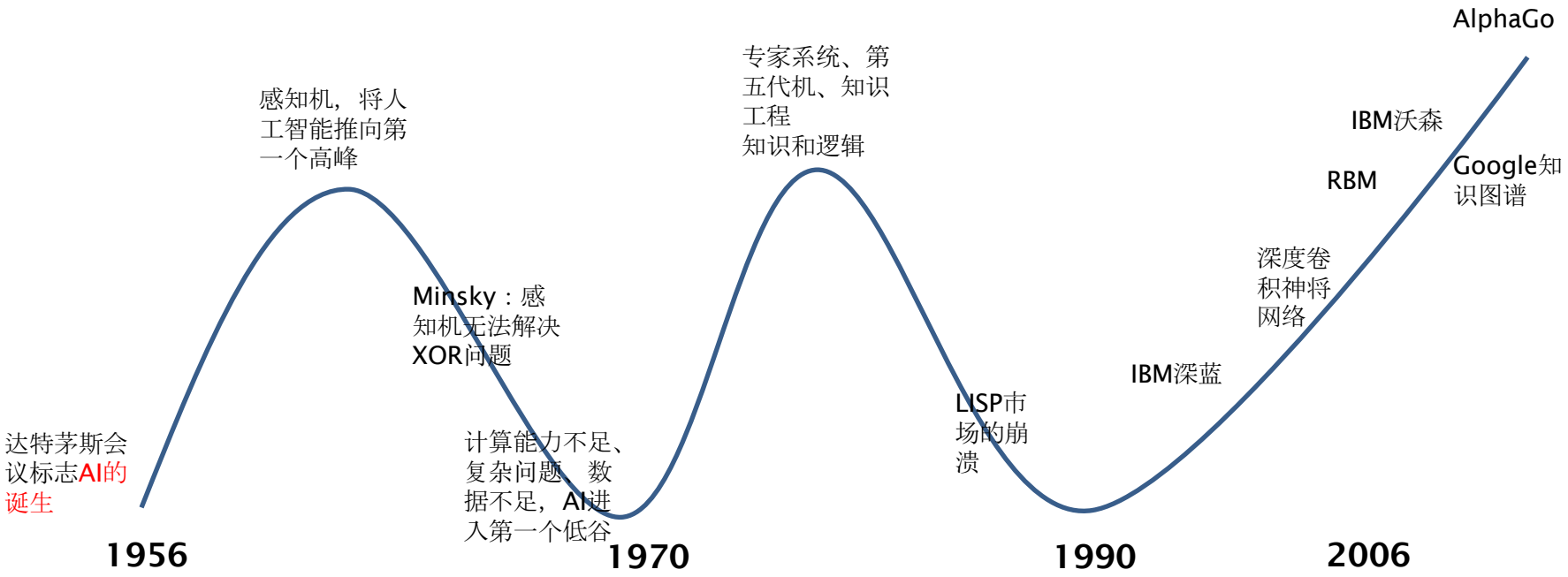
2.5知识图谱表示法

- 知识图谱发展历史与现有应用
- 知识图谱基本概念
- 知识图谱的生命周期
- 代表性知识图谱

目录

人工智能的发展历史

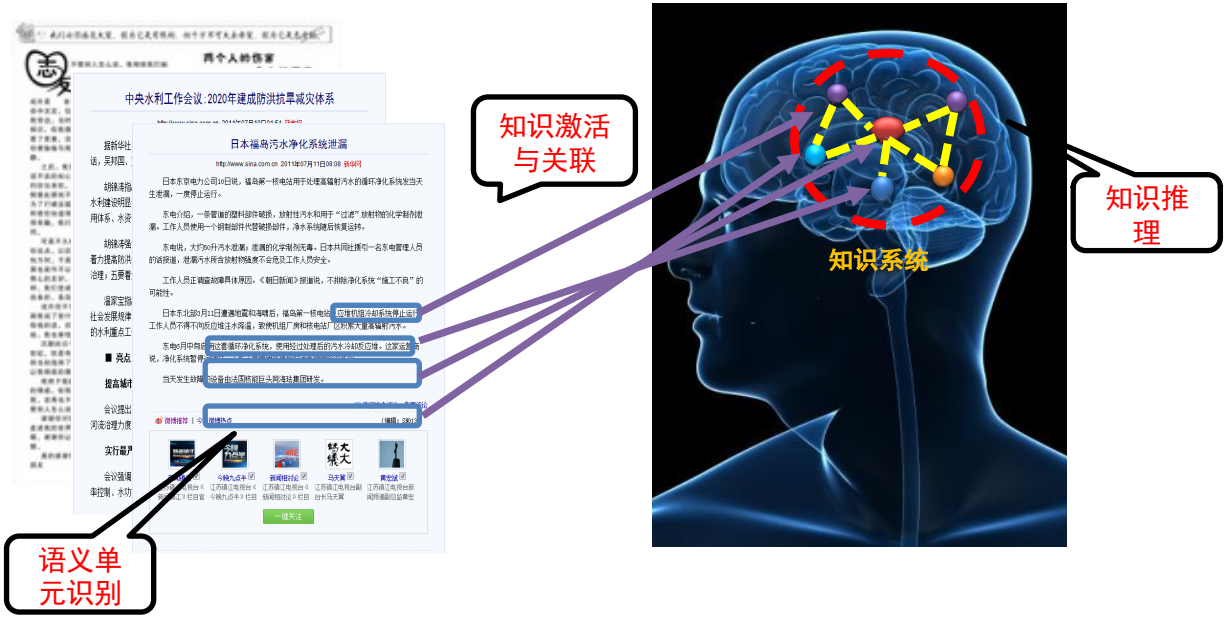

AlphaGo
感知机，将人工智能推向第一个高峰
专家系统、第五代机、知识工程
知识和逻辑
IBM沃森
RBM
Google知识图谱
深度卷积神将网络
Minsky：感知机无法解决XOR问题
IBM深蓝
LISP市场的崩溃
达特茅斯会议标志AI的诞生
计算能力不足、复杂问题、数据不足，AI进入第一个低谷
1956
1970
1990
2006

知识就是力量

Knowledge is power, and the computer is an amplifier of that power. We are now at the dawn of a new computer revolution... Knowledge itself is to become the new wealth of nations.

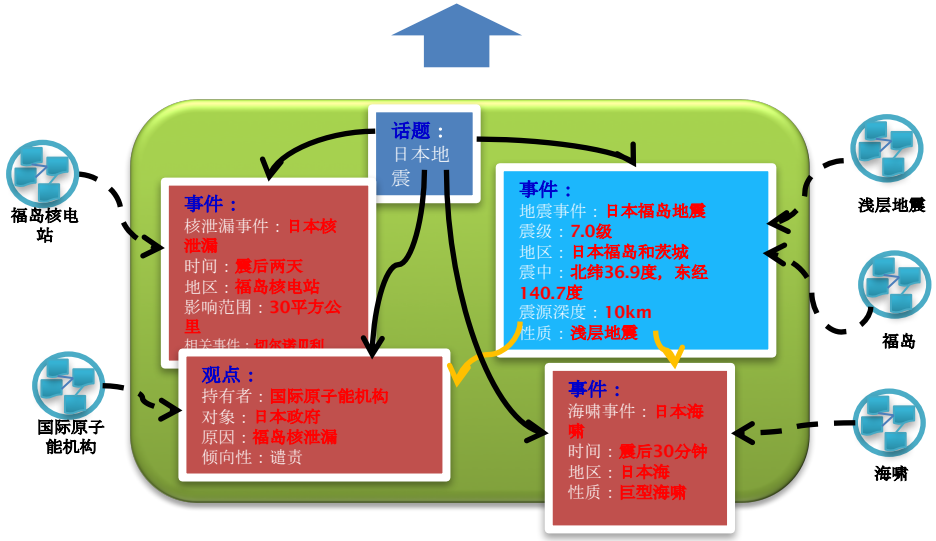
-Edward Feigenbaum (专家系统之父，1994年图灵奖)

深度自然语言理解需要知识的支撑



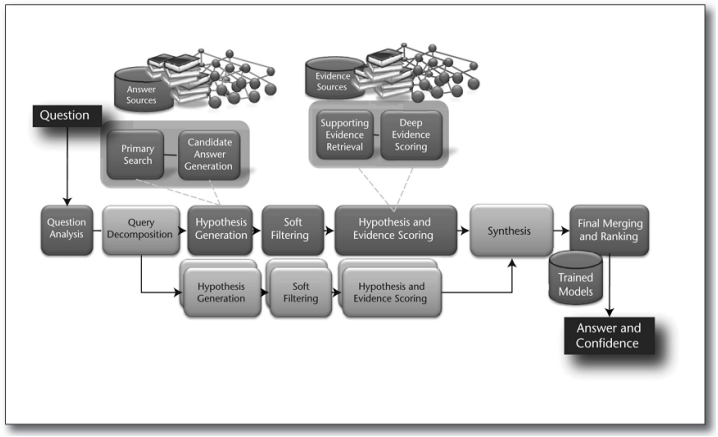
深度自然语言理解需要知识的支撑

2011年4月11日17点16分，日本东北部的福岛和茨城地区发生里氏7.0级强烈地震（震中北纬36.9度、东经140.7度，即福岛西南30公里左右的地方，震源深度10公里，属于浅层地震）。当局已经发布海啸预警。震后约30分钟后在日本海地区发生巨型海啸，同时造成福岛核电站出现核泄漏。震后第十天，国际原子能机构对于日本政府反应迟钝进行了谴责。



IBM Watson

- 沃森(Watson)：2011年，IBM研发的超级计算机“沃森”在美国知识竞赛节目《危险边缘Jeopardy!》中上演“人机问答大战”，战胜人类选手Ken和Brad



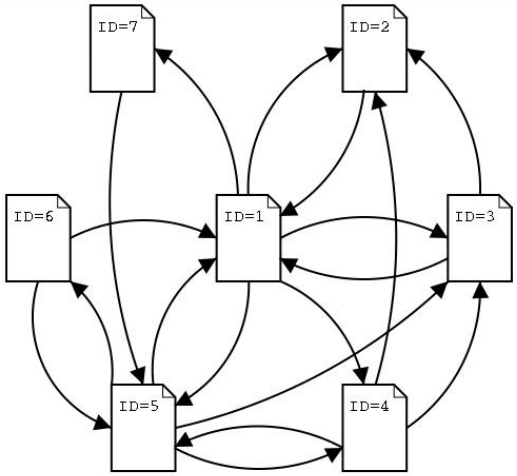
辅助医疗

金融辅助决策

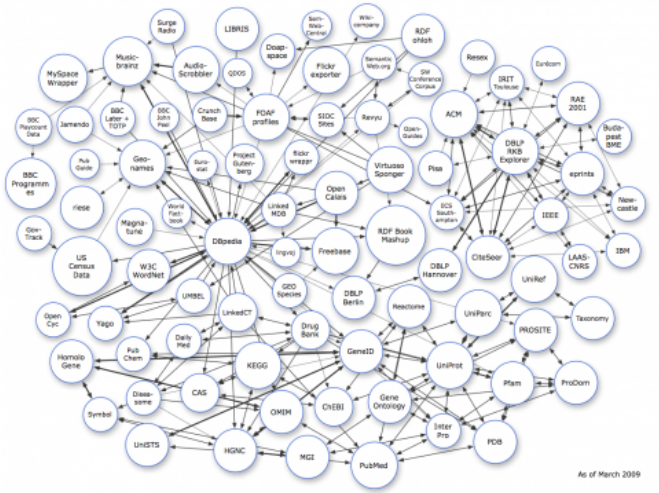
企业服务

Semantic Web

- Tim Berners-Lee 1998年提出语义网的概念
  - 通过给全球信息网上的文档（如：标准通用标记语言下的一个应用HTML）添加能够被计算器所理解的语义“元数据”（Meta data），从而使整个互联网成为一个通用的信息交换媒介



Page/Document web



Data web

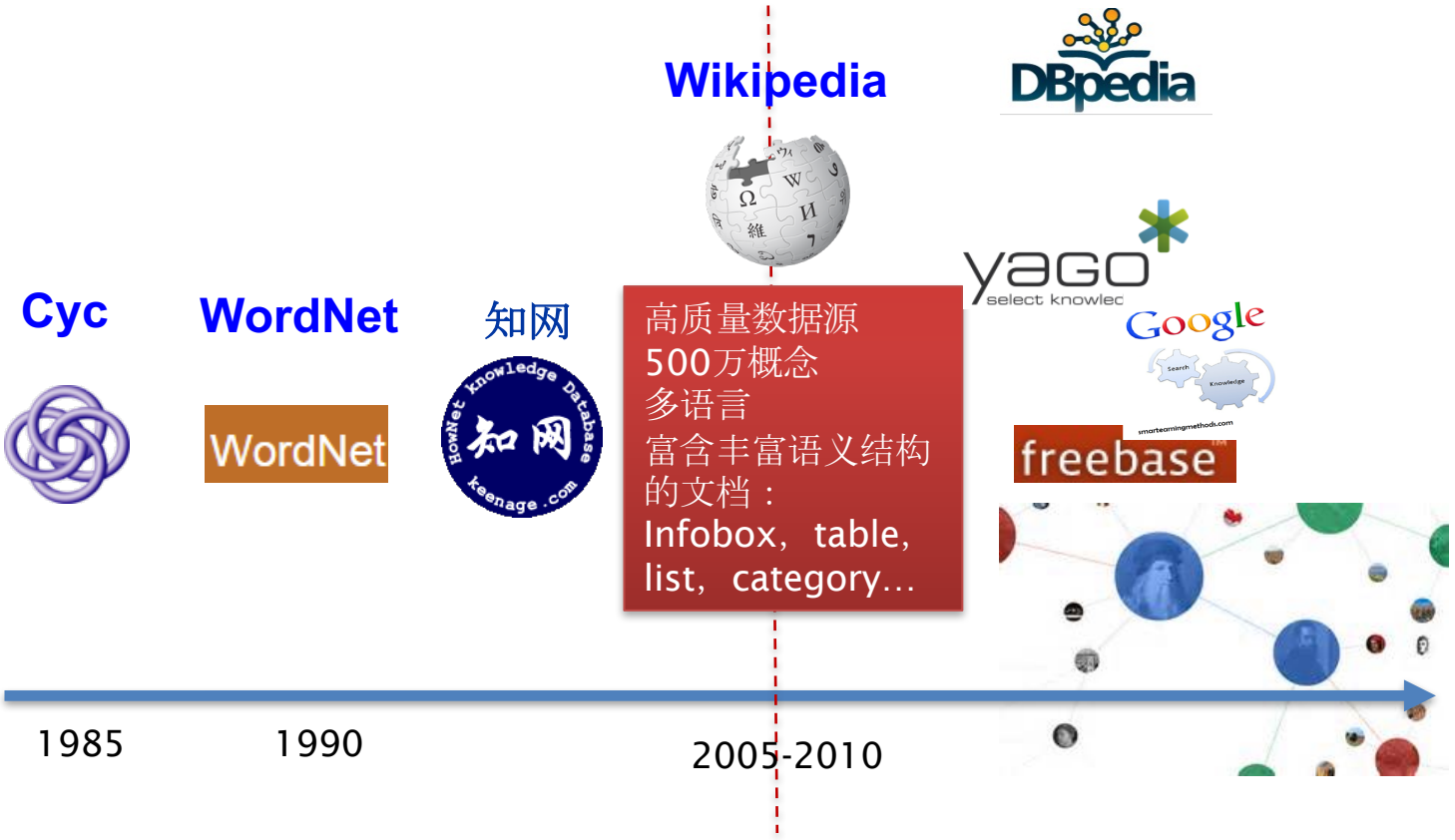
已有的知识图谱

- 语言知识图谱
  - WordNet：155, 327个单词，同义词集117,597个，同义词集之间由22种关系连接
- 事实性知识图谱
  - OpenCyc： 23.9万个实体，1.5万个关系属性，209.3万个事实三元组
  - Freebase：4000多万实体，上万个属性关系，24多亿个事实三元组
  - DBpedia：400多万实体，48,293种属性关系，10亿个事实三元组
  - YAGO2 ：980万实体，超过100个属性关系， 1亿多个事实三元组
  - 百度百科：词条数1000万个
  - 互动百科：800万词条，5万个分类，68亿文字

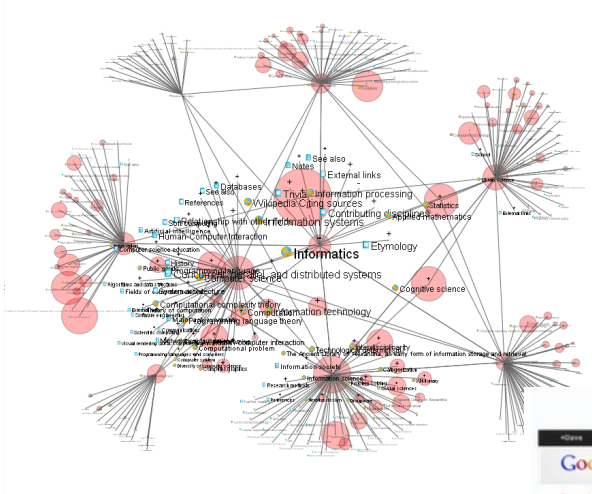
已有的知识图谱

- 领域知识图谱
  - Kinships：描述人物之间的亲属关系，104个实体，26种关系，10,800个三元组
  - UMLS：医学领域，描述医学概念之间的联系，135个实体，49种关系，6,800个三元组。
  - Cora：2,497个实体，7种关系，39,255个三元组
- 机器自动构建的知识图谱
  - NELL： 519万实体，306种关系， 5亿候选三元组
  - Knowledge Vault: 4500万实体，4469种关系，2.7亿三元组

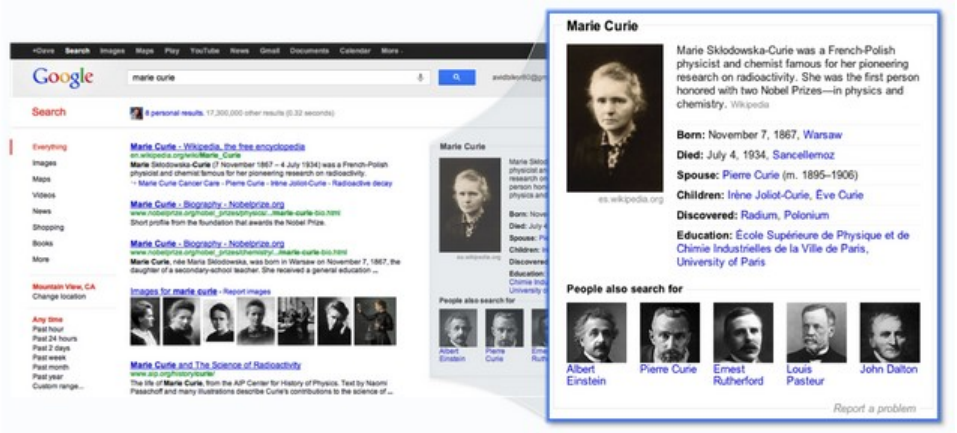
知识图谱历史


Wikipedia
DBpedia
Cyc
WordNet
知网
yago
Google
freebase
高质量数据源
500万概念
多语言
富含丰富语义结构的文档：
Infobox，table，list，category…
1985
1990
2005-2010

Knowledge Graph



超过5亿实体
超过35亿条关系

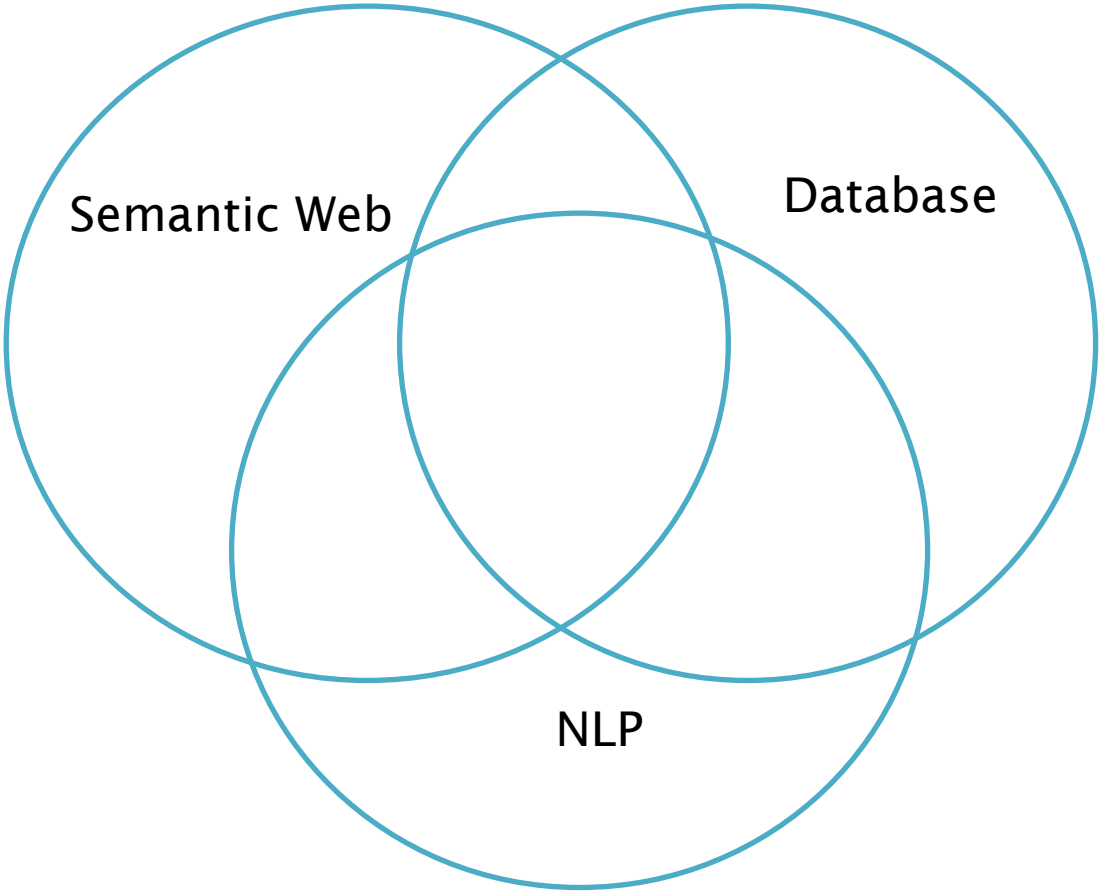


ProBase

百度知心

搜狗知立方

Knowledge Graph 涉及的领域

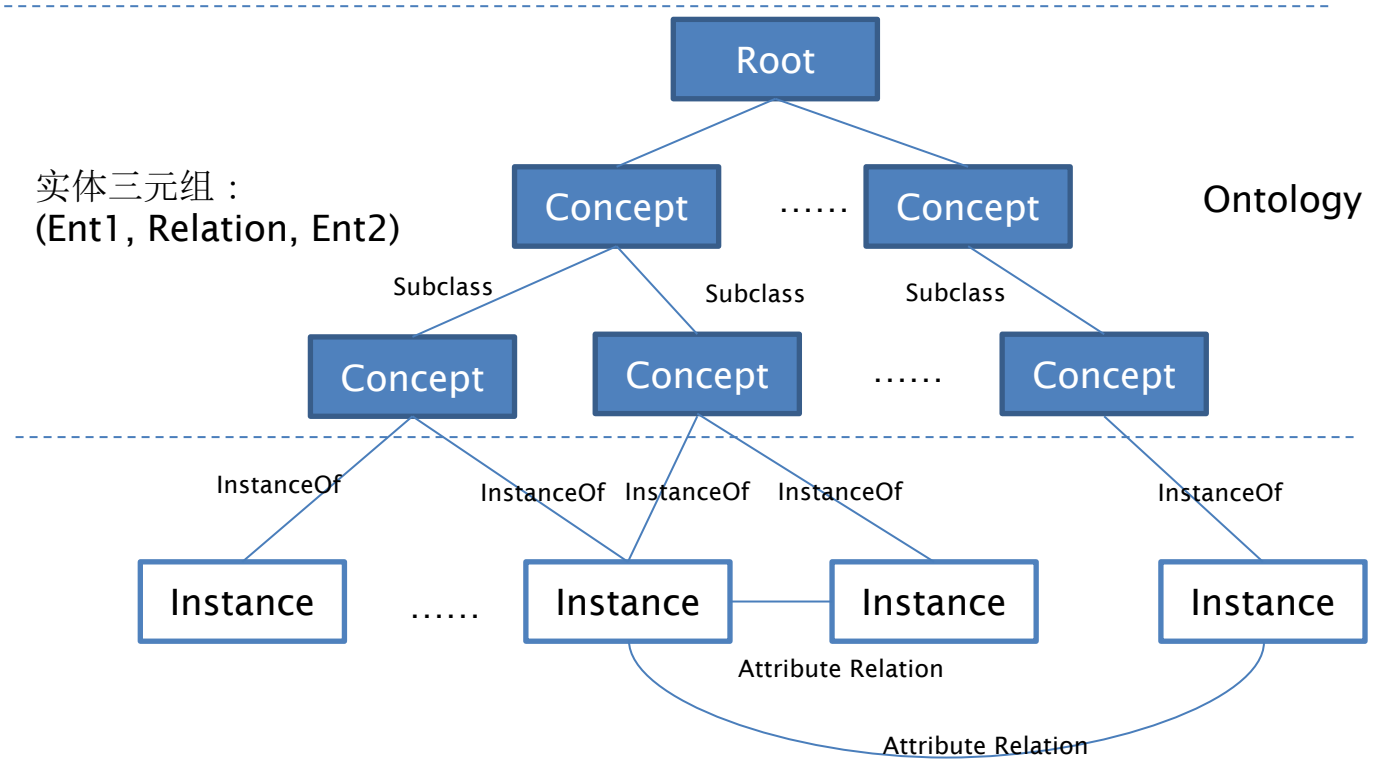


什么是知识图谱

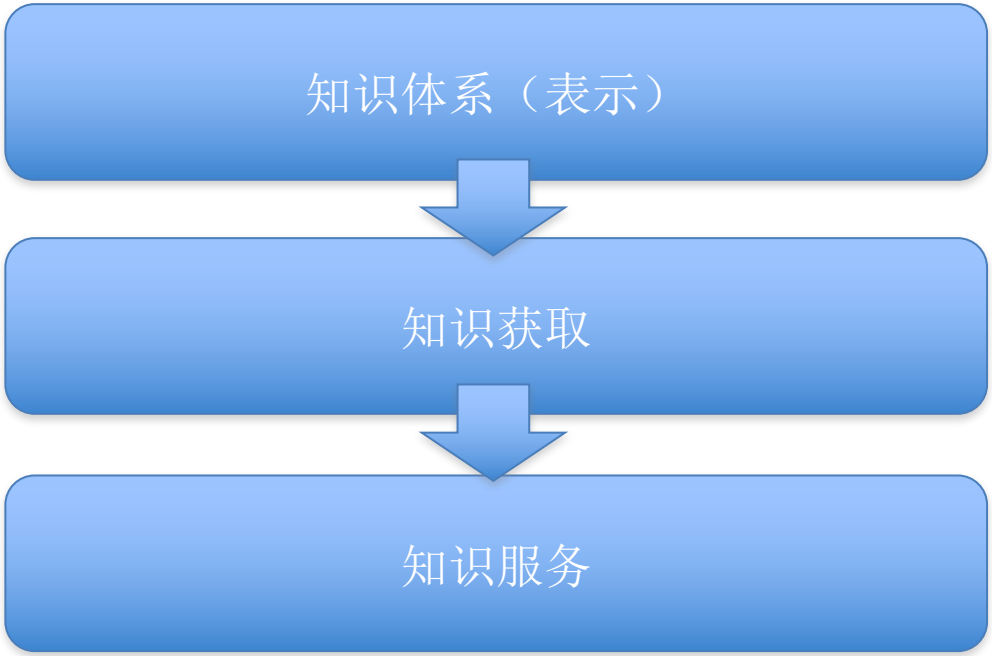
- The Knowledge Graph is a system that understands facts about people, places and things and how these entities are all connected.

- 知识图谱本质上是一种语义网络。其结点代表实体（entity）或者概念（concept），边代表实体/概念之间的各种语义关系

知识图谱包含哪些内容

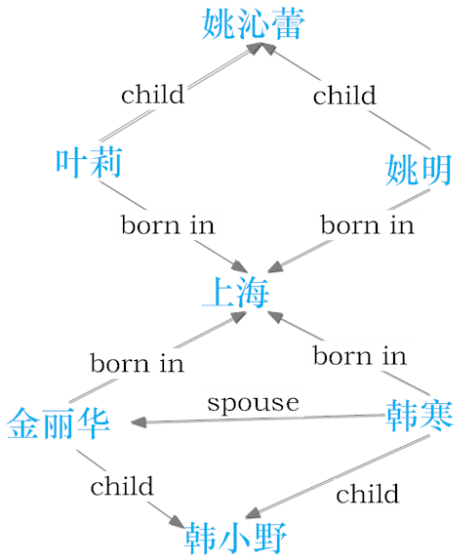

实体三元组：
(Ent1, Relation, Ent2)
Root
Concept
……
Concept
Ontology
Subclass
Subclass
Subclass
Concept
Concept
……
Concept
InstanceOf
InstanceOf
InstanceOf
InstanceOf
InstanceOf
Instance
……
Instance
Instance
Instance
Attribute Relation
Attribute Relation

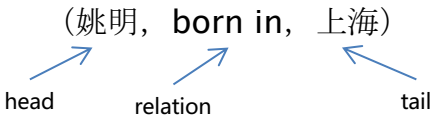
三个层面问题



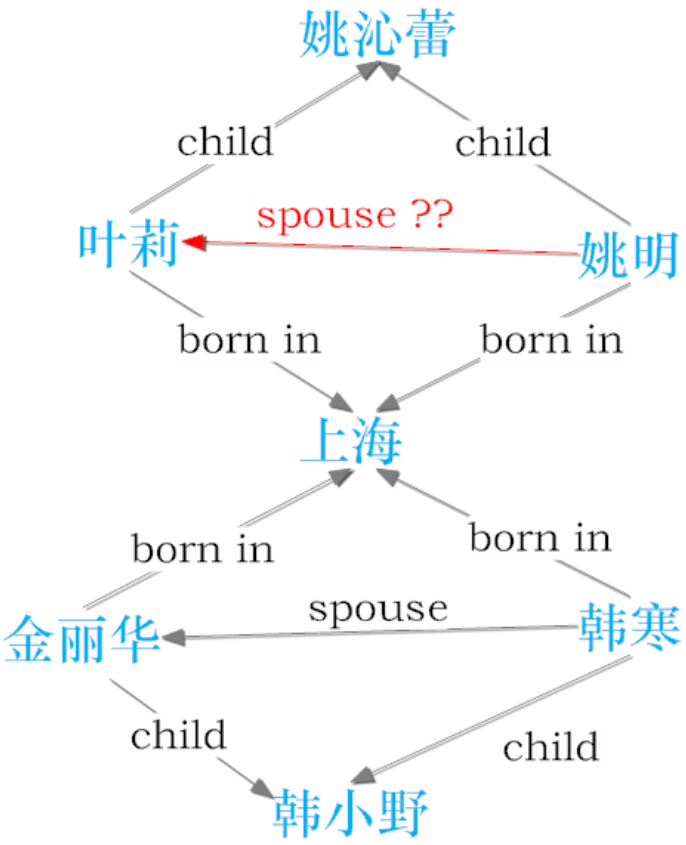
知识图谱概览（基于符号的表示）

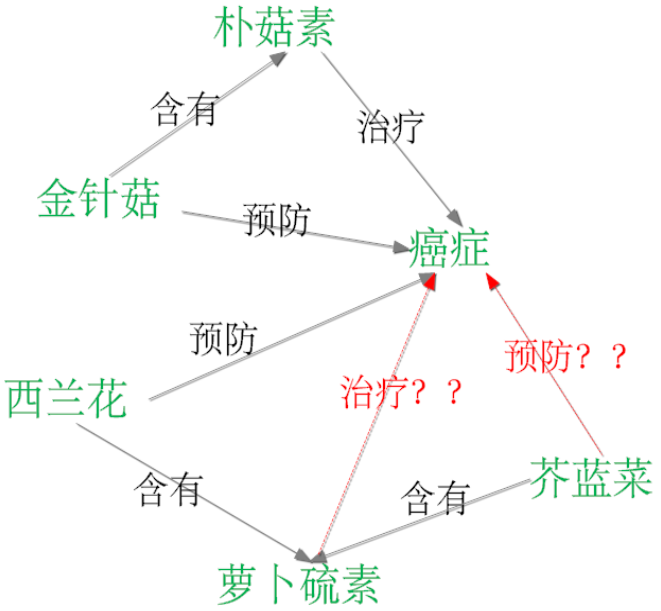
- 知识库是一个有向图
  - 多关系数据(multi-relational data)
  - 节点：实体/概念
  - 边：关系/属性
  - 关系事实 = *(head, relation, tail)*
    - head：头部实体
    - relation：关系/属性
    - tail：尾部实体





知识图谱概览（基于符号的表示）

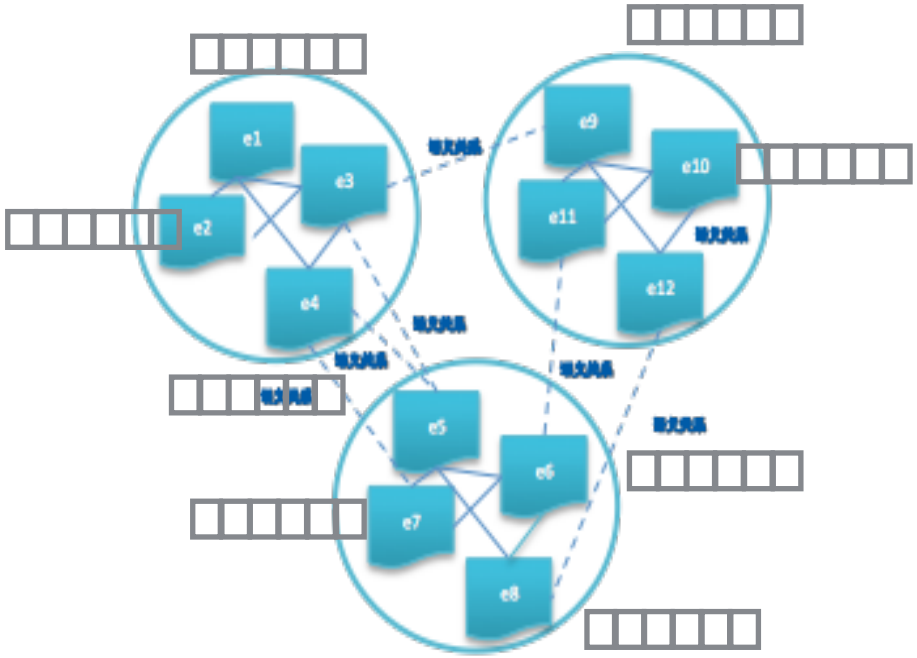




$\text{child}(A,B) \wedge \text{child}(A,C) \Rightarrow \text{spouse}(B,C)$

$\text{含有}(A,B) \wedge \text{治疗}(B,C) \Rightarrow \text{预防}(A,C)$

知识图谱概览（分布式表示）



知识体系组织形式

- Ontology vs. Knowledge Base
  - Ontology：共享概念化的规范，涉及概念、关系和公理三个要素
  - Knowledge：服从于ontology 控制的知识单元的载体
  - Ontology是蛋糕的模具，Knowledge Base是蛋糕

- 公理：Formal Ontology vs. Lightweight Ontology
  - Formal Ontology: 大量使用公理
  - Lightweight Ontology: 不用或很少使用公理

知识体系组织形式

- Ontology
  - 树状结构，不同层节点之间是严格的IsA关系
  - 优点：可以适用于知识的推理
  - 缺点：无法表示概念的二义性（运动员：体育？人物？）
- Taxonomy
  - 树状结构，上下位节点之间非严格的IsA关系
  - 优点：可以表示概念的二义性（体育→运动员）
  - 缺点：不适用于推理，无法避免概念冗余（餐厅：美食？机构？地点？）
- Folksonomy
  - 类别标签，更加开放
  - 优点：能够涵盖更多的概念
  - 缺点：如何进行标签管理？

知识体系组织形式

- 目前的知识资源多是采用Folksonomy与Taxonomy相结合的组织形式
  - 但是能够覆盖的类别还很少

全部	含有开放分类(Folksonomy)的页面数比例
互动百科	70.19%
百度百科	64.38%

  - 这些开放式类别标签存在冗余、不规范的问题，标签之间也缺乏关联
    - 体育、人物
    - 1980年、购房、房产、房地产……

知识体系组织形式

- 类别属性定义不统一
  - 已有的体系框架
    - GeoNames
    - DBpedia Ontology
    - TexonConcept Ontology
    - KOS
    - Schema.org
      - 1）面对站长，而不是面对知识
      - 2）体系覆盖度不足，局限于英文
      - 3）细致化不足

- Creative works: CreativeWork, Book, Movie, MusicRecording, Recipe, TVSeries ...
- Embedded non-text objects: AudioObject, ImageObject, VideoObject
- Event
- Health and medical types: notes on the health and medical types under MedicalEntity.
- Organization
- Person
- Place, LocalBusiness, Restaurant ...
- Product, Offer, AggregateOffer
- Review, AggregateRating

Schema.org

中文名:	李娜	籍贯:	武汉市
性别:	女	民族:	汉族
国籍:	中国	出生年月:	1982年2月26日
星座:	双鱼座	职业:	运动员 女子网球选手
毕业院校:	华中科技大学	身高:	172厘米

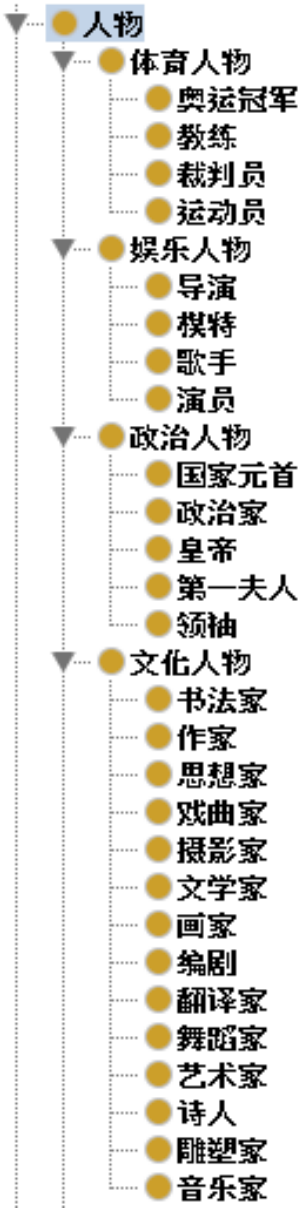
互动百科

中文名	李娜	主要奖项	
外文名	Li Na	重要事件	
别　名	娜姐		
国　籍	中国		
民　族	汉	启蒙教练	
出生地	湖北省武汉市江岸区	训练地	
出生日期	1982年2月26日	教　练	
毕业院校	华中科技大学（新闻学专业）	丈　夫	

百度百科

KG基本概念

- Node：概念（Concept）

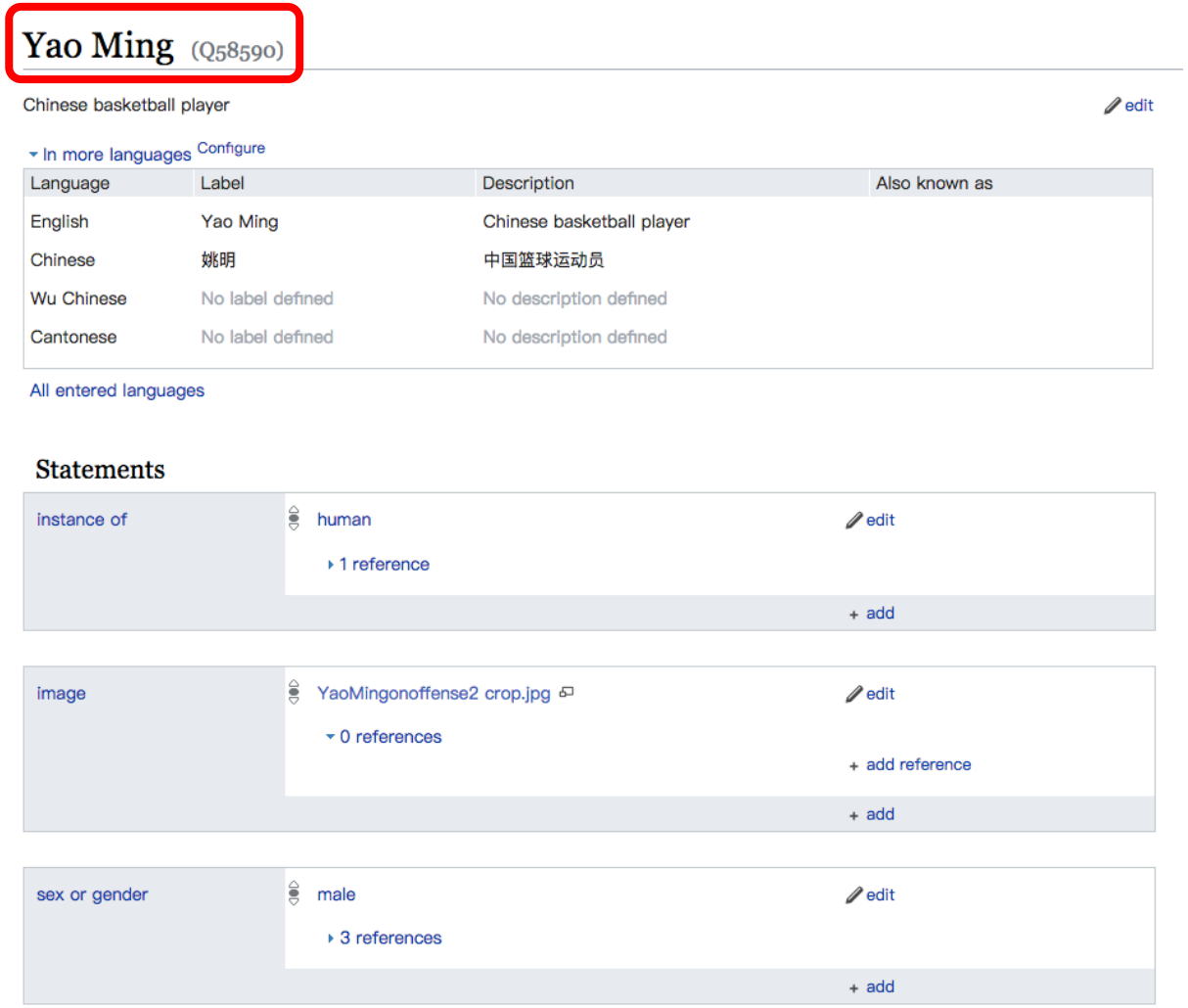


KG基本概念

- Node：领域（Domain/Topic）

KG基本概念

- Node：实例/实体（Entity/Objects/Instance）



Yao Ming (Q58590)

Chinese basketball player — edit

In more languages Configure

Language	Label	Description	Also known as
English	Yao Ming	Chinese basketball player	
Chinese	姚明	中国篮球运动员	
Wu Chinese	No label defined	No description defined	
Cantonese	No label defined	No description defined	

All entered languages

Statements

instance of	human 1 reference	edit + add
image	YaoMingonoffense2 crop.jpg 0 references	edit + add reference + add
sex or gender	male 3 references	edit + add

KG基本概念

- Node：值（Value）
  - 实体（Entity）
    - （姚明，出生地，上海市）
  - 字符串（String）
    - （北京大学，学术传统，兼容并包、思想自由）
  - 数字（Number）
    - 平方公里：（北京市，面积， 1.641万）
    - 公斤：（姚明，体重，140公斤）
    - 米：（姚明，身高， 2.29米）
    - …
  - 时间（Date）
    - （姚明，出生年份，1981年）
  - 枚举（Enumerate）
    - （姚明，性别，男）
  - ……

KG基本概念

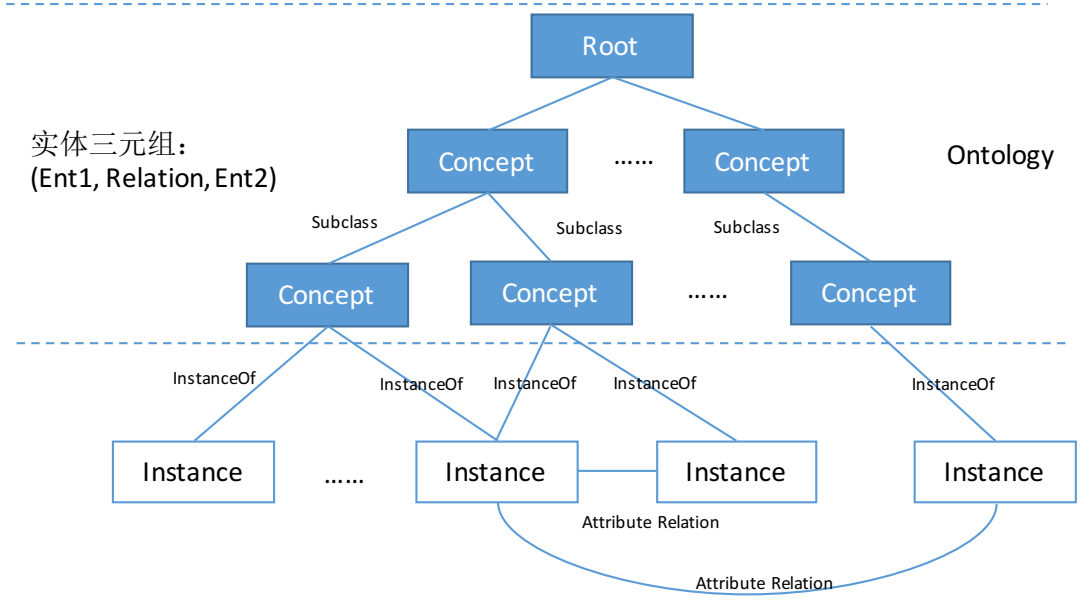
- 边：关系
  - Subclass
  - Type
  - Relation
  - Property、Attribute

(狗， Is-A， 哺乳动物)

(旺财， Is-A， 狗)

(旺财， 朋友， 小白)

(旺财， 颜色， 黄色)



KG基本概念

- 关系：Taxonomic Relation vs. Non-taxonomic Relation
  - Taxonomic Relation： is-a/Hypernym-Hyponym上下位
  - Non-taxonomic Relation：概念之间的相互作用
    - Meoronymy 部分整体
    - Thematic roles 论旨角色
    - Attribute 属性
    - Possession 领属
    - Casuality 因果
    - ……

KG基本概念

- Node：高阶三元组
  - 与时间、地点相关
    - （（美国，总统，特朗普），开始时间，2017）
  - 事件
    - Compound Value Type复合值类型

A Compound Value Type is a Type within Freebase which is used to represent data where each entry consists of multiple fields. Compound value types, or CVT's are used in Freebase to represent complex data.

Example of CVT

Donald Trump (Q22686)

spouse

Ivana Trump
- start time: 7 April 1977
- end time: 22 March 1992

1 reference
- reference URL: http://hollywoodlife.com/celeb/ivana-trump/
- quote: Ivana married Donald Trump on April 7, 1977. (English)

Melania Trump
- start time: 22 January 2005
- place of marriage: Mar-a-Lago

1 reference
- reference URL: http://www.hollywoodreporter.com/features/trumps-wedding-melania-bill-hill-880088
- quote: on Jan. 22, 2005, it was a different story. Trump married model Melania Knauss (English)

Marla Maples
- end time: 8 June 1999
- start time: 19 December 1993

0 references

知识分类

- 百科知识

标注系统 管理中心 导入体系 导入数据 浏览体系 共指列表 高级管理 你好：admin 修改日志 Log out

-root
-创造性工作 [外文名,中文名]
+绘画 [别名,作者,尺寸,类型,年代,收藏单位]
+雕塑 [材料,高度,别名,作者,寓意,类型,年代]
+电视剧 [集数,导演,颜色,语言,编剧,地区,主演,别名,发行时间,获得荣誉,上映时间,制片人,类型,出品公司]
+音乐 [语言,地区,曲长,别名,发行时间,作曲者,类型]
+书籍 [isbn,语言,装帧,页数,开本,出版社,发行时间,作者,别名,字数,类型,价格]
+电影 [导演,颜色,语言,编剧,片长,主演,别名,发行公司,发行时间,获得荣誉,上映时间,imdb编码,制片人,分级,类型,出品公司,地区]
+软件应用 [语言,发行时间,开发者,编程语言,操作系统]
-组织 [外文名,所在地,中文名,地址,别名,创建时间]
+教育组织 [类型]
+运动队 [主教练,代表队员,获得荣誉,所在联赛,运动项目]
+非政府组织 [创始人]
+政府机构 []
+公司 [注册资本,上市代码,公司口号,经营范围,法人代表,上市市场,证券简称,年盈利,总部所在地,宗旨理念,员工数,产品,创始人,总资产,行业,净利润,性质]
-人物 [出生地,外文名,毕业院校,政党,职业,籍贯,去世日期,别名,信仰,国籍,中文名,体重,血型,星座,出生日期,性别,身高,相关事件,民族]
+网络人物 [代表作品,艺名,获得荣誉,主要成就]
+文化人物 [代表作品,获得荣誉,主要成就]
+娱乐人物 [艺名,获得荣誉,主要成就,经纪公司]
+政治人物 [获得荣誉,主要成就]
+虚拟人物 []
+体育人物 [运动项目,获得荣誉,主要成就]
+经济人物 [获得荣誉,主要成就]
+社会科学人物 [获得荣誉,主要成就]
+自然科学人物 [获得荣誉,主要成就]
-地点 [外文名,所在地,面积,中文名,别名,位置]
+公共设施 []
+旅游景点 [主要景点,开园时间,闭园时间,邮政区码,地址,电话区码,门票,海拔,创建时间,分级,竣工日期]
+行政区域 [GDP,主要景点,民族,现任领导人,著名高校,下辖地区,方言,邮政区码,知名企业,车牌代码,火车站,电话区码,创建时间,机场,时区,人口,政府驻地,知名产业,特产,名人]
+地形地貌 [气候]

知识分类

- 领域知识

股票： tags［行业，地区，板块，股票种类］

{

基本属性：{

股票代码（ID）[String] (六位整形数字):

股票种类 [String] (A B H N S股):

股票简称 [String]

股票英文简称 [String]:

上市日期 [Date] (1980.01.01-now):

上市地点 [String]:

上市板 [String]:

交易币种 [String]:

股票面值 [double] (>=0):

摘牌日期 [Date] (1980.01.01-now):

}

发行属性：{

公司名称: [公司实体 Id]

成立日期 [Date]:

上市日期 [Date]:

发行数量(万股) [Double]

发行价格(元) [Double]:

发行市盈率 [Double]:

预计募资(万元) [Double]:

实际募资(万元) [Double]:

主承销商 [String]:

上市保荐人 [String]:

}

知识分类

- 事实性知识
  - (乔布斯，CEO，苹果）
  - (中华人民共和国，首都，北京）

- 主观性知识

“我今年天让入手诺基亚5800，把玩不到24小时，目前感觉5800屏幕很好，操作也很方便，通话质量也不错，但是外形有些偏女性化，不适合男生。这些都是小问题，最主要的问题是电池不耐用，只能坚持一天，反正我觉得对不起这个价格。”

- 外形
- 电池

- 屏幕
- 操作
- 通话质量

知识分类

- 场景知识
  - 打人，打篮球
  - MJ出版了三部专辑→Michael Jackson
  - MJ获得了NBA总冠军→Michael Jordan
  - 订机票的步骤，红烧肉的做法
- 语言知识
  - (乔丹，SameAs， 佐敦)
  - (乔丹，SameAs, Jordan)
  - (Microsoft， SameAs, MS)
  - (hasFounded, SameAs, isFounderof)
- 常识知识 (Common-sense Knowledge)
  - hasAbility(鸟，飞）， hasAbility(人，说话）
  - hasProperties(水，透明）， hasShape(球，圆的）
  - moreHeavy(大象，小狗）
  - Mother(x,y) ^ Brother(z,x) ➔ Uncle(z,y)

目录

知识图谱系统的架构

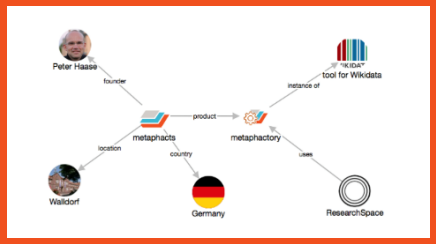
Applications

- Semantic Search
- Analytics
- Knowledge Sharing
- Question Answering
- Dashboards
- Knowledge Management

Algorithms

- Inferencing
- Entity Recognition
- Text Understanding
- Machine Learning
- Disambiguation
- Recommendations

Knowledge Graph



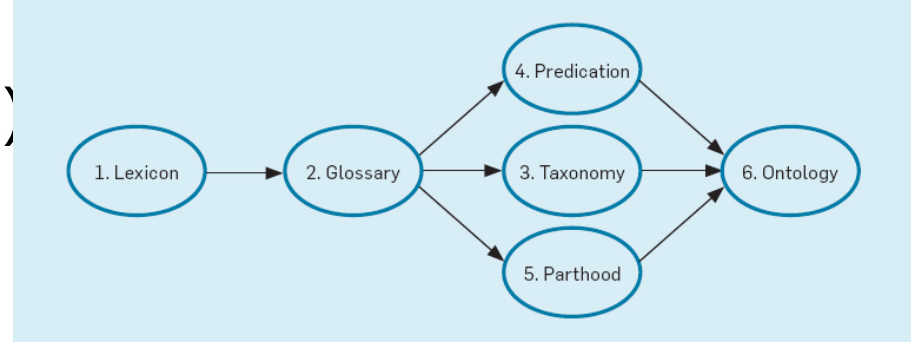
- Entities
- Relationships
- Semantic Descriptions

Data Sources

Data Transformation, Integration
Natural Language Processing

生命周期—领域知识建模

- **输入：**
  - 目标领域 (医疗、金融...)
  - 应用场景



- **输出：**领域知识本体
  - 领域实体类别体系
  - 实体属性
  - 领域语义关系
  - 语义关系之间的关系

- **关键技术：**
  - Ontology Engineering

Ontology engineering

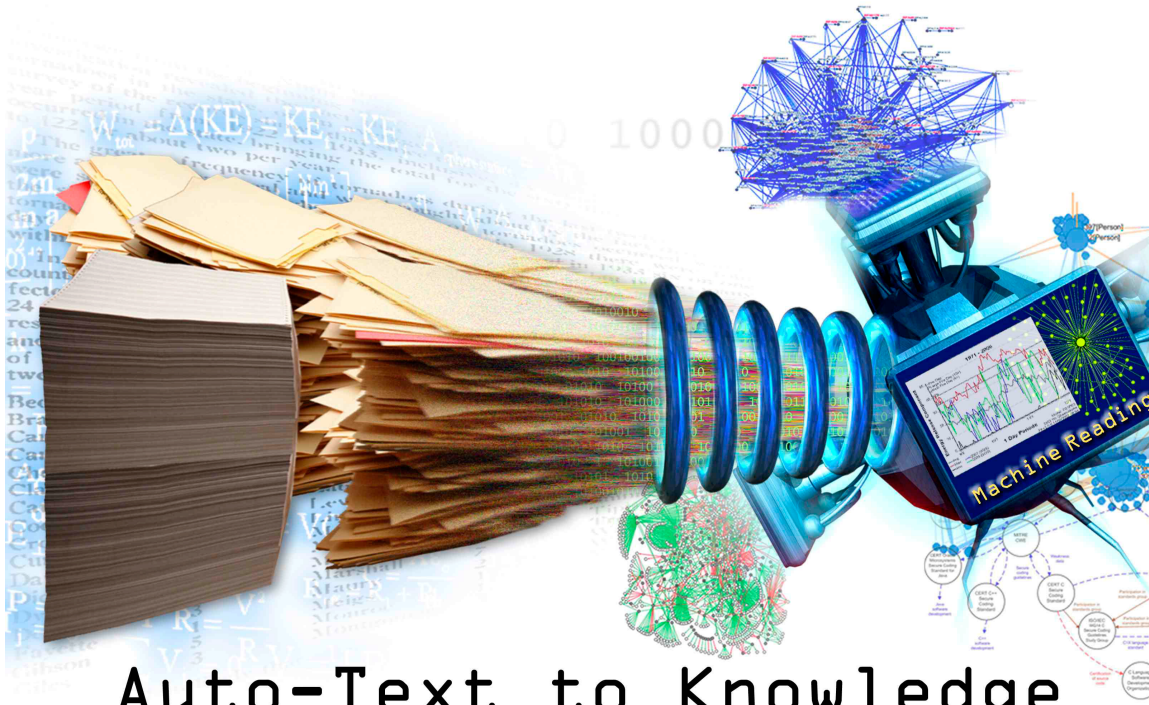
File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

	A	B	C	D
1	**Term**	**Synonyms**	**Kind**	**Description [source]**
2	Delivery address	Shipping address	Complex property	Location to which goods are to be sent [1].
3	Invoice	Bill	Object	Itemized list of goods shipped, usually specifying the price and the terms of sale [2].
4	Postal address	Address	Complex property	Information that locates and identifies a specific address, as defined by the postal service [3].
5	Purchasing conditions	Purchase terms and conditions	Object	Conditions related to the transaction and the trade [4].
6	Purchase order	PO	Object	Commercial document issued by a buyer to a seller, indicating types, quantities, and agreed prices for products or services the seller will provide to the buyer [5].
7	Customer	Client	Actor	One who purchases a commodity or service [2].
8	Invoicing	Issuing invoice	Process	Making or issuing an invoice for goods or services [6].
9	Purchasing	Buying	Process	Acquisition of something for payment [6].

1. Lexicon | 2. Glossary | 3. Taxonomy | 4. Predication | 5. Parthood

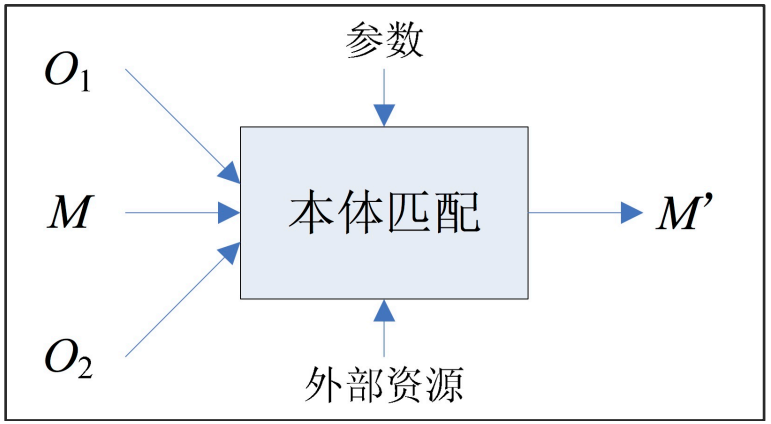
生命周期—知识获取

- **输入：**
  - 领域知识本体
  - 海量数据：文本、垂直站点、百科
- **输出：**领域知识
  - 实体集合
  - 实体关系/属性
- **主要技术：**
  - 信息抽取
  - 文本挖掘



生命周期—知识集成

- **输入：**
  - 抽取出来的知识
  - 现有知识库
  - 知识本体

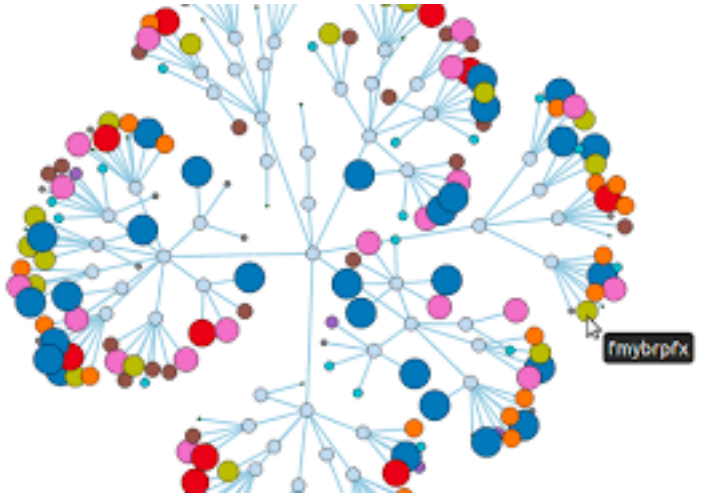


- **输出：**
  - 知识置信度
  - 统一知识库

- **关键技术：**
  - Ontology Matching
  - Entity Linking

生命周期—知识存储/查询/推理

- **输入：**
  - 大规模知识库知识
- **输出：**
  - 知识库存储/查询/推理服务
- **主要技术：**
  - 知识表示
  - 知识查询语言
  - 存储/检索引擎
  - 推理引擎



知识图谱的生命周期

- **知识建模**
  - 建模领域知识结构
- **知识获取**
  - 获取领域内的事实知识
- **知识集成**
  - 估计知识的可信度，将碎片知识组装成知识网络
- **知识存储**
  - 提供高性能知识服务

目录

- 知识图谱发展历史与现有应用
- 知识图谱基本概念
- 知识图谱的生命周期
- 代表性知识图谱

代表性知识图谱

- **人工构建知识图谱**
  - WordNet
  - CYC
- **基于Wikipedia的知识图谱**
  - Yago
  - DBPedia
  - Freebase
- **文本抽取知识图谱**
  - NELL

代表性知识图谱

- **人工构建知识图谱**
  - WordNet
  - CYC
- **基于Wikipedia的知识图谱**
  - Yago
  - DBPedia
  - Freebase
- **文本抽取知识图谱**
  - NELL

WordNet

WordNet
A lexical database for English

- **WordNet是什么?**
  - 一部在线词典数据库系统，采用了与传统词典不同的方式，即按照词义而不是词形来组织
  - 词语被聚类成词义簇(synset)，词义之间通过语义关系连接成大的概念网络
- 由普林斯顿大学认知科学实验室在1985年建立

WordNet包含的知识

描述的对象

- compound（复合词）
- phrasal verb（短语动词）
- collocation（搭配词）
- idiomatic phrase（成语）
- word（单词）

对象之间的语义关系

- 同义反义关系（synonymy，antonymy）
- 上下位关系（hyponymy，hypernym，troponymy）
- 部分整体关系（entailment，meronymy）

部分句法信息

- 简单的动词基本句式信息（Verb Sentence Frames）

WordNet的核心概念

- **Synset：** WordNet 将英语的名词、动词、形容词、和副词组织为Synsets，每一个Synset表示一个基本的词汇概念

- **概念关系**
  - 同义关系
  - 反义关系
  - 上位关系
  - 下位关系
  - 整体关系（名词）
  - 部分关系（名词）
  - 蕴含关系（动词）
  - 因果关系（动词）
  - 近似关系（形容词）

newspaper词义的上位synsets

newspaper, paper
=> press, public press
=> print media
=> medium
=> instrumentality
=> artifact, artefact
=> whole, unit
=> object, physical object
=> physical entity
=> entity

WordNet组织示例

Concept hierarchy

{equipment}

Concept ("synset")

Synonymous words for the concept

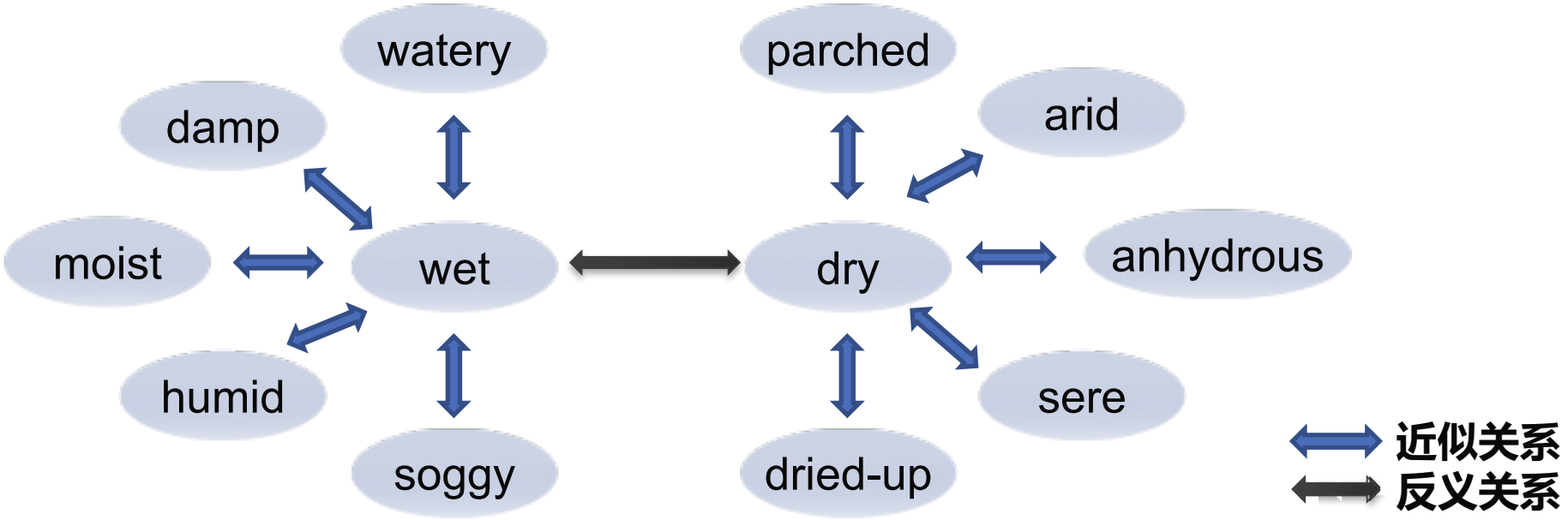
{ camera, photographic camera}

{ camera, television camera}

Gloss describing the concept

"television equipment consisting of a lens system that focuses an image on a photosensitive mosaic"

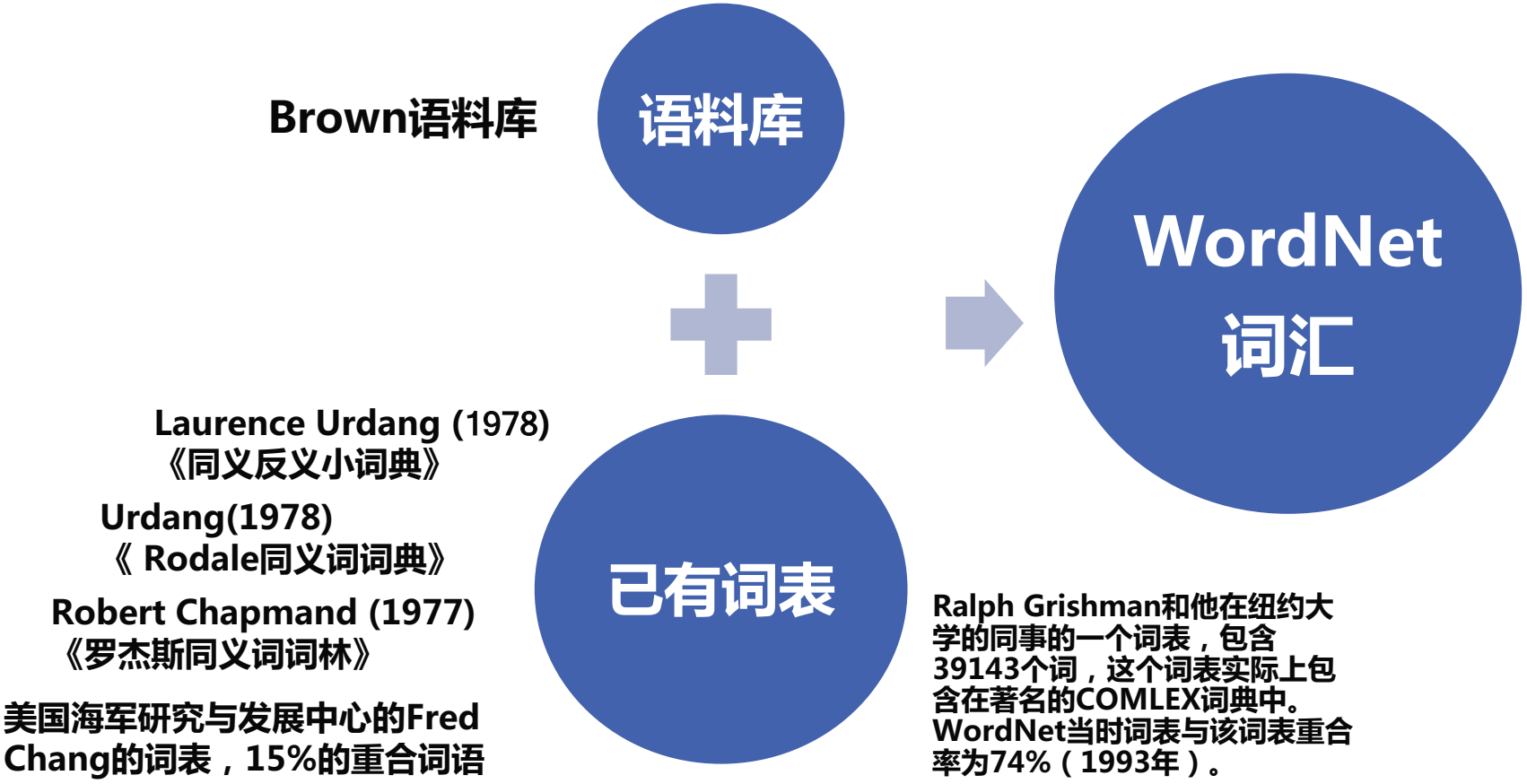
WordNet组织示例



基于反义、近义组织的形容词synset

WordNet的构建方法

- **人工构建+机器辅助（后续有很多自动构建技术研究）**



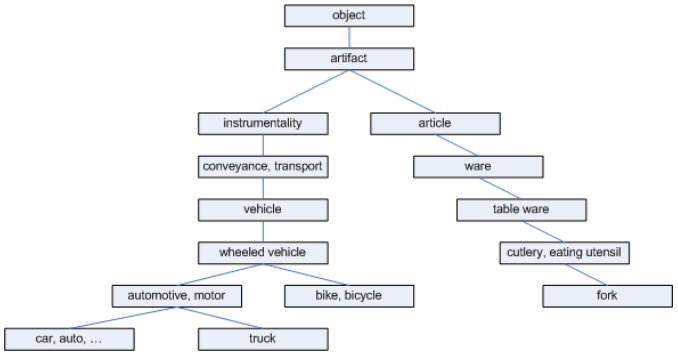
WordNet规模

POS	Unique Terms	Synsets	Total Word-Sense Pairs
Noun	109195	75804	134716
Verb	11088	13214	24169
Adjective	21460	18576	31184
Adverb	4607	3629	5748
Totals	146350	111223	195817

WordNet的应用

- **WordNet在自然语言处理中被广泛应用**
  - 作为词义消歧的目标知识库
  - 作为高质量的Taxonomy
  - 用于计算语义相似度

Word 1	Word 2	lin	wup	path	remarks
genuineness	genuine	0	0	0	needs explanation
Valid	reality	0	0	0	needs explanation
Painter	paint	0.15	0.62	0.09	Good Enough
Really	fact	0	0	0	needs explanation
Real	reality	0.1	0.3	0.09	Good Enough
really	reality	0	0	0	needs explanation
paint	painting	0.3	0.8	0.12	Good Enough



WordNet: a lexical database for English

GA Miller - Communications of the ACM, 1995 - dl.acm.org

Abstract Because meaningful sentences are composed of meaningful words, any system that hopes to process natural languages as people do must have information about words and their meanings. This information is traditionally provided through dictionaries, and

被引用次数：9323　相关文章　所有 34 个版本　引用　保存　更多

WordNet综述

Content	Adjectives, verbs, nouns and adverbs of the English language
Format	Visualization tool data downloadable in Prolog-like format
Main strength	High quality lexicon for English
Technique	Manual
Size	Words: 155k
	Senses: 117k
	Word-sense pairs: 207k
License	Proprietary, free use
Reference	[Miller, Comm ACM 1995]
URL	http://wordnet.princeton.edu

一个高质量英文电子词典和本体

代表性知识图谱

- **人工构建知识图谱**
  - WordNet
  - **CYC**
- **基于Wikipedia的知识图谱**
  - Yago
  - DBPedia
  - Freebase
- **文本抽取知识图谱**
  - NELL

Cyc

- **Cyc** 是一个由Douglas Lenat 在1984年启动的人工智能项目，其目的是构建**一个完整的、机器可使用的本体体系和人类常识知识库**
  - 500万条知识
  - 50万概念
- **OpenCyc** 是其开放出来免费供大众使用的一部分知识
  - 24万概念
  - 200万条知识
- **ResearchCyc**是提供Research Liscence供研究使用的完整版

Cyc构建方法--人手工构建

- 1986年, Doug Lenat估计整个Cyc需要包括**25万条规则，耗费350人年**
  - 这是一个明显的低估

- 近年来，也开始使用自动构建方法，从自然语言中抽取知识

- 2008年开始，Cyc开始将其资源与Wikipedia、DBpedia和Freebase等资源开始建立Link

Cyc中包含的知识

- Cyc中包含的大部分知识都是**常识**
  - *Every tree is a plant(所有树都都是植物)*
  - *Plants die eventually（植物都会死）*

- 核心概念：
  - **Individuals**: #$BillClinton, #$France
  - **Collections**:
    - #$Tree-ThePlant（包含了所有的树）
    - #$EquivalenceRelation(包括所有的等同关系)

Cyc中包含的知识-Sentences

- **事实通过CycL sentences 来表示**
  - *(#$isa #$BillClinton #$UnitedStatesPresident)*
  - *(#$genls #$Tree-ThePlant #$Plant):"All trees are plants".*
- **Rules:** 包含变量(以?开头)的句子
  - 如果一个对象是SUBSET的成员，同时SUBSET是SUPERSET的子类，那么OBJ是SUPERSET的成员

```
(#$implies
   (#$and
     (#$isa ?OBJ ?SUBSET)
     (#$genls ?SUBSET ?SUPERSET))
   (#$isa ?OBJ ?SUPERSET))
```

推理引擎与应用

- Cyc提供了非常多的推理引擎，支持演绎推理和归纳推理；同时也提供了扩展推理机制的模块
- 支持自然语言的解析

English Words	18,796
Syntactic Frame Links	23,336
Single-word Denotation Mappings	27,681
Multi-word Phrase Denotation Mappings	44,298
Verbal Semantic Frame Links	3,701
Noun Semantic Frame Links	2,578
WordNet 2.0 Links	11,322
Names (Includes chemical symbols, person/place/organization names, acronyms, etc.)	100,811
Predicate-based Phrasal Links (genTemplates for paraphrase)	9,637

基于Cyc的自然语言解析示例

As of Feb (#$February). 24 (24), Air Force (#$UnitedStatesAirForce) officials (#$PublicOfficial #$OrganizationRepresentative) reported (#$RegisteringAComplaint #$Reporting) that personnel (#$Employee) in the area (#$Area 0 #$FieldOfStudy #$Region-Underspecified) numbered (#$Counting) close to 8,000 (8000). The 100 (100) aircraft (#$AirTransportationDevice) based (#$Base-Support #$MilitaryBase-Grounds #$BaseOfLandProtrusion #$NitrogenBase #$ChemicallyBasicSubstance) in Saudi Arabia (#$SaudiArabia) for patrols (#$Patrolling) over southern Iraq ((#$SouthernRegionFn #$Iraq)) has (#$possesses) seen (#$VisualPerception #$MeetingSomeone #$sees) the addition (#$DoingAddition) of two (2) dozen (12) F-15 (#$FighterPlane-F15) and F-16 fighter jets (#$FighterPlane-F16) to Bahrain (#$Bahrain-TheIsland #$Bahrain (#$CityNamedFn Bahrain #$Bahrain)). The Air Force (#$UnitedStatesAirForce) has (#$possesses) also authorized (#$GrantingPermission) the dispatch (#$SendingSomething) of 12 (12) F-117 (#$FighterPlane-F117) stealth (#$DodgeStealthCar) fighter jets (#$JetOfFluid #$JetPropelledAircraft) to Kuwait (#$CityOfKuwaitKuwait (#$ProperSubcollectionNamedFn-Ternary kuwait #$Individual 34057665-f4ed-11d9-9bea-0002b3a85b0b) #$Kuwait), three (3) B-1 bombers (#$B-1-Bomber) to Bahrain (#$Bahrain-TheIsland #$Bahrain (#$CityNamedFn Bahrain #$Bahrain)) and 14 (14) B-52 (#$B-52-Bomber) bombers (#$SubmarineSandwich #$BomberPlane #$Bomber) to the island (#$Island) of Diego Garcia. It also has (#$possesses) diverted (#$AmusingSomeone #$DivertingSomething) dozens (#$Dozens-Quant 12) of support (#$SupportingSomething #$ShowingSupportForSomeone (#$SubcollectionOfWithRelationFromTypeFn #$PartiallyTangible #$supportingObject #$SupportingSomething)) aircraft (#$AirTransportationDevice) to the region (#$TheRegion) for refueling (#$Refueling (#$MakingAvailableFn #$CombustibleFuelSubstance)),

Cyc综述

一个人工撰写的常识知识库

	Cyc
Content	Common sense knowledge, axioms
Main strength	Huge ontology, with tools
Technique	Manual
License	proprietary, OpenCyc is Apache License V2.0
Entities	500k
Assertions	5m
Relations	15k
Tools	Reasoner, NL tool
URL	http://cyc.com
References	[Lenat, Comm. ACM 1995]

代表性知识图谱

- **人工构建知识图谱**
  - WordNet
  - CYC
- **基于Wikipedia的知识图谱**
  - Yago
  - DBPedia
  - Freebase
- **文本抽取知识图谱**
  - NELL

Wikipedia

- 免费的在线百科全书
  - 2001年开始
  - crowdsource的方式构建
  - 目标：构建全世界最大的百科全书

主要特点

高质量数据源

500万概念

多语言

富含丰富语义结构的文档：Infobox，table，list，category...

Wikipedia: 文档结构

标题=概念

唐太宗

唐太宗李世民（598年1月28日－649年7月10日[1][2][3]），中国唐朝第二任皇帝。祖籍陇西郡成纪县（今甘肃省天水市秦安县北），生于陕西武功县，626年至649年在位。父亲是唐高祖李渊，母亲是窦皇后

概念文本描述

Infobox:以(属性，值)对形式呈现的信息表格

每个页面有多个类别，类别组成Taxonomy

- 分类：598年出生，649年逝世　唐朝皇帝

唐太宗

概要

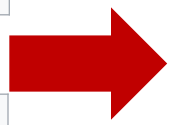
姓名 李世民
庙号 太宗
谥号 文皇帝（649年初谥）
文武圣皇帝（674年加谥）
文武大圣皇帝（749年加谥）
文武大圣大广孝皇帝（754年加谥）
陵墓 昭陵
政权 唐朝
在世 598年1月28日－649年7月10日（52岁）
在位 626年9月4日－649年7月10日
年号 贞观：627年－649年

出发点

- 基于Wikipedia的知识库都基于几乎相同的思路：
  - **从Wikipedia丰富的半结构化信息中挖掘知识**
  - 包括：Infobox，Category，超链接, Table，List...
- 不同之处在于
  - 如何处理有歧义的属性映射
  - 如何构建知识库的Taxonomy

在世	598年1月28日 - 649年7月10日（52岁）

出生	隋文帝仁寿四年604年



BirthDate（李世民，598年）

BirthDate（李元吉，604年）

这些知识库具有相同的数据模型

- **一个知识库包含一个集合的实体**
  - 猫王、李世民、赵高、唐朝、分封制...
- **实体被划分到不同的类别中**
  - 歌手(猫王)，皇帝(李世民), 朝代(唐朝), 制度(分封制)
- **类别通过上下文关系等关系相互关联**
  - SubClassOf(歌手，人)， SubClassOf(皇帝，人)
- **实体和类别都通过属性和相互之间的关系来描述**
  - BirthDate（李世民，598年），Has(歌手，歌曲)
- **关系可以通过蕴含关系来进行推理**
  - 歌曲→作品， 收购 → 持有

DBPedia

- 2007年开始，其主要目标是构建一个社区，通过社区成员来**定义和撰写准确的抽取模板，从维基百科中抽取结构信息，并将其发布到Web上**

- 社区通过人工的方式构建了Taxonomy
  - 280个类别
  - 覆盖约50%的维基百科实体

DBpedia

抽取方法

- **DIEF - DBpedia Information Extraction Framework**
  - 目标：抽取WIkipedia中的结构化信息
  - 方法：基于属性mapping的Infobox抽取，Raw Infobox Extraction, Feature Extraction, Statistical Extraction
  - 编程语言：Scala & Java

- DBPediaLive： 持续保持与Wikipedia的同步
  - 2013年六月，英语维基百科有将近330万次编辑(每分钟越77次)

DBPedia综述

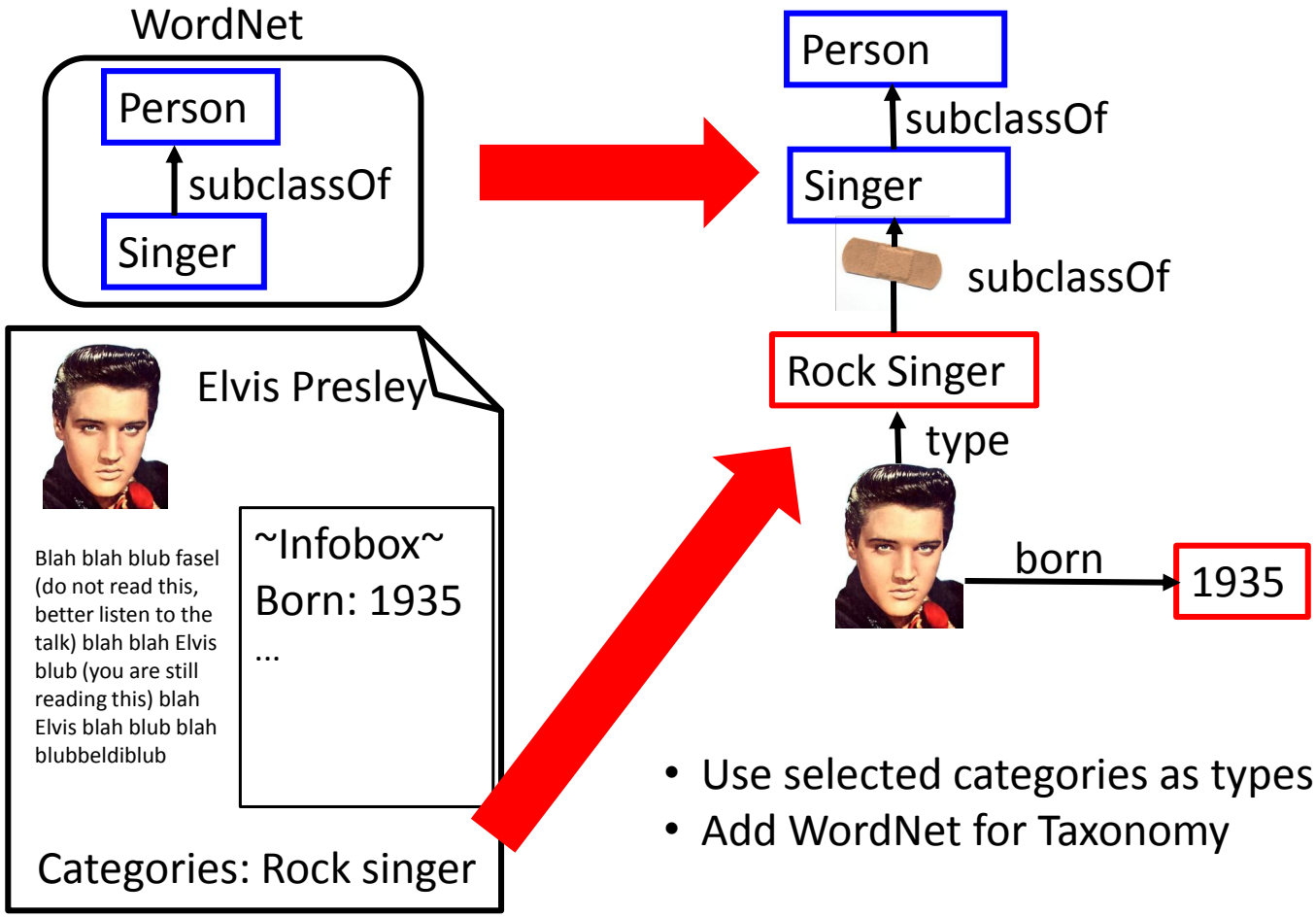
Content	Entities of public interest
Format	RDF, API, SPARQL
Sources	Wikipedia, YAGO/WordNet
Main strengths	Focus on coverage, interlinking with other data sets
Technique	Extraction from Wikipedia + manual supervision by the community
Size	Entities: 3.5m (in manual taxonomy: 1.7m) Facts: 670m Attributes: 9k (manually defined: 1k) Manual Classes: 280
License	CC-BY-SA & GNU FDL
URL	http://dbpedia.org
Reference	[Auer, ISWC 2007], [Bizer09, JWS 2009]

Yago(Yet Another Great Ontology)

- 德国马普研究所从2007年开始的一个项目
- 融合WordNet和Wikipedia
  - 从Wikipedia的结构中抽取信息：Infoboxes,类别...
  - 包括时间和地点标注
  - 人工采样评估
  - >1亿事实和100种关系

Yago Taxonomy构建

- 使用WordNet的Taxonomy作为基础
- 将Wikipedia中的类别加入到WordNet中



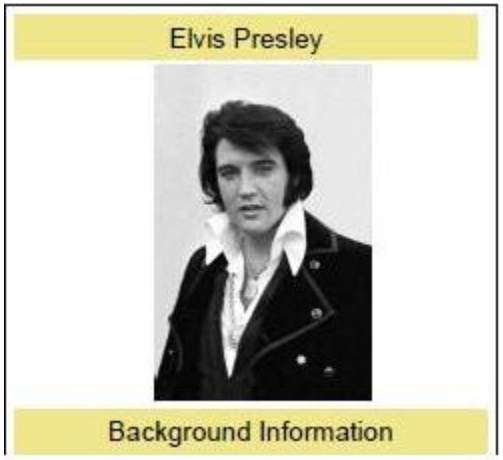
Yago语义关系

- **人工定义了100多种语义关系**
  - wasBornOnDate，locatedIn， hasPopulation

- **抽取方法：主要采用手写的规则抽取**
  - Infobox Harvesting:信息框
  - Word-Level Techniques:重定向页
  - Category Harvesting:类别信息抽取
  - Type Extraction: 维基类别,WordNet类别

抽取示例

Infobox



Died: August 16, 1977

Attribute Map

Attribute	Relation	Inverse	Manifold	Indirect
	…			
Died	diedOnDate			
	…			

Relation Map

Relation	Domain	Range
	…	
diedOnDate	person	yagoDate
	…	

Elvis Presley **diedOnDate** **August 16, 1977**

Yago综述

Content	Entities of public interest
Format	TSV, RDF, XML, N3, Web Interface
Sources	Wikipedia, WordNet, Geonames
Main strength	Focus on precision, geotemporal annotations, multilingual
Precision	95%
Technique	Extraction from Wikipedia + matching with WordNet & Geonames + consistency checks
Size	Entities: 3 m (+ geonames -> 10m) Facts: 120m (+geonames -> 460m) Relations: 100, Classes: 200k, Languages: 200
License	Creative Commons BY-SA
URL	http://mpii.de/yago
References	[Suchanek, WWW 2007] [Hoffart, WWW 2011] [deMelo, CIKM 2010]

Freebase

- Metaweb公司2000年开始构建，2010年被Google收购

- 从Wikipedia和其他数据源（如IMDB、MusicBrainz）中导入知识

- 核心想法：
  - 在Wikipedia中，人们编辑文章
  - **在Freebase中，人们编辑结构化知识**

用户是Freebase知识构建的核心

- **编辑实体**
  - 创建实体
  - 将实体分到类别
  - 增加/修改属性/关系
  - 上传图片
- **编辑Schema**
  - 定义新类别
  - 定义类别的属性
- **Review**
  - 验证知识准确性
  - 投票
  - 删除错误知识
- **DataGame**
  - 寻找别名
  - 抽取事件日期
  - 使用Yahoo图片搜索加入图片

用户在Freebase中的作用

Freebase综述

Content	Entities with public information
Format	API, RDF
Construction	by the community data import from public sources
Sources	Wikipedia, Libraries, WordNet, MusicBrainz…
Main strength	free and large
Size	Facts: several millions
	Entities: 20 m
License	Creative Commons Attribution (CC-BY)
URL	http://download.freebase.com

一个大规模协同构建知识库，目前归Google所有

代表性知识图谱

- **人工构建知识图谱**
  - WordNet
  - CYC
- **基于Wikipedia的知识图谱**
  - Yago
  - DBPedia
  - Freebase
- **文本抽取知识图谱**
  - NELL

NELL

Read the Web
Research Project at Carnegie Mellon University

- 2009年开始的CMU项目
- **输入：**
  - 初始本体（~800类别和关系）
  - 每个谓词的一些实例（~10-20个种子实例）
  - web(~10亿页面, ClueWeb)
  - 间歇性的人工干预
- **任务：**
  - 24 x 7 持续运行(从2010年开始)
  - 每天
    - 抽取更多知识来补充给定本体
    - 学习如何更好的构建抽取模型
- **结果：**超过9千万实例（不同置信度）

抽取结果示例

Recently-Learned Facts

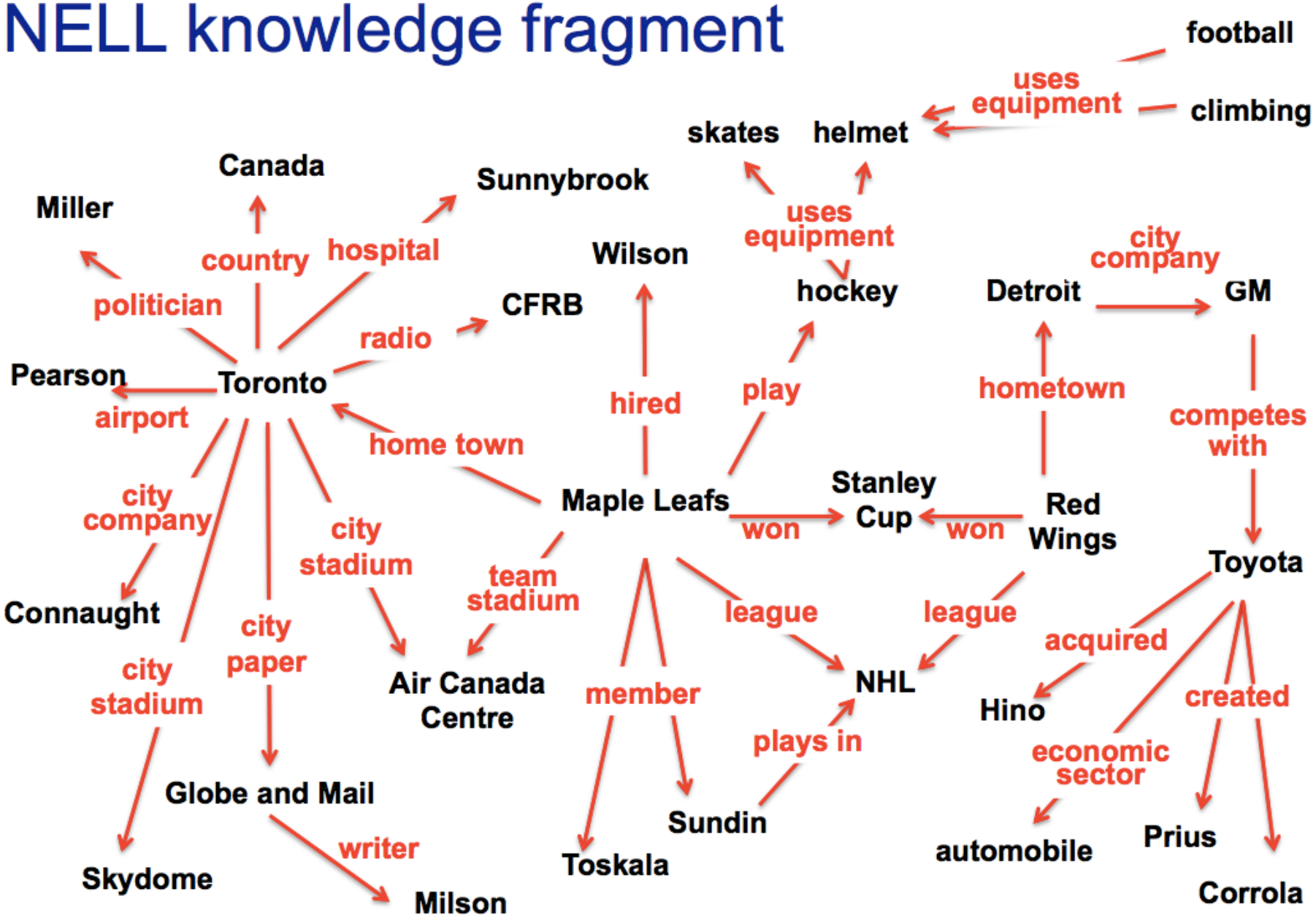
twitter

Refresh

instance	iteration	date learned	confidence
rillito_river is a river	1064	26-jun-2017	99.9
james_dexter is a CEO	1064	26-jun-2017	99.7
richard_thompson is a Mexican person	1064	26-jun-2017	100.0
htc_droid_eris is a consumer electronic device	1064	26-jun-2017	92.8
radiant_snow is a weather phenomenon	1066	15-jul-2017	99.8
the_london_eye001 is a tourist attraction in the city london	1069	03-aug-2017	100.0
state_university is a sports team also known as michigan_state_university	1067	21-jul-2017	100.0
state_university is an organization also known as clemson	1066	15-jul-2017	96.9
irene_kirkaldy belongs to the religion seven_day_adventist	1069	03-aug-2017	100.0
kcal is a TV station in the city los_banos	1069	03-aug-2017	100.0

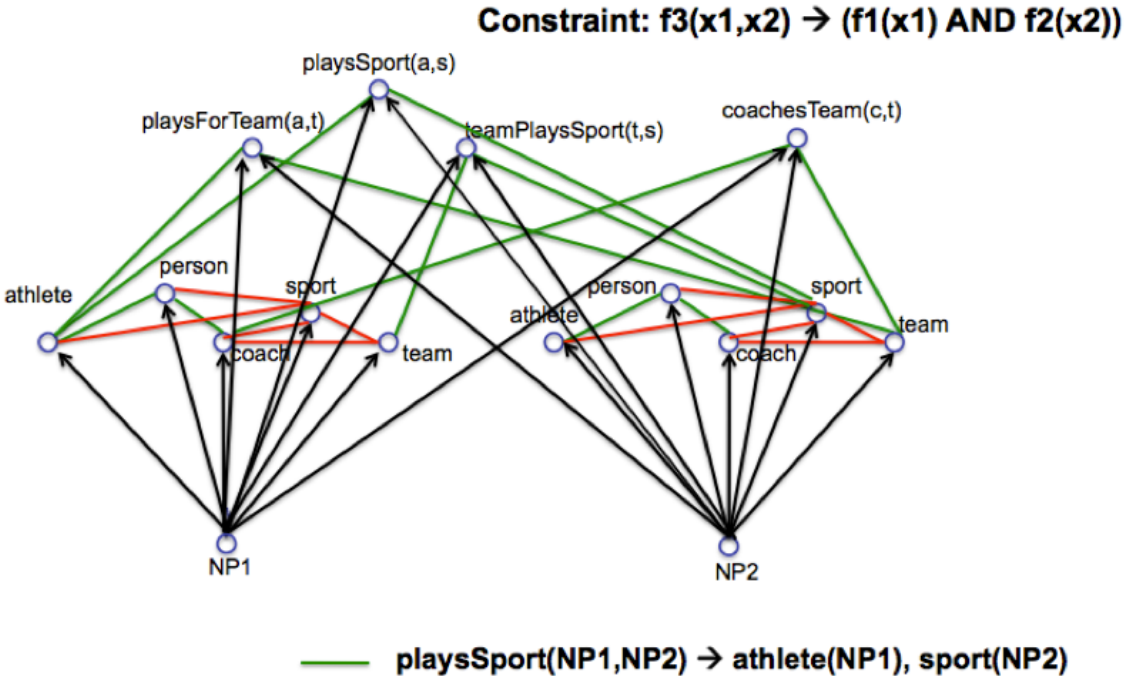
http://rtw.ml.cmu.edu

NELL knowledge fragment



抽取步骤

- **1. 把名词短语划分到给定类别**
  - Entity Set Expansion: 基于Pattern的Bootstrapping
- **2. 分类名词短语之间的语义关系**
  - Coupling Learning



抽取步骤

- **3. 识别新的推理规则，用于发现新的关系实例**
  - PathRank

0.95 athletePlaysSport(?x,basketball) :- athleteInLeague(?x,NBA)

0.93 athletePlaysSport(?x,?y) :- athletePlaysForTeam(?x,?z)
teamPlaysSport(?z,?y)

0.91 teamPlaysInLeague(?x,NHL) :- teamWonTrophy(?x,Stanley_Cup)

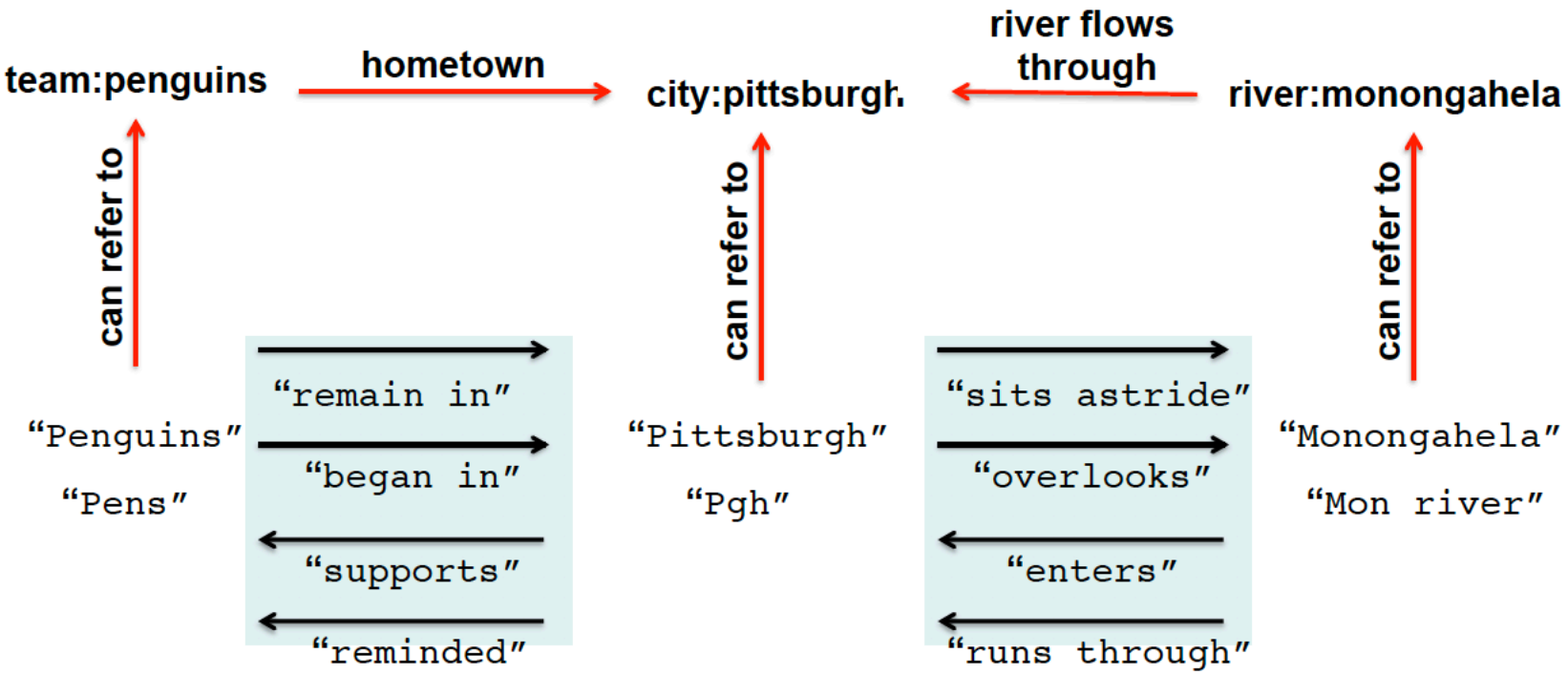
0.90 athleteInLeague(?x,?y):- athletePlaysForTeam(?x,?z),
teamPlaysInLeague(?z,?y)

0.88 cityInState(?x,?y) :- cityCapitalOfState(?x,?y),
cityInCountry(?y,USA)

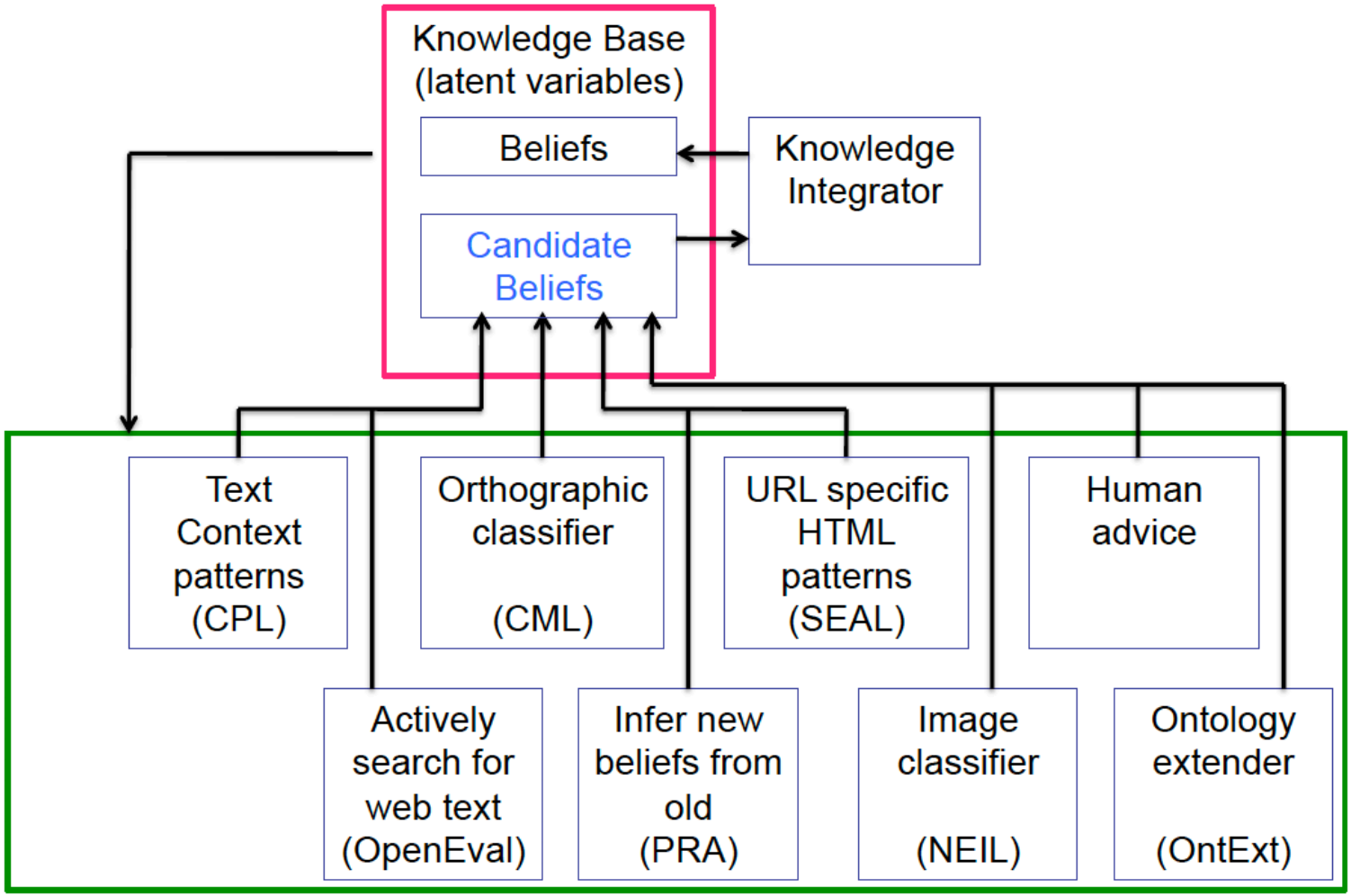
0.62* newspaperInCity(?x,New_York) :- companyEconomicSector(?x,media),
generalizations(?x,blog)

抽取步骤

- **4. 名词短语被映射到概念，动词短语被映射到关系**
  - Entity Linking, Entity Resolution



完整的抽取框架


Knowledge Base
(latent variables)
Beliefs
Candidate
Beliefs
Knowledge
Integrator
Text
Context
patterns
(CPL)
Orthographic
classifier
(CML)
URL specific
HTML
patterns
(SEAL)
Human
advice
Actively
search for
web text
(OpenEval)
Infer new
beliefs from
old
(PRA)
Image
classifier
(NEIL)
Ontology
extender
(OntExt)

NELL综述

Content	Entities mentioned on Web pages
Format	TSV
Construction	by a perpetual extractor
Sources	The Web
Main strength	Not limited to a specific source
Size	Facts: 800k
	Categories & relations: 633
Reference	[Carlson, AAAI 2010]
URL	http://rtw.ml.cmu.edu/

一个基于文本信息抽取技术持续不断更新的知识库

代表性知识图谱总结

知识图谱

- **实体及其之间关系的语义描述**
  - 使用形式化知识表示（如RDF，RDFS，OWL)
- **Entities**
  - 真实世界对象 (things, places, people) 和抽象概念 (genres, religions, professions)
- **Relationships**
  - 将实体按语义关系链接成一张大网
- **Semantic descriptions**
  - 类别和属性
- 有时包含支持推理的公理知识 (如规则)

代表性知识图谱综述

- **人工构建知识图谱**
  - **WordNet**： 英文电子词典
  - **CYC**：常识知识库
- **基于Wikipedia的知识图谱**
  - **Yago**： Wikipedia+WordNet
  - **DBPedia**：基于社区抽取Wikipedia结构化信息
  - **Freebase**：知识编辑社区协同构建
- **文本抽取知识图谱**
  - **NELL**：从文本中持续自动抽取海量知识
- **还有许多其他知识库没有覆盖**：Wikidata、BabelNet、OpenMind、Probase...

当前KG的限制和不足

- **领域限制**
  - 一些知识库侧重于语言: WordNet,BabelNet
  - 一些知识库侧重于Schema： Cyc, UMBEL
  - 一些知识库侧重于Fact: DBPedia, Yago
- **对时空属性的建模**
  - 对动态性的实体，如Event建模不足
  - Yago 3在一定程度上考虑时间和地理属性
- **自动构建**
  - 自动构建是维护和保持KG质量和覆盖的核心技术
- **与LOD的集成**
  - 缺乏Schema之间的alignment
  - 往往只用到底层的表达能力，OWL的高级功能很少涉及

KG展望

- **新的知识表示模型**
  - Ontology engineering已经被用了超过15年
- **新类型的知识图谱**
  - 不再围绕实体和关系的存储
  - 如Event-centric KG
- **知识图谱自动构建技术**
  - 在Freebase中，71%的人没有出生日期
  - 新技术：Distant Supervision, KG embedding, 知识集成（如Google的Knowledge Vault）