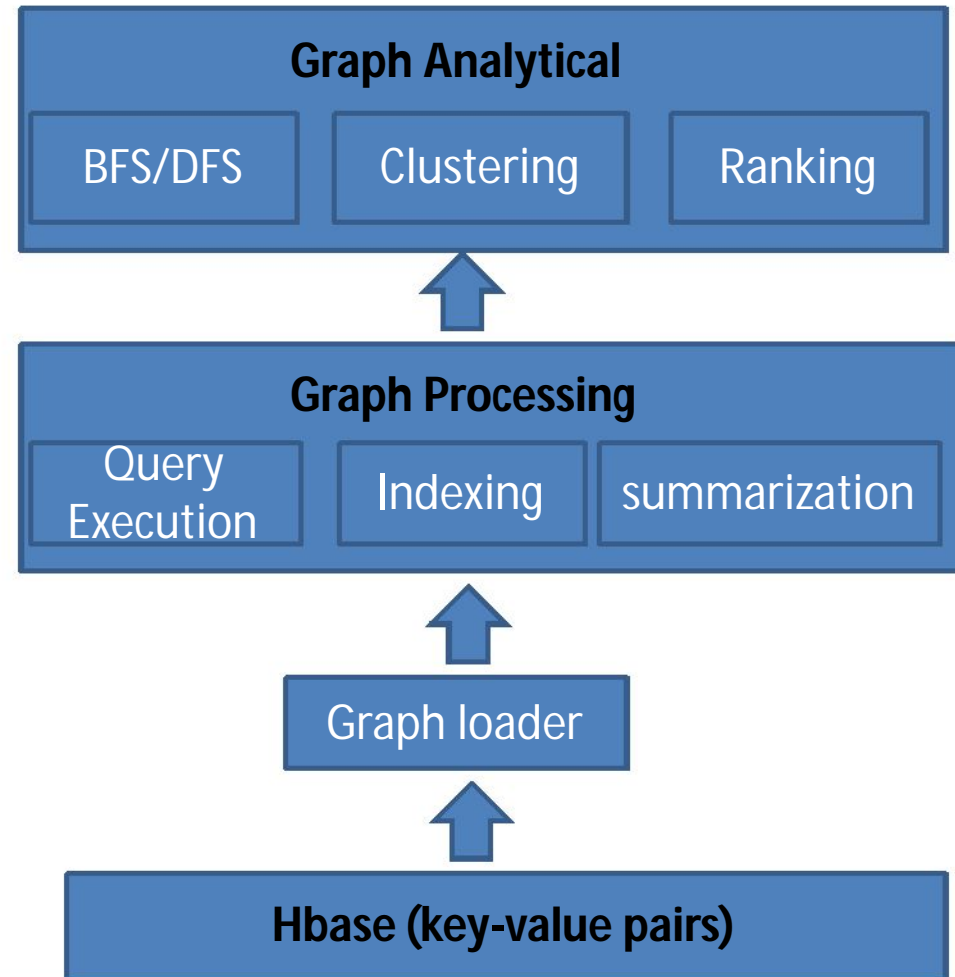
美国第44任总统 <div>简介</div>	
	
巴拉克·侯赛因·奥巴马二世 ， 美国第44任总统 ，出生于 美国夏威夷州火奴鲁鲁 ，祖籍为 古巴 （The Republic of Kenya）。奥巴马是首位拥有黑人血统，并且曾在非洲成长的美国总统。与不同种族与不同文化背景的人共同生活过。2010年3月27日美国官方发布了“国家安全战略报告”。奥巴马在该报告中将军事作为外交努力无效的最后手段。普国家安全战略认为世界充满了多种威胁，放弃了布什政府“反恐战争”的说法。2012年3月25日，抵韩并参加首尔核安全峰会的奥巴马访问了朝鲜非军事区，这是两位领袖31年来与朝鲜半岛关系的一次访问。	
Infobox:	
中文名	贝拉克·侯赛因·奥巴马
外文名	Barack Hussein Obama II
别号	奥巴马
国籍	美国
出生地	美国夏威夷
出生日期	1961年8月4日
职业	总统、参议员
毕业院校	哥伦比亚大学、哈佛大學
信仰	新教
主要成就	1996年当选伊利诺伊州参议员 2009年当选美国第三十三任总统 2009年获得诺贝尔和平奖
代表作品	总统就职演说《美国的变革时代已到来》
所属政党	美国民主党
专业	政治学和国际关系（哥伦比亚大学）
专业	法学博士（哈佛大学）

Framework



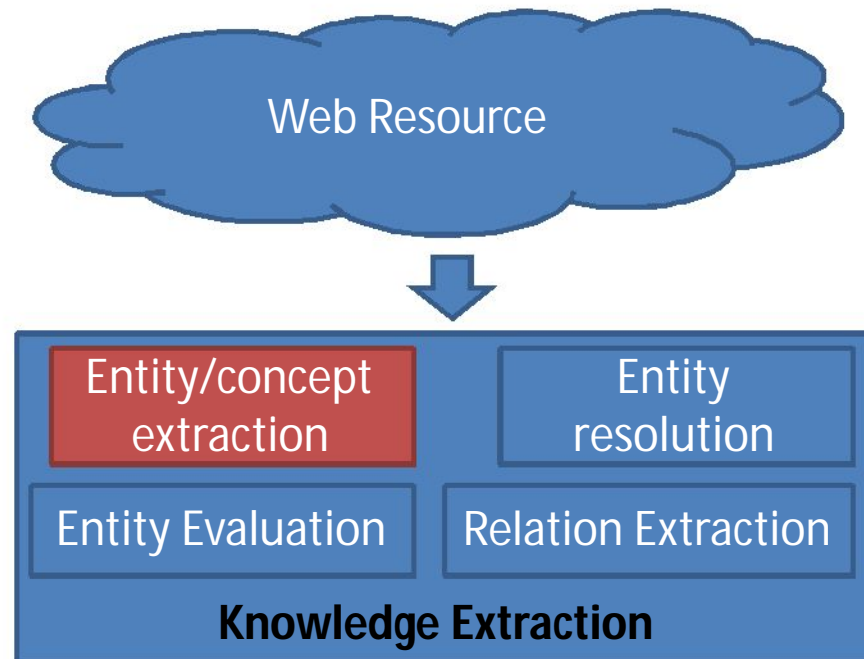
Cloud based Graph DB engine

- A graph db system based on Hbase
 - Support big knowledge graph data
 - Billion -node graphs



Entity Extraction

- Extract Chinese entities and concepts
 - 100 millions
 - High quality
 - With frequency
 - To support various applications
- Basic solution
 - Use some high quality entities to construct core entity/concept sets
 - Bootstrapping enrichment



Construct core entity/concept sets

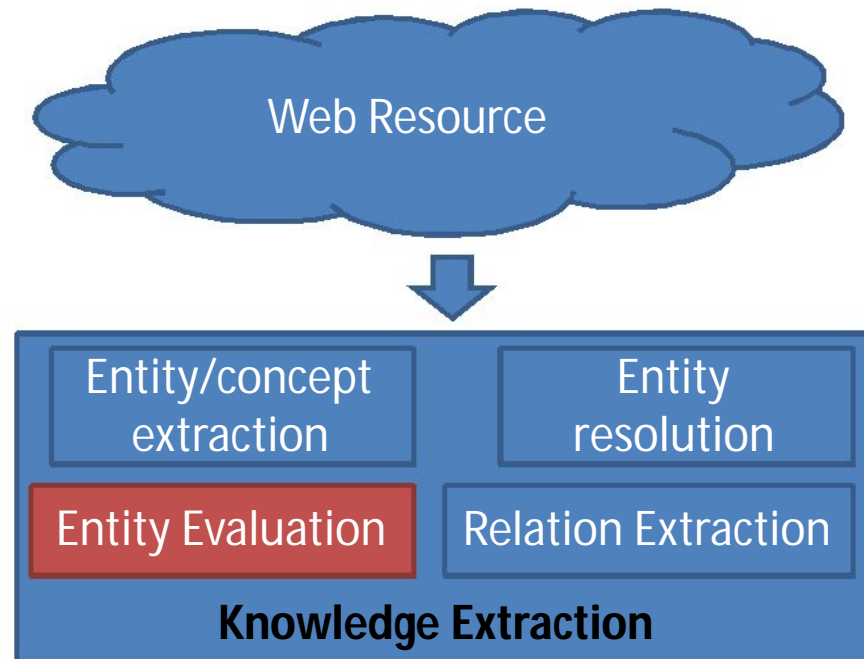
- Harvest high quality entities
 - human annotated entities
 - online encyclopedia
 - online geographic location information
 - Chinese input dictionary

Bootstrapping enrichment

- Two main method
 - Enumeration pattern or punctuation pattern in Chinese to extract more high-quality entities
 - *Segment and tag*
 - 我/r 国/n 多数/a 地区/n 先后/d 遭遇/v 大风/n 降温/vn 天气/n 过程/n
 - *DUN*
 - 菠萝、杨梅、葡萄、香蕉、苹果、梨、荔枝、柠檬等水果
 - *Punctuation 《》*
 - 《三国演义》
 - Extract entities/concepts from web structured information
 - *HTML Tables*

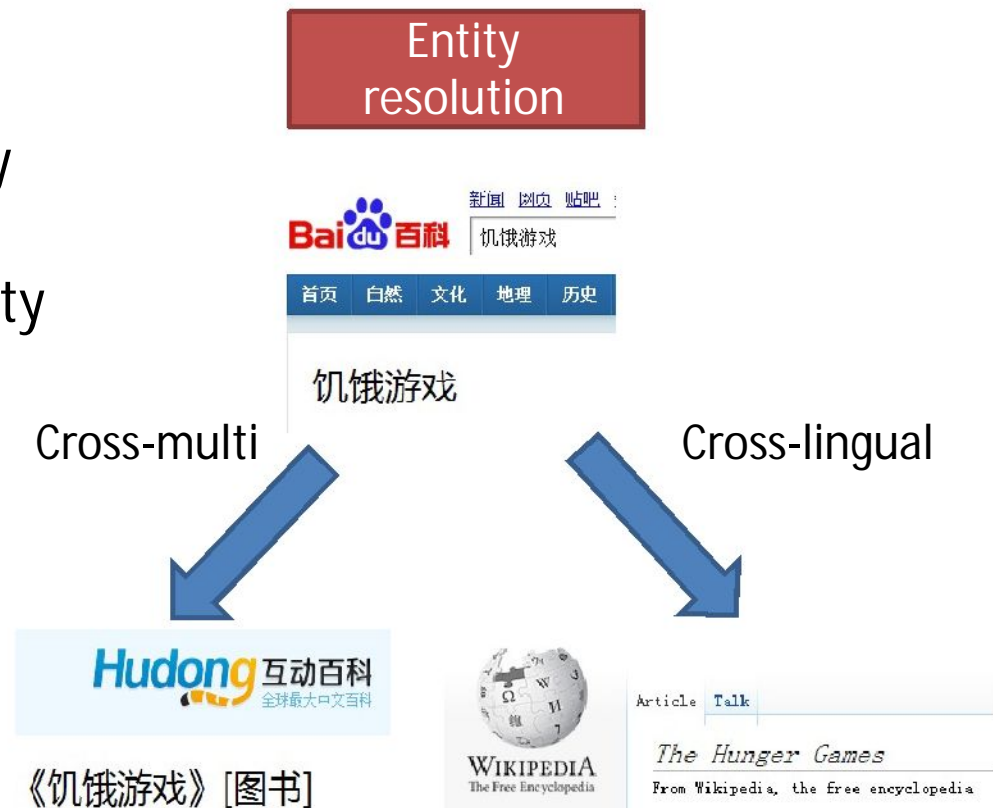
Entity Evaluation

- Evaluation Chinese entities/concepts quality
 - Precise
 - Above 90%
- Solution
 - Build the quality index
 - TF/IDF
 - Measure the specificity of an entity



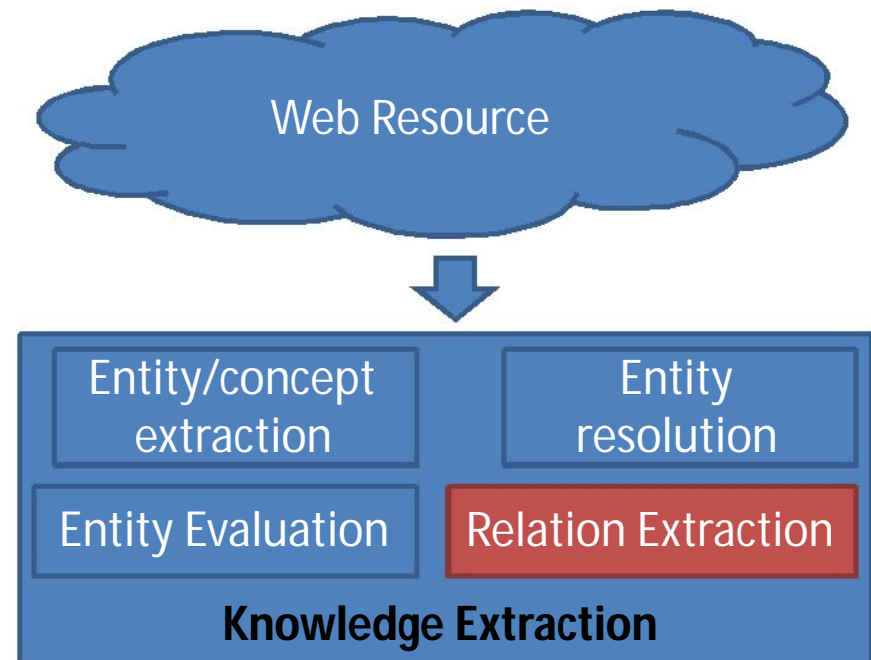
Entity Resolution

- Two types of Entity Resolution problem
 - Cross-multi source entity resolution
 - Cross-lingual source entity resolution
- Solution
 - Build a heterogeneous information network
 - Extract features and integrate them into a factor graph model



Relation Extraction

- Two types of Relation Extraction Problem
 - Typed Relation Extraction
 - Untyped Relation Extraction



Typed relationship extraction

- Is -a relationship extraction
- Solution framework
 - Using baidu open category to construct the core data set
 - Use syntax patterns to extract the isa candidate pairs
 - Use a bootstrapping procedure
 - to filter noisy isa pair
 - To expand isa pair

Patterns

1. NP1 是一 NP2
2. (NP_i)*NP2(和NP3)?等 NP4
3. NP1[属于|是|为|指]NP2的一[种|类|个]
4. NP1[是|为|指]NP2之一
5. NP1[这|那][种|类|些]NP2
6. NP1[有|含|含有|包括|包含|如|例如|像]NP2等
7. NP1分为(NP_i)*NP3[数词][种|类]

Untyped relationship extraction

- Extract undefined relationship from
 - Web texts
 - We tables
- Entity pair relationship
 - 俄罗斯，中国，接壤关系
- Entity-attribute relationship
 - 俄罗斯民族(民族，人口，百分比，宗教信仰)

俄罗斯

 编辑词条

百科名片



俄罗斯联邦，简称俄罗斯或俄联邦，是世界上领土面积最大的国家，地域跨越欧亚两个大洲，绵延的海岸线从北冰洋一直伸展到北太平洋，还包括了内陆海黑海和里海。与中国、蒙古、朝鲜等国接壤，同时，俄罗斯还与日本、美国、加拿大等国隔海相望。为前苏联的主要加盟共和国，1991年，苏联解体，俄罗斯继承苏联成为联合国安全理事会常任理事国，对安理会议案拥有否决权。俄罗斯是仅次于美国的世界第二军事强国。亦是在世界范围内有巨大影响力的世界性强国，是G8成员国之一。已成为全球最大的天然气出口国及OPEC以外最大的原油输出国。

2010年俄罗斯主要民族数量

民族	人口(万)	百分比	宗教信仰
俄罗斯族	11100	80%	东正教
鞑靼族	531	3.83%	伊斯兰教、东正教
乌克兰族	193	1.39%	东正教、天主教、联合教派
巴什基尔族	160	1.15%	伊斯兰教(逊尼派，下同)
楚瓦什族	160	1.15%	东正教
车臣族	130	0.94%	伊斯兰教
达格斯坦族	110	0.78%	基督教新教派、东正教



Demo

- <http://gdm.fudan.edu.cn/Soso/index.jsp>

Billion -node graph processing

- 中国计算机学会《学科前沿讲习班》
 - The CCF Advanced Disciplines Lectures
 - 主题 图数据管理和挖掘
 - 2012年11月16-17日 北京
 - My talk: 大图数据管理中的关键算法研究
- “云海会”报告专场
 - 时间：11/23（周五）13:30~17:00
 - 地点:SAP中国研究院（张江）。



Thanks for your attentions!!

QA