When Knowledge Graph meet Python

Yongpan Sheng

目录 CONTENTS

Preliminaries

AI system = Knowledge + Reasoning

Preliminaries

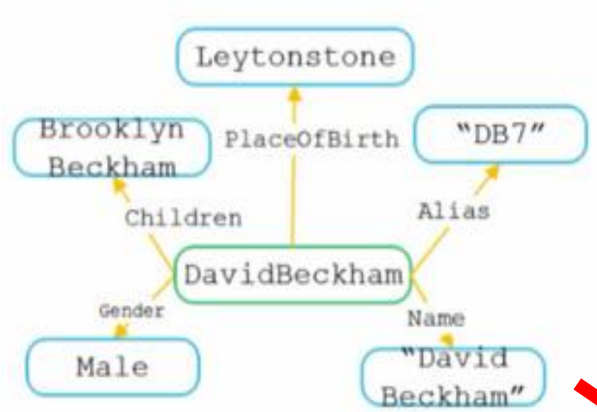
Q: 1M = ? B → 1024

Q: Where was David Beckham born? → Leytonstone

How dose the AI system works?

Q: 1M = 1024 B → 机器运算的过程即是符号操作的过程 （机器的潜台词：“我”有储备，so easy！）。

Q: Where was David Beckham born?



SPARQL

```
SELECT ?object
WHERE { <DavidBeckham>
<PlaceOfBirth> ?object }
```

Knowledge as triples

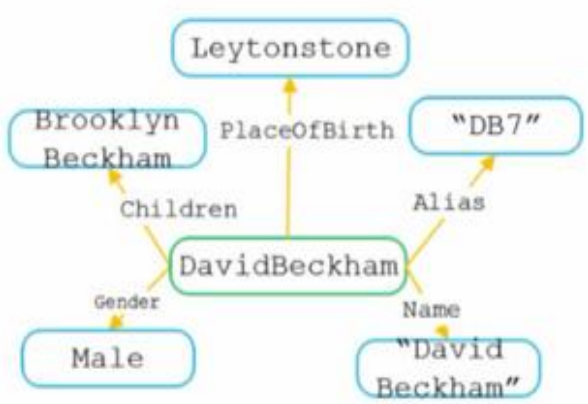
<DavidBeckham, Name, “David Beckham”>

<DavidBeckham, PlaceOfBirth, Leytonstone>

<subject, relation, object>

Preliminaries

Q: Where was David Beckham born?



SPARQL

SELECT ?object
WHERE { <DavidBeckham>
<PlaceOfBirth> ?object }

Knowledge as triples

<DavidBeckham, Name, "David Beckham">

<DavidBeckham, PlaceOfBirth, Leytonstone>

<subject, relation, object>

- **Mapping from natural questions to structured queries executable on knowledge graph**

（机器的潜台词：“我”会推理，so easy！）。

所以，通俗的来说，在AI system中：要么从原有的知识体系中直接提取信息来使用，要么进行推理。

将知识融合在机器中，使机器能够利用我们人类知识、专家知识解决问题，这就是早期知识工程（**Knowledge Engineering**）的核心内涵。

本页PPT借鉴于复旦大学肖仰华老师《大数据时代的知识工程与知识管理》

Preliminaries

Explaining AI system from the perspective of KE – Symbolism

- **符号主义的主要观点**
  - 认知即计算
  - 知识是信息的一种形式，是构成智能的基础
  - 知识表示、知识推理、知识运用是人工智能的核心
- **Physical Symbol System**
  - A physical symbol system has the necessary and sufficient means of general intelligent action. [R1]
  - The mind can be viewed as a device operating on bits of information according to formal rules. [R2]
- **A special entity category (“good old fashioned AI”, proposed by John Haugeland)**
  - Focused on these kind of high level symbols, such as <dog> and <tail>

Newell Simon

AI System = Knowledge + Reasoning

R1：Newell, Allen; Simon, H. A. (1976), “Computer Science as Empirical Inquiry: Symbols and Search”, Communications of the ACM, 19 (3)
R2：Dreyfus, Hubert (1979), What Computers Still Can't Do, New York: MIT Press.

Preliminaries

Conventional KE – Features and Challenges

自上而下：严重依赖专家和用户的干预（规模有限、质量存疑）

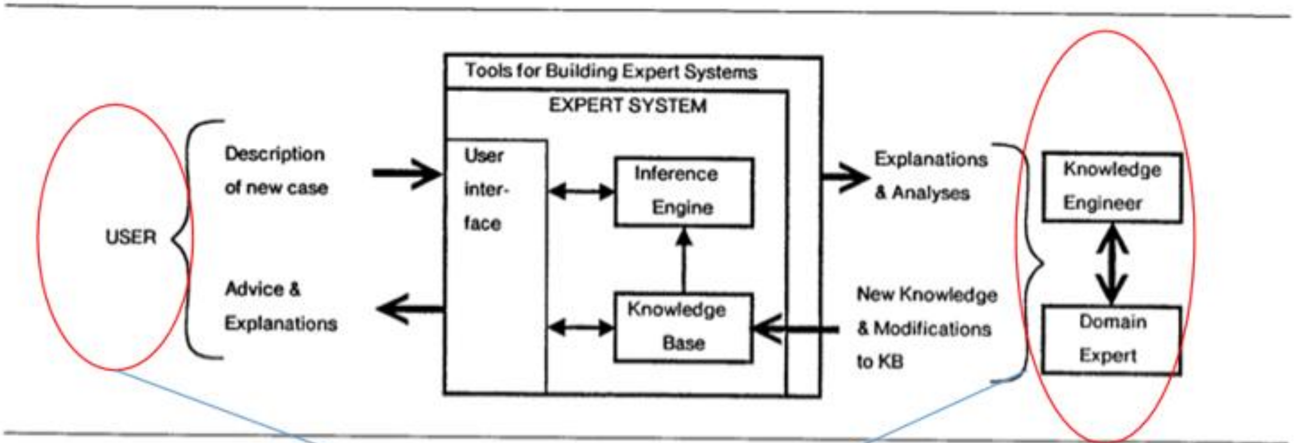


FIGURE 1-2 Interaction of a knowledge engineer and domain expert with software tools that aid in building an expert system. Arrows indicate information flow.

MYCIN专家系统中的人工参与部分

Major difficulties:

1、知识获取困难

e.g., 领域知识难以表达（形式化），因为它往往是一种隐性知识、过程知识。

2、知识应用困难

（1）开放性应用易于超出预先设定的知识边界；（2）有的应用需要尝试知识的支撑，而常识知识往往难以定义、表达、表征。

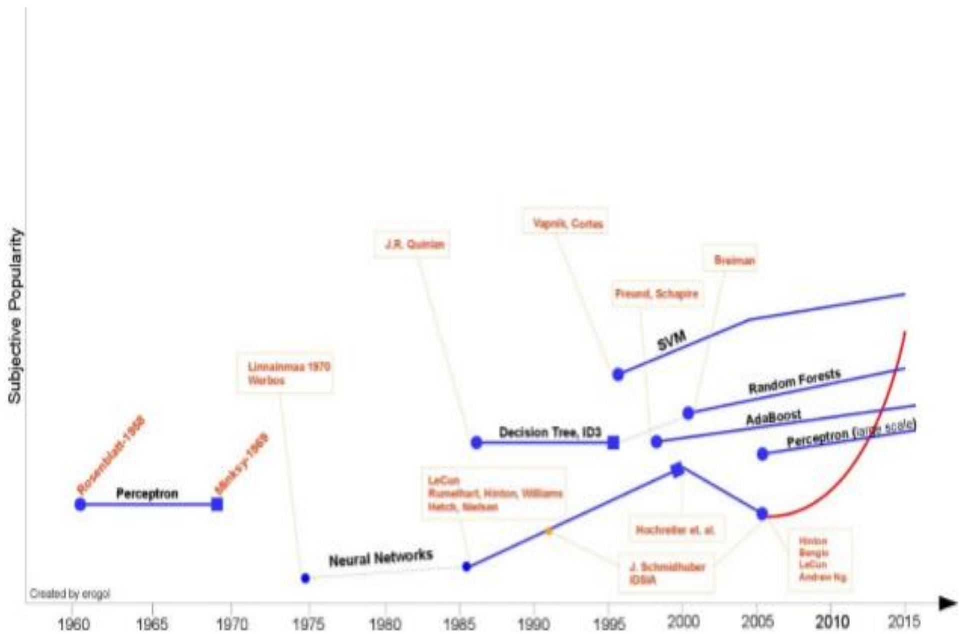
3、很难处理异常情况 e.g., 鸵鸟不会飞

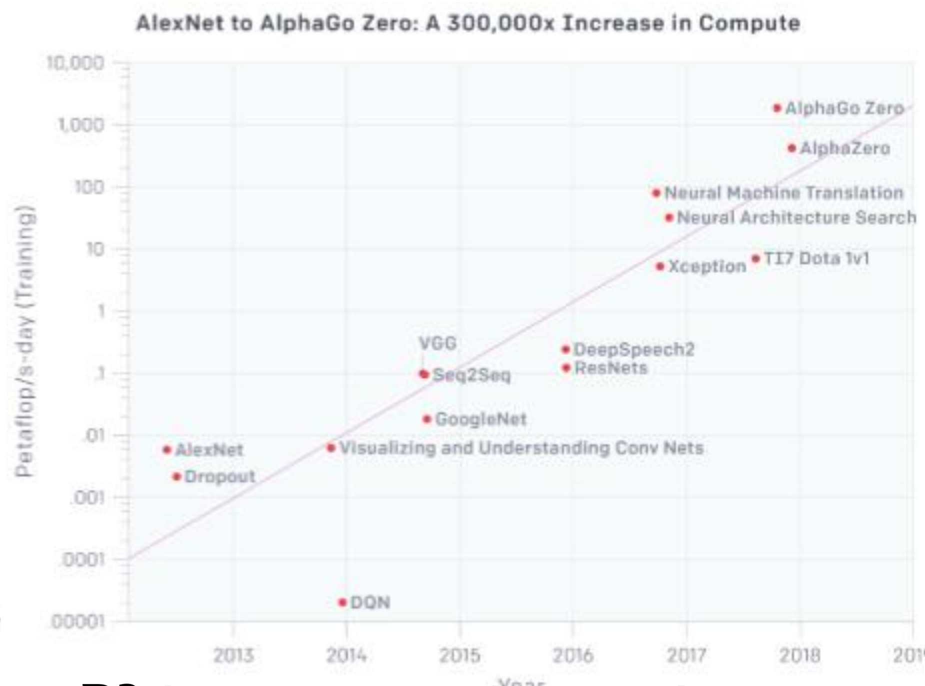
大数据时代催生KE飞速前进发展

Preliminaries

大数据时代的机遇 – 大规模知识自动获取

- Big Data + Machine Learning[R1] + Powerful Computation[R2]
  - 完全意义上的自下而上的方式
  - 从海量的数据中去挖掘**异构、动态、碎片化的知识**
    e.g., 从Web corpora、搜索日志等都可挖掘出有价值的知识





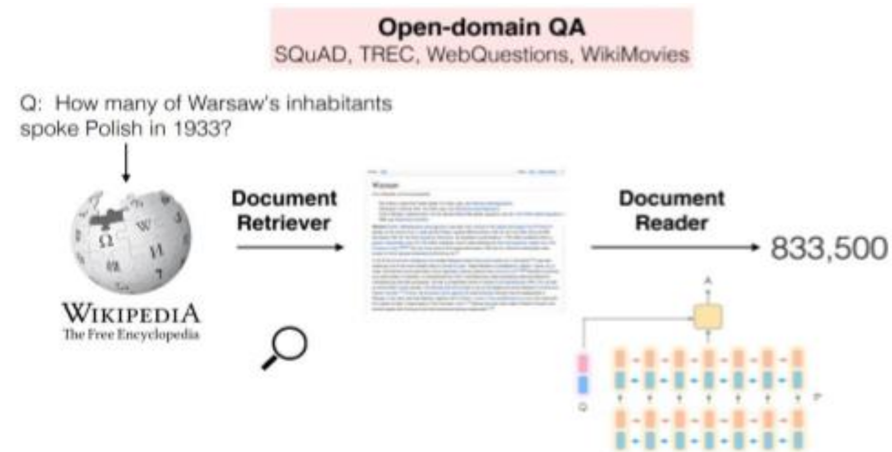
R1, http://www.erogol.com/brief-history-machine-learning/

R2, https://openai.com/blog/ai-and-compute/

Preliminaries

大数据时代的机遇 – 大量UGC

- Web 2.0时代，存在大量UGC（User Generated Content）
  - 提供获得广大用户一致认可的高质量数据源 e.g., Wikipedia，百度百科
  - 为自动挖掘知识提供了高质量的数据源
  - 为构建抽取模型提供了高质量的样本



Ref：Danqi Chen, etc. Reading Wikipedia to Answer Open-Domain Questions

周杰伦

版本对比	更新时间	全部版本	贡献者	修改原因	区块链信息
	2018-06-06 03:36	查看	w_ou	内链修复	查看
	2018-03-11 16:14	查看	[illegible]	内容扩充 内链	查看
	2018-03-01 20:20	查看	[illegible]	图片	查看
	2018-02-28 18:59	查看	[illegible]	内容扩充 参考资料	查看
	2018-02-15 08:05	查看	[illegible]510	内容扩充 参考资料	查看
	2018-02-11 20:54	查看	5ssax	更正错误 图片	查看
	2018-02-10 11:31	查看	Mini小北1992	完善作品信息	查看

Wiki和百科的编辑机制保证了UGC内容的质量

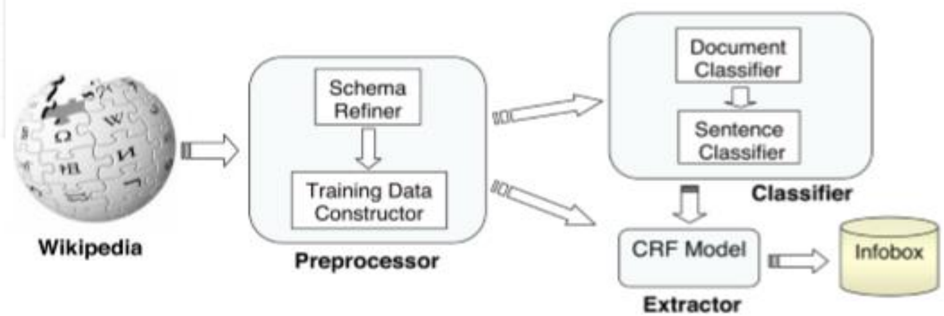


Figure 3: Architecture of KYLIN's infobox generator.

Ref： Fei Wu, etc. Autonomously Semantifying Wikipedia

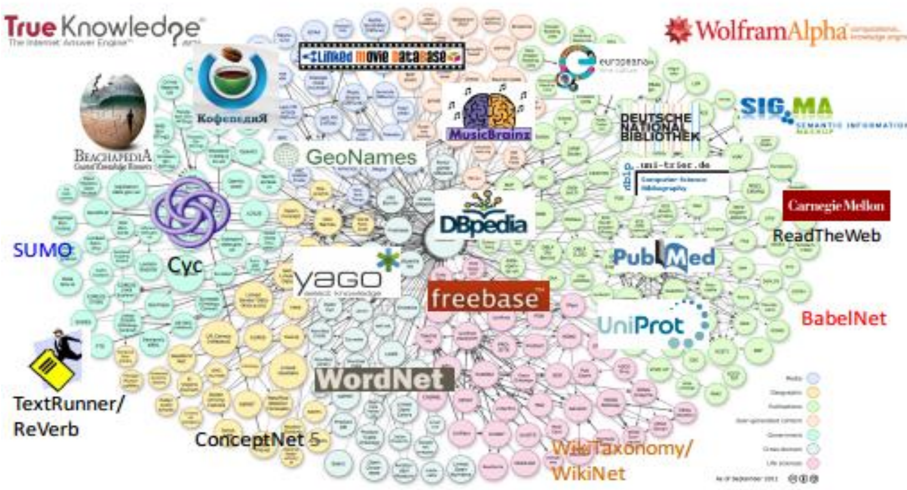
Preliminaries

大数据时代的到来，使得知识库技术突破了长久以来制约其发展的**规模**与**质量**瓶颈。知识图谱是这一突破的代表性产物。知识工程（KE）在知识图谱（KG）技术的引领下进入了全新的阶段（大数据时代的知识工程BigKE），BigKE将显著提升机器的认知水平。

本页PPT借鉴于复旦大学肖仰华老师
《Data-driven Approaches for Large-scale Knowledge Graph Construction》

Knowledge Graph – KG引领KE复兴

- Knowledge graph is a large-scale semantic network consisting of entities and concepts as well as the semantic relationships among them
  - Large scale
  - Semantically rich
  - Friendly structure
  - High quality



- Why knowledge graphs?
  - Understanding the semantic of text needs background knowledge
  - A robot brain needs knowledge base to understand the word

Preliminaries

Knowledge Graph – KG引领KE复兴

- Common large-scale KG

名称	开始时间	依赖资源	规模#（实体/概念/关系/事实）
Cyc/OpenCyc	1984	专家知识	239,261/116,822/18,014/2,093,000
WordNet	1985	专家知识	155,287/117,659/18/-
ConceptNet	1999	群体智能（多语言）	-/8,000,000/36/21,000,000
YAGO	2007	WordNet + Wikipedia	4,595,906/488,469/77/
DBpedia	2007	Wikipedia + 专家知识	17,315,785/754/2843/79,030,098
Freebase	2008	Wikipedia + 领域知识+ 群体智能	58,726,427/2,209/39,151/ 3,197,653,841
NELL	2010	机器学习	-/287/327/2,309,095
BabelNet	2012	WordNet + Wikipedia （多语言）	9,671,518/6,117,108/1,307,706,673 /-
Wikipedia	2012	Freebase + 群体智能	45,766,755/-/-/-
Google Knowledge Graph	2012	基于Freebase	570M/1500/35000/18000M
Knowledge Vault	2014	机器学习	45M/1100/4469/271M

本页PPT借鉴于复旦大学肖仰华老师《大数据时代的知识工程与知识管理》

Preliminaries

Knowledge Graph – KG引领KE复兴

知识图谱有着广泛的应用场景

Preliminaries

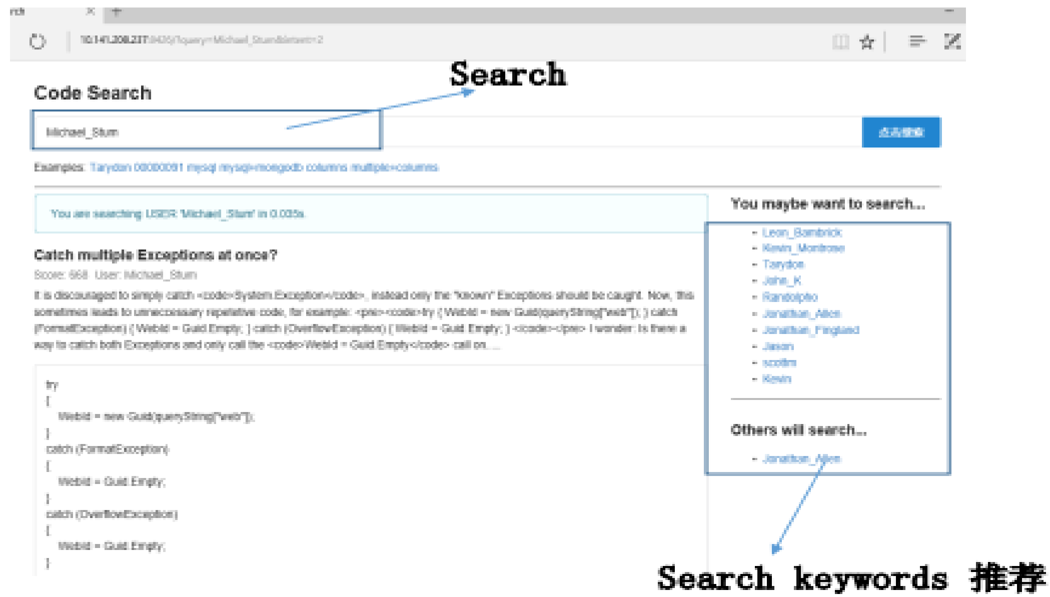
Knowledge Graph – 智慧搜索

- **精准搜索意图理解**
  - 精准分类
  - 语义理解
  - 个性化

- **Why knowledge graphs?**
  - 表格、文本、图片、视频
  - 文案、素材、代码、专家

- **多粒度搜索**
  - 篇章级、段落级、语句级

- **跨媒体搜索**
  - 不同媒体数据联合完成搜索任务



一切皆可搜索，搜索必达

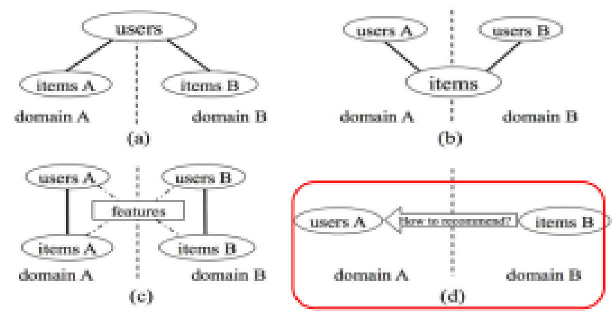
本页PPT借鉴于复旦大学肖仰华老师《大数据时代的知识工程与知识管理》

Preliminaries

Knowledge Graph – KG引领KE复兴

- 场景化推荐
- 任务型推荐
- 冷启动环境下的推荐
- 跨领域推荐
- 知识型推荐

电商领域的场景化推荐



跨领域推荐，比如给微博用户推荐Taobao商品，存在巨大的Vocabulary Gap

精准感知任务与场景，想用户之未想
从基于用户行为的推荐发展到行为与语义融合的智能推荐

本页PPT借鉴于复旦大学肖仰华老师《大数据时代的知识工程与知识管理》

Preliminaries

Knowledge Graph – 智能问答

人机交互方式将更加自然，对话式交互取代关键词搜索成为主流交互方式

一切皆可回答：图片问答、新闻问答、百科问答

目录 CONTENTS

1 The Pipeline of Knowledge Graph Construction by Data-driven manner

Data-driven **approaches** for large-scale KG construction

Pipeline of Knowledge Graph Construction by Data-driven manner

Data-driven VS Hand crafted

- **Manually constructed KG**
  - Examples：WordNet, Cyc
  - Size：**Small** (Huge human cost)
  - Quality：Almost **Perfect** (Each relation is checked by experts)

- **Auto-constructed KG**
  - Automatically extracted from huge Web Resource
  - Examples：Probase、WikiTaxonomy, etc
  - Size：**Huge** (From huge corpus)
  - Quality：**Good** (The accuracy can’t reach 100%)

    Because of the huge size, there are many wrong facts

Pipeline of Knowledge Graph Construction by Data-driven manner

Data-driven approaches for large-scale KG construction

- **知识抽取**
  - 属性抽取
  - 关系抽取
  - 实体抽取
- **知识融合**
  - 知识合并
  - 共指消解
  - 实体消歧
- **知识加工**
  - 知识推理
  - 质量评估
  - 本体构建
- **知识更新**

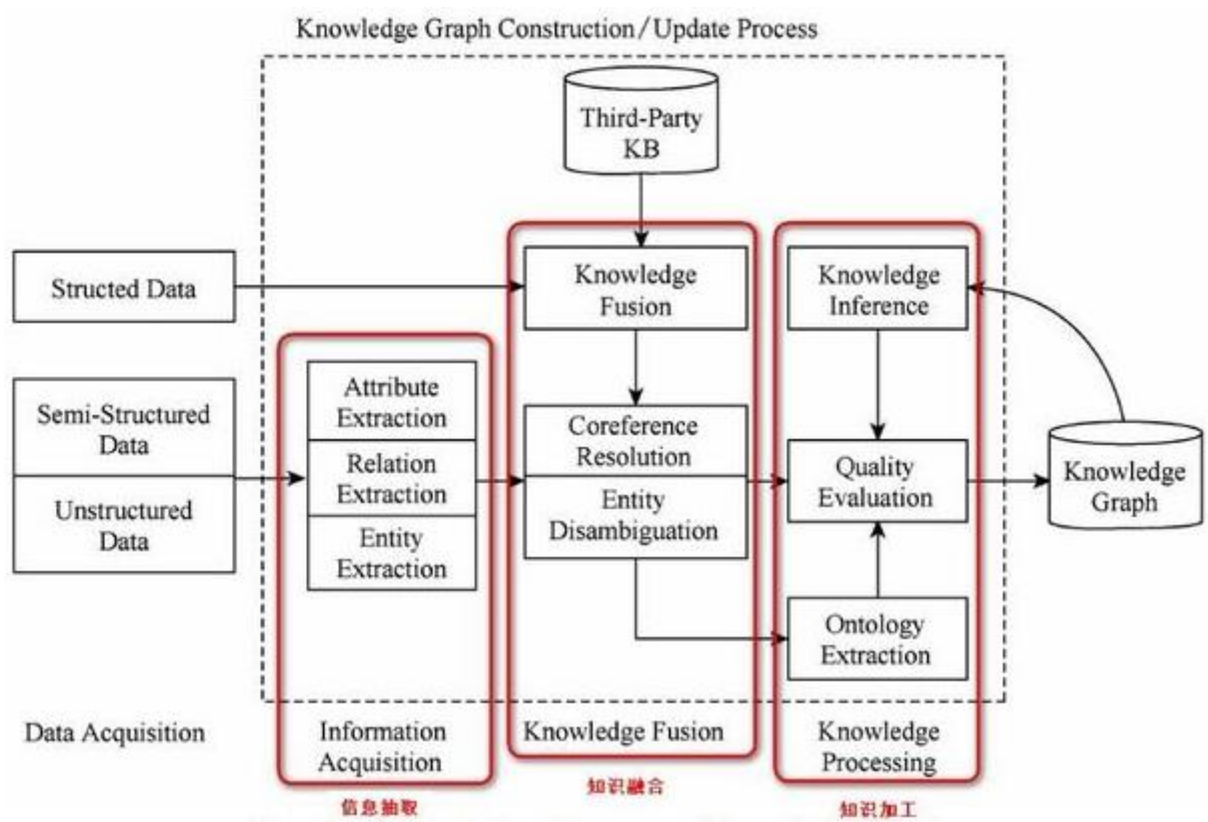


Figure 1: Data-driven KG construction techniques

刘峤，李杨等人, 知识图谱构建技术综述, 计算机研究与发展, 53 (3)，2016

Pipeline of Knowledge Graph Construction by Data-driven manner

Data-driven approaches for large-scale KG construction

- Data acquisition
  - 结构化的数据（工业界常用）
  - 半结构化的数据（工业界常用）
  - 非结构化的数据（学术界常用）

信息抽取方法相对简单，数据噪声小，经过人工过滤后能够得到高质量的三元组事实。

涉及的NLP分析与处理技术，难度较大。互联网的更多信息都是以非结构化的文本形式存在的。

- 知识抽取
  - **限定域关系抽取（判别的语义关系是预先定义的）**
    输入一个句子以及标识句子中所出现的实体指称的条件下，系统将其分类到所属的语义类别上（已有研究常把这一任务看成是一个文本分类问题）。
  - **开放域关系抽取（不需要预先定义关系，而是使用实体上下文中的一些词语来描述实体之间的关系）**
    e.g., 在语句“姚明出身在上海”中，通过开放域关系抽取方法抽取出的结果为（姚明，出生于，上海）
- 限定域关系抽取
  - **基于模板的关系抽取方法**
  - **基于机器学习的关系抽取方法**

Pipeline of Knowledge Graph Construction by Data-driven manner

Data-driven approaches for large-scale KG construction

- **基于模板的关系抽取**

  e.g., 用以下模板表示收购关系（acquisition）

  X is acquired by Y

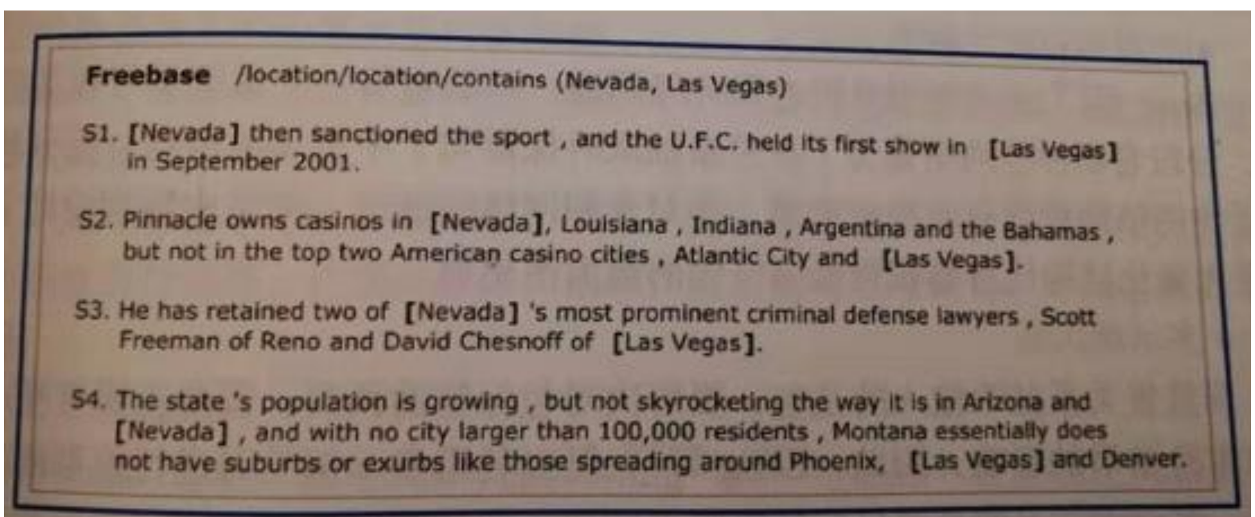
  X is purchased by Y

  X is bought by Y

- **基于机器学习的关系抽取方法**

- 有监督的关系抽取方法（e.g., 基于特征工程的方法，基于核函数的方法，基于神经网络的方法）
- 弱监督的关系抽取方法

Distant Supervision（远程监督），即如果两个实体之间存在某种关系，则所有包含这两个实体的句子都表达了这种关系，这些句子的集合被称为一个“包”。


Freebase /location/location/contains (Nevada, Las Vegas)

S1. [Nevada] then sanctioned the sport , and the U.F.C. held its first show in [Las Vegas] in September 2001.

S2. Pinnacle owns casinos in [Nevada], Louisiana , Indiana , Argentina and the Bahamas , but not in the top two American casino cities , Atlantic City and [Las Vegas].

S3. He has retained two of [Nevada] 's most prominent criminal defense lawyers , Scott Freeman of Reno and David Chesnoff of [Las Vegas].

S4. The state 's population is growing , but not skyrocketing the way it is in Arizona and [Nevada] , and with no city larger than 100,000 residents , Montana essentially does not have suburbs or exurbs like those spreading around Phoenix, [Las Vegas] and Denver.

Pipeline of Knowledge Graph Construction by Data-driven manner

Data-driven approaches for large-scale KG construction

- 开放域关系抽取（Open-domain Information Extraction, Open IE）
  - 华盛顿大学的AI研究小组最早提出Open IE的想法
  - TextRunner、Kylin、WOE、ReVerb等系统相继被开发

- **以TextRunner为例进行介绍（核心：将动词作为关系名称，通过动词链接两个论元，从而挖掘论元之间的关系）**

1、**语料自动生成**：主要通过依存句法分析，结合启发式的规则自动生成语料
E.g., 启发式的规则包括：关系指代词是两个实体之间依存路径上的**动词**或**动词短语**。

2、**分类器的训练**：利用朴素贝叶斯分类器进行训练，其使用的特征包括：关系指示词的词性、实体的类型等。

3、**关系三元组的抽取**：利用训练好的分类器对Web文本上的三元组进行抽取。

4、**关系三元组可信度计算**：将存储起来的相似三元组进行合并，然后根据网络数据的冗余性，计算合并后的三元组在Web文本中出现的次数。

- Open IE方法普遍存在的问题：（1）**三元组识别错误**（incoherent extractions）；（2）**无信息三元组抽取**（un-informative extractions）

Michele, Michael, Stephen, et al., Open information extraction from the web, IJCAI, 2007.

Pipeline of Knowledge Graph Construction by Data-driven manner

Data-driven approaches for large-scale KG construction

- 知识融合（为跨领域的信息需求提供服务）

  从融合的知识图谱类型来看：

- **垂直方向的融合**（融合较高层通用本体与较低层领域本体或实例数据）

- **水平方向的融合**（融合相同层次的知识图谱）

- 知识融合中的关键技术
- **匹配框架**（元素级、结构级的匹配）
- **实体对齐**（e.g., **等价关系合并**；互动百科与百度百科中的实体“刘洋”描述的是同一个对象）
- **冲突检测与消解**（使多个知识图谱形成一致的结果）

- 典型的知识融合系统
- AgreementMaker：一个集成系统，包含了若干自动对齐的方法
- Falcon：一个采用分治法设计的对齐系统
- RiMOM：一个采用动态多策略的对齐框架

Pipeline of Knowledge Graph Construction by Data-driven manner

Data-driven approaches for large-scale KG construction

- 知识加工
  - **知识推理**
  - **质量评估**
  - **本体构建**

- 知识推理
  - **基于符号演算的推理**（逻辑上）
  - **基于数值计算的推理**（基于张量分解的方法、基于能量函数的方法）
  - **符号演算和数值计算的融合推理**
  - **常识知识推理**

- 质量评估
  - 对知识的可信度进行量化，通过舍弃置信度较低的知识，从而保障知识库的质量

Pipeline of Knowledge Graph Construction by Data-driven manner

Data-driven approaches for large-scale KG construction

- 知识更新
  - 知识更新是一个不断迭代更新的过程
  - 从逻辑与内容两个方面来分析

- 从逻辑层面来分析
  - **概念层的更新**（往往需要借助于专业团队来完成）
  - **数据层的更新**（包括新增或更新实体、关系，以及属性）

    当前流行的方法是选择百科等可靠知识库，将其中出现频率较高的事实和属性加入到知识库

- 从知识图谱的内容层面来分析
- **数据驱动的全面更新**（资源消耗极大，从1到全部）
- **增量式更新**（添加新增知识，资源消耗小）

2 Python Tools of Graph Data Management

Stanford CoreNLP

KnowItAll system

networkx

Gephi

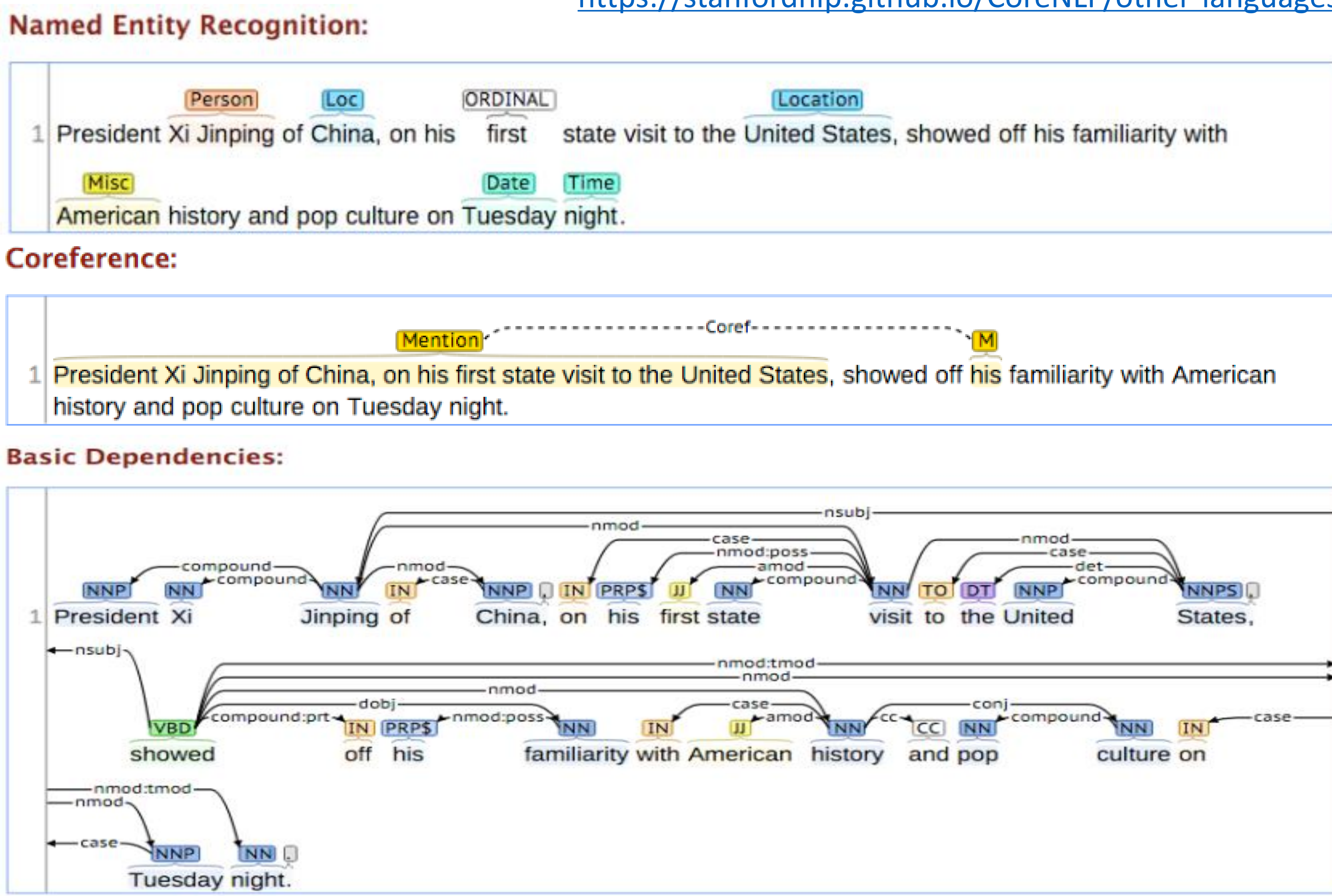
Python Tools of Graph Data Management

Stanford CoreNLP

https://stanfordnlp.github.io/CoreNLP/

Python API interface:

https://stanfordnlp.github.io/CoreNLP/other-languages.html



Python Tools of Graph Data Management

KnowItAll system: https://github.com/knowitall/openie

Python 调用 Java KnowItAll接口（Jpype实现python调用）：
https://blog.csdn.net/fengmm521/article/details/78446431

KnowItAll system

Example 1:

```
The U.S. president Barack Obama gave his speech on Tuesday to thousands of people.
```

```
(Barack Obama, is the president of, the U.S.)
(Barack Obama, gave, his speech)
(Barack Obama, gave his speech, on Tuesday)
(Barack Obama, gave his speech, to thousands of people)
```

N-ary relation extraction

Example 2:

```
> John ran down the road to fetch a pail of water.
John ran down the road to fetch a pail of water.
0.86 (John; ran; down the road; to fetch a pail of water)
0.82 John ran:(John; ran down the road to fetch; a pail of water)
```

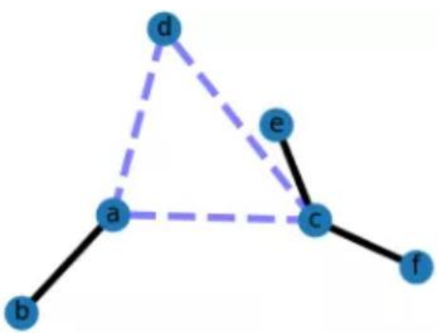
Each extraction associates a confidence score

Python Tools of Graph Data Management

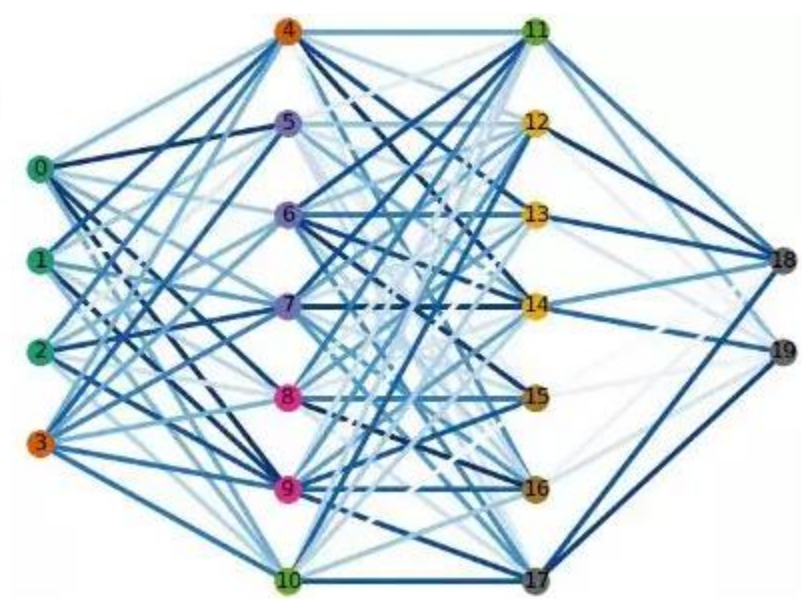
networkx

https://mp.weixin.qq.com/s/WYM7k9gddAndlLBuQWTbSA

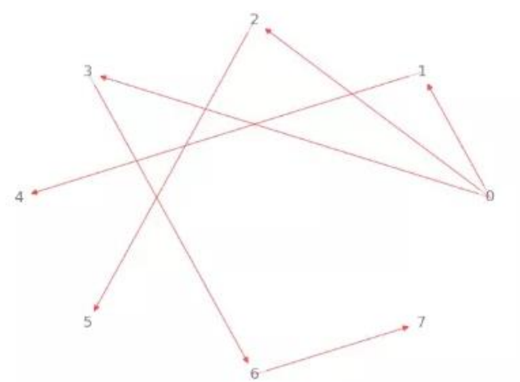
- Networkx是一个基于python的复杂网络分析库，内置了常用的图与复杂网络分析算法，可以方便的进行复杂网络数据分析、仿真建模等工作。
- 生成随机网络、经典网络、建立网络模型、网络绘制
- 以图（简单无向图、有向图、多重图等）为基本数据结构，支持通过在线数据源生成图结构



权重图



多层感知机网络



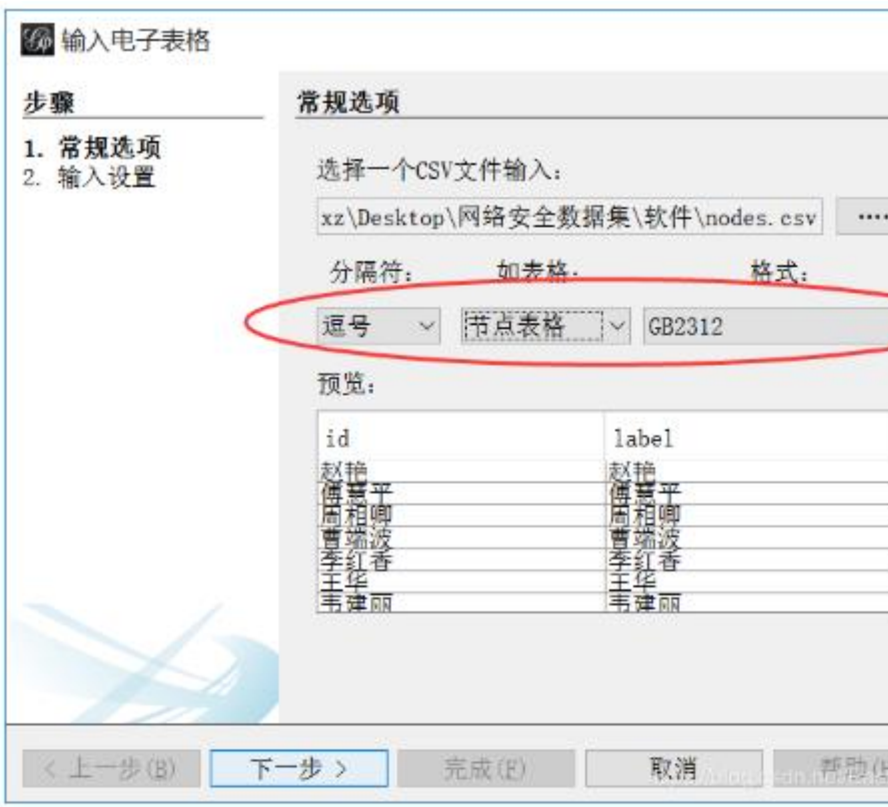
最短路径算法

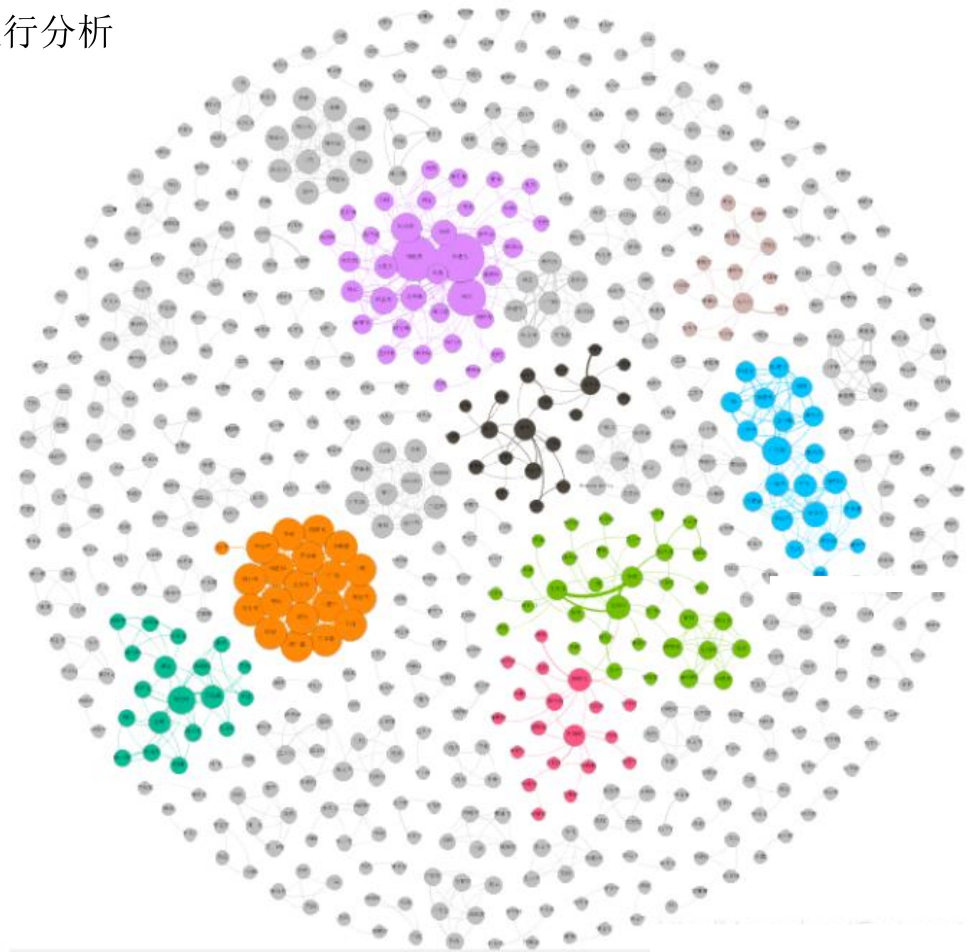
Python Tools of Graph Data Management

Gephi

https://download.csdn.net/download/u011217593/8091061

- Gephi是一款基于JVM的复杂网络分析软件
- 支持多种复杂的网络结构
- 能够通过图密度分析、PageRank的算法对网络进行分析





3 Domain-specific Knowledge Graph Construction

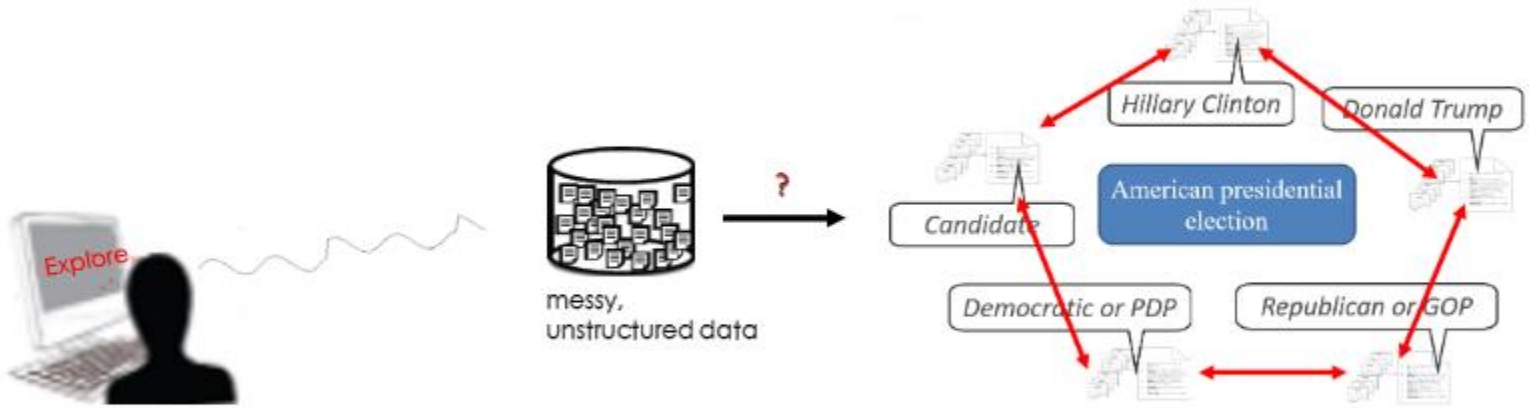
A Conceptual Knowledge Graph oriented News Data

A Conceptual Knowledge Graph oriented News Data

PyConChina 2019

Motivation

With the emergence of massive text corpora in many domains and languages, the sheer size and rapid growth of this new data poses many challenges understanding and connecting significant insights from these massive unstructured texts.



PYTHON 30th

A Conceptual Knowledge Graph oriented News Data

Motivation

How to mine and organize meaningful concepts and their semantic connections from a set of related documents under the same topic.

Traditional relation extraction systems require people to the pre-specify the set of relations of interest. Obviously, it is not appropriate for the news documents with diverse relation schemas.

Given a query topic, a user is often expected to understand core topic information serving by a large conceptual graph, rather than having a collection of relevant documents.

A Conceptual Knowledge Graph oriented News Data

We present **a system** that extracts salient entities, concepts, and their relationships from a set of related documents, discovers connections within and across them, and presents the resulting information in a graph-based visualization.

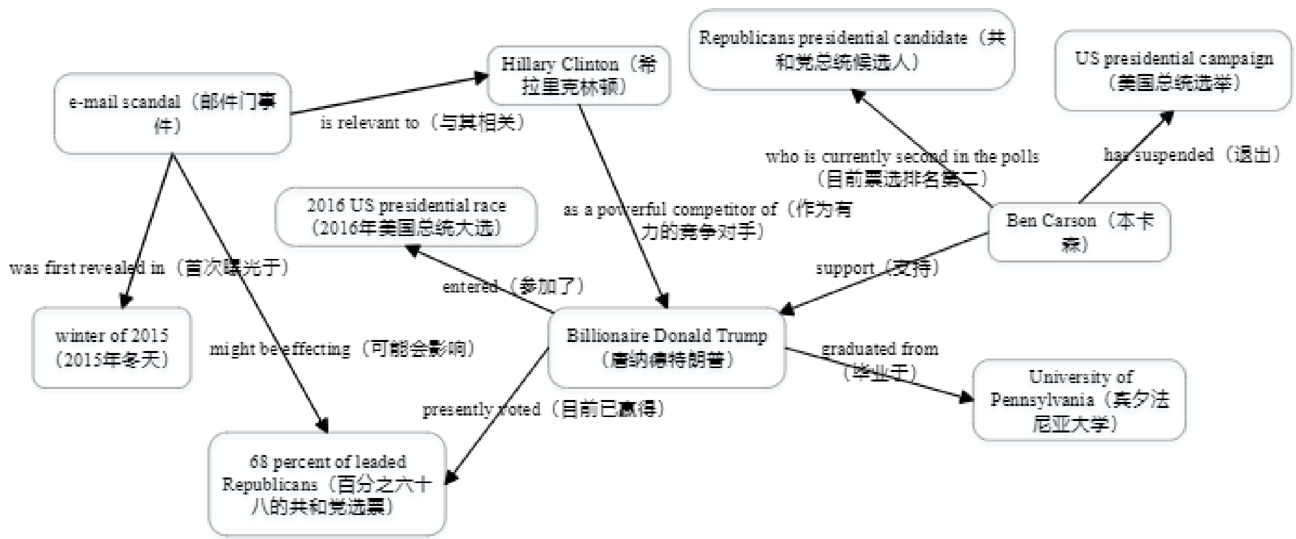


Figure 1: Example of a conceptual graph on the topic “presidential election of the US”

A Conceptual Knowledge Graph oriented News Data

The objective of our system is to assist users in quickly finding meaningful and salient connections and facts from a collection of relevant documents, and in summary, it can be best described as a combination of **three major subtasks**:

- **Subtask 1: Candidate Fact Extraction**

  Given a collection of documents $D = \{d_1, d_2, ..., d_M\}$ clustered around a topic T. The goal of this subtask is to extract a set of facts $F_c = \{f_1, f_2, ..., f_N\}$ from D. Each of facts is essentially (s, r, o) triple, for *subject* s, *relation* r, and *object* o. Since we need to estimate the coherence of these preferred facts for T, we refer to them as ***candidate facts***.

- **Subtask 2: Fact Filtering**

  Given a specified document topic T, the goal of the subtask is to find a subset of $F_t \subseteq F_c$ and each of them should be coherent with T.

- **Conceptual Knowledge Graph Construction**

  The goal of the subtask is to determine which of the facts from $F_t \subseteq F_c$ generated by the previous subtask are more likely to be salient, which of their entities and concepts to merge and, when merging, which of the available labels to leverage in the final conceptual graph G.

A Conceptual Knowledge Graph oriented News Data

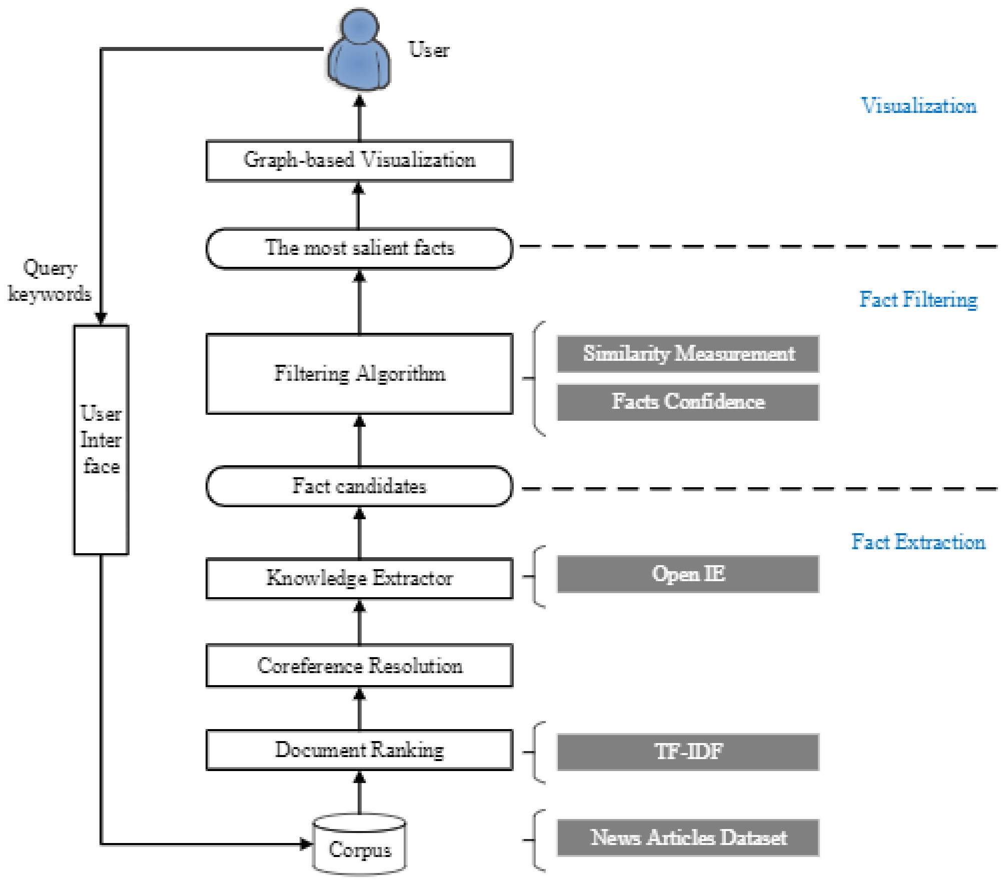


Figure 2: System architecture

Approaches and Implementation

- News data
- Fact Extraction
- Fact Filtering
- Conceptual Graph Construction

A Conceptual Knowledge Graph oriented News Data

News data

- Our dataset include **5 categories**, and for each category we have **2 popular events** and each of which represents a document topic. Every topic cluster comprises approximately **30 documents** with on average 1,316 tokens, which leads to an average topic cluster size of 2,632 tokens. It is **3 times** larger than typical DUC[1] clusters of 10 documents.

- The articles in our dataset stem from a larger news document collection released by **Signal Media** as well as crawled from **Web Blogs** by ourselves, we rely on event keywords to filter them so as to retain related ones for different topics.

[1] Document Understanding Conference, https://duc.nist.gov/

A Conceptual Knowledge Graph oriented News Data

News data

Table 1: Dataset description

Category	Topic ID	Document topic	Time period	Docs	Doc.Size	Source
Armed conflicts and attacks	1	Syria refugee crisis	2015-09-01 - 2015-09-30	30	2179 ± 506	News, Blog
	2	North Korea nuclear test	2017-08-09 - 2017-11-20	30	1713 ± 122	News
Business and economy	3	Chinese cooperation with Sudan	2015-09-01 - 2015-09-30	30	768 ± 132	News, Blog
	4	Trump TPP	2016-12-23 - 2017-02-23	30	879 ± 306	News
Politics and elections	5	US presidential election	2016-06-14 - 2016-08-14	30	1175 ± 207	News, Blog
	6	US-China trade war	2018-03-23 - 2018-06-15	30	2412 ± 542	News, Blog
Arts and culture	7	Muslim culture	2013-02-01 - 2013-05-01	30	972 ± 161	News, Blog
	8	Turing Award winner	2019-03-15 - 2019-04-01	30	1563 ± 464	News, Blog
Information technology and application software	9	Next-generation search engine	2016-11-07 - 2017-01-03	30	729 ± 280	News, Blog
	10	Program repair for Android system	2018-02-01 - 2018-05-10	30	772 ± 453	Blog

Approaches and Implementation

- News data
- Fact Extraction
- Fact Filtering
- Conceptual Graph Construction

Open-Domain Knowledge Extraction

- **Document Ranking**. The system first select the words appearing in the document collection with sufficiently high frequency as topic words, and computes standard **TF-IDF weights**[2] for each word. Documents under the same topic are ranked according to the TF-IDF weights of the topic words in each document. The top-k documents for every topic are selected for further processing.
- **Coreference Resolution**. Pronouns and other form of coreference are resolved in each document using **Stanford CoreNLP system**[1]. "she" may be replaced by "Angela Merkel", for instance.
- Sentence Ranking. Our system computes the **TextRank importance scores**[4] for all sentences within the ranked top-k document list. It then considers only those sentences with sufficiently high scores.

[1]https://stanfordnlp.github.io/CoreNLP/index.html
[2]https://en.wikipedia.org/wiki/Tf%E2%80%93idf
[4]https://github.com/letiantian/TextRank4ZH/blob/master/README.md

Open-Domain Knowledge Extraction

- Our candidate fact extraction is based on a publicly available system for open information extraction, namely the KnowItAll project's Open IE 4[4].

Considering an example consisting of the following two sentences:

"George Washington was the first President of the United States, the Commander-in-Chief of the Continental Army during the American Revolutionary War."

- 0.95 ("George Washington", "was", "the first President of the United States")
- 0.88 ("George Washington", "was", "the Commander-in-Chief of the Continental Army")

[4]https://github.com/knowitall/openie

Open-Domain Knowledge Extraction

Considering an example consisting of the following two sentences:

*"**He** presided over the convention that drafted the current United States Constitution and during his lifetime was called the 'father of his country' "*

- 0.45 ("**He**", "presided", "over the convention")
- 0.90 ("the convention", "drafted", "the current United States Constitution")

Noting that:

When the ambiguous pronoun "**He**" is replaced with "**George Washington**",

- 0.93 ("George Washington", "presided", "over the convention")

Approaches and Implementation

- News data
- Fact Extraction
- Fact Filtering
- Conceptual Graph Construction

A Conceptual Knowledge Graph oriented News Data

Fact Filtering

Our candidate fact The filtering algorithm aims at hiding less representative facts in the visualization, seeking to retain only the most **salient**, **confident** and **compatible** facts. This is achieved by optimizing for a high degree of coherence between facts with high confidence.

The joint optimization problem can be solved via integer linear programming (ILP), as follows:

$$\max_{x,y} \quad \alpha^T x + \beta^T y \tag{1}$$

$$\text{s.t.} \quad 1^T y \leq n_{\max} \tag{2}$$

$$x_k \leq \min\{y_i, y_j\} \tag{3}$$

$$\forall \; i < j, i, j \in \{1, \ldots, M\},$$

$$k = (2M - i)(i - 1)/2 + j - i$$

$$x_k, y_i \in \{0, 1\} \, \forall i \in \{1, \ldots, M\}, k \tag{4}$$

Fact Filtering

ILP method:

Here, $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{y} \in \mathbb{R}^M$ with $N = (M+1)(M-2)/2 + 1$. The y_i are indicator variables for facts t_i: If y_i is true, t_i is selected to be retained. x_k represents the compatibility between two facts $t_i, t_j \in T$ $(i, j \leq M,\ i \neq j)$, where $T = \{t_1, \ldots, t_M\}$ is a set of fact triples containing M elements. β_i denotes the confidence of a fact, and $n_{\max}$ is the number of representative facts *desired by the user*. α_k is weighted by similarity scores $sim(t_i, t_j)$ between two facts t_i, t_j, defined as $\alpha_k = sim(t_i, t_j) = \gamma \dot{s}_k + (1-\gamma)\dot{l}_k$. Here, s_k, l_k denote the semantic similarity and literal similarity scores between the facts, respectively. We compute s_k using the *Align, Disambiguate and Walk* algorithm, l_k are computed using the Jaccard index. $\gamma = 0.8$ denotes the relative degree to which the semantic similarity contributes to the overall similarity score, as opposed to the literal similarity. The constraints guarantee that the number of results is not larger than $n_{\max}$. If x_k is true, the two connected facts t_i, t_j should be selected, which entails $y_i = 1$, $y_j = 1$.

Approaches and Implementation

- News data
- Fact Extraction
- Fact Filtering
- Conceptual Graph Construction

Conceptual Knowledge Graph Construction

- **Merge Equivalent Concepts and Add Relations**.
- **Literal features of entities**. e.g., Billionaire Donald Trump, Donald Trump, Donald John Trump, Trump, etc. all refer to the same person.
- **Entity linking from search engine**. For **NER**, they can use the powerful entity linking ability from a search engine for deciding on coreference. ADW[5] tool is used for semantically similarity computation between concepts for coreference.
- Annotators were able to add up to **three synthetic relations with freely defined labels** to connect the subgraphs into a fully connected graph.

[5]Align, Disambiguate and Walk, https://github.com/pilehvar/ADW

Experiments – Experimental setting

Parameter setting

- **Sentence-level extractions**. We first randomly sample 10 documents from every document topics (100 documents in total) and perform coreference resolution. Then, once again a random sample of 10 sentences from every extracted document (1,000 sentences in total) for further analysis. Each sentence is examined by three expert annotators with NLP background independently to annotate all of correct triples[a].
- **An empirical study**. We further conduct to investigate the quality of the final generated conceptual graph towards different document topics on its coverage rate of topic entities and concepts, confidence score, and the compatibility of involved facts.

[a] A triple is annotated as correct if the following conditions are met: i) it is entailed by its corresponding clause; ii) it is reasonable or meaningful without any context and iii) when these three annotators mark it correct simultaneously (The inter-annotator agreement was 82% ($\kappa = 0.60$))

A Conceptual Knowledge Graph oriented News Data

Experiments – Evaluation measures

Evaluation measures

Three standard metrics

Sentence-level extractions

- Precision (P)
- Recall (R)
- F-score (F1)

An empirical study

Quality of the conceptual graph

- $TopicCon\_Coverage$
- $Avg\_Confidence$
- $Avg\_Compatibility$

Experiments – Performance analysis of our extraction approach

Table 1: Evaluation of precision, recall, and F-score on five independent document topics (including topic 1 to topic 5) from two datasets

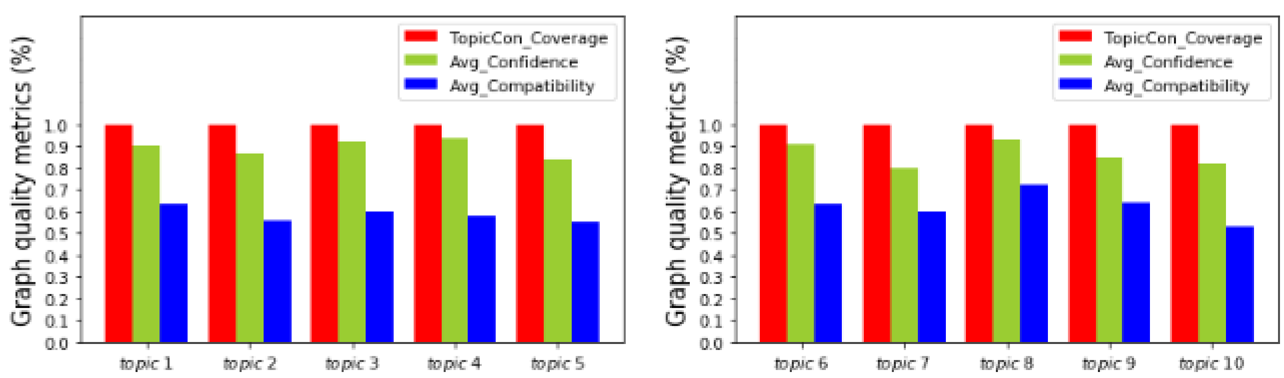
OpenIE methods	#Topic 1			#Topic 2			#Topic 3			#Topic 4			#Topic 5		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Our approach (without coref)	0.43	0.29	0.56	0.44	0.27	0.33	0.65	0.24	0.35	0.47	0.33	0.39	0.45	0.30	0.36
Our approach	0.86	0.85	**0.85**	0.78	0.74	**0.76**	0.95	0.92	**0.93**	0.95	0.82	**0.88**	0.92	0.78	**0.84**

Experiments – Performance analysis of our extraction approach

Table 1: Evaluation of precision, recall, and F-score on five independent document topics (including topic 6 to topic 10) from two datasets

OpenIE methods	#Topic 6			#Topic 7			#Topic 8			#Topic 9			#Topic 10		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Our approach (without coref)	0.43	0.29	0.35	0.44	0.32	0.37	0.47	0.30	0.37	0.55	0.42	0.48	0.40	0.29	0.34
Our approach	0.90	0.73	**0.81**	0.78	0.69	**0.73**	0.95	0.78	**0.86**	0.88	0.73	**0.80**	0.78	0.74	**0.76**

Experiments – Quality analysis of the conceptual knowledge graph



(a) (b)

A Conceptual Knowledge Graph oriented News Data

The results indicates that:

Our approach achieved 100% coverage rate of topic entities and concepts ($TopicCon\_Coverage$), 87% confidence score ($Avg\_Confidence$), and 68% fact compatibility ($Avg\_Compatibility$) over ten document topics.

- The proposed fact filtering approach is capable to select high confident and salient facts from the extracted candidate facts, however, may not guarantee their better compatibility, which needs to be further explored.
- The final generated conceptual graph has higher coverage rate of topic entities and concepts, which demonstrate the importance of the heuristic strategy in the process of conceptual graph construction.

A Conceptual Knowledge Graph oriented News Data

Conclusions

- Our system is intended to aid users in quickly discerning salient connections in a collection of documents, including via graph-based visualizations. Experiments on two real-world datasets demonstrate the effectiveness of our proposed approach.

Future Work

- The fact filtering algorithm will give greater consideration to the context of the triples, to enhance compact connections.
- The fact fusion problem in generating the final conceptual graph needs to be further explored for the fully automated conceptual graph construction for specified domain is possible.

Codes and Datasets

- We release the codes and datasets related to this system at: https://shengyp.github.io/vmse.

THANK YOU

shengyp2011@163.com