知识图谱融合方法

胡伟

南京大学

whu@nju.edu.cn

提纲

1. 概述
2. 预备知识
3. 本体匹配

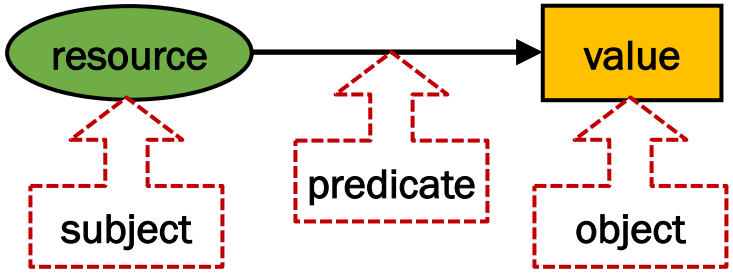
4. 实体对齐
5. 知识融合
6. 总结与展望

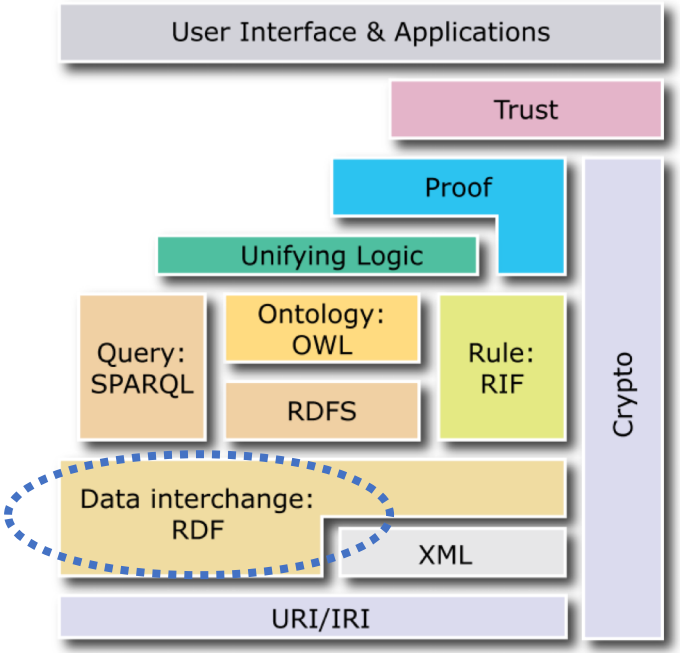
1. 概述

语义 (万维) 网

- Semantic Web 源自 Tim Berners-Lee
  - Give formal meanings to web info → semantics
    - Web 1.0 → Web 2.0 → Web 3.0 (a web of data)
- Semantic Web is about
  - Common formats for
    - integration and combination of data drawn from diverse sources
  - Languages for
    - recording how the data relates to real-world objects

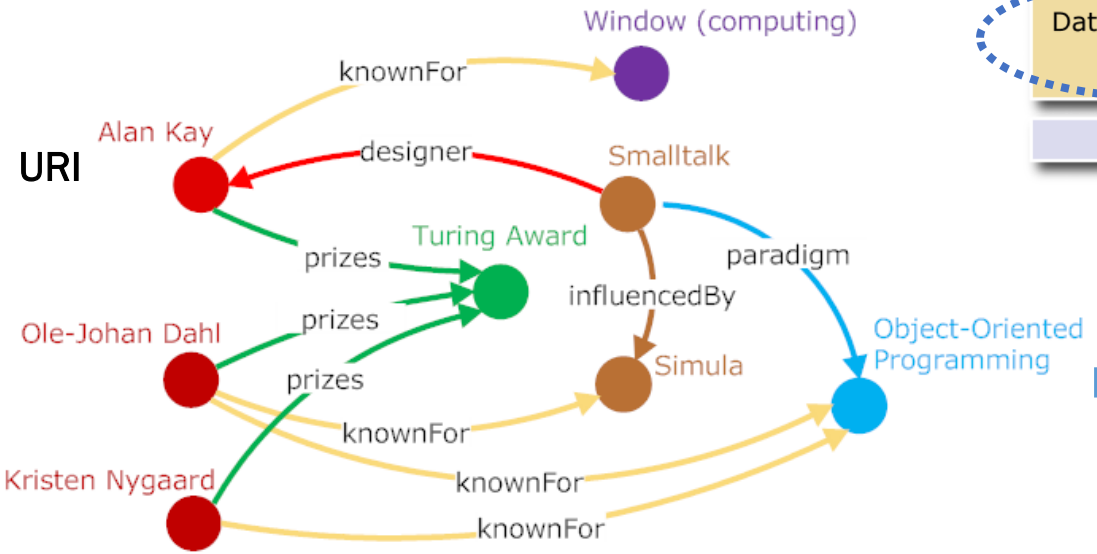
RDF 数据模型





Layer cake

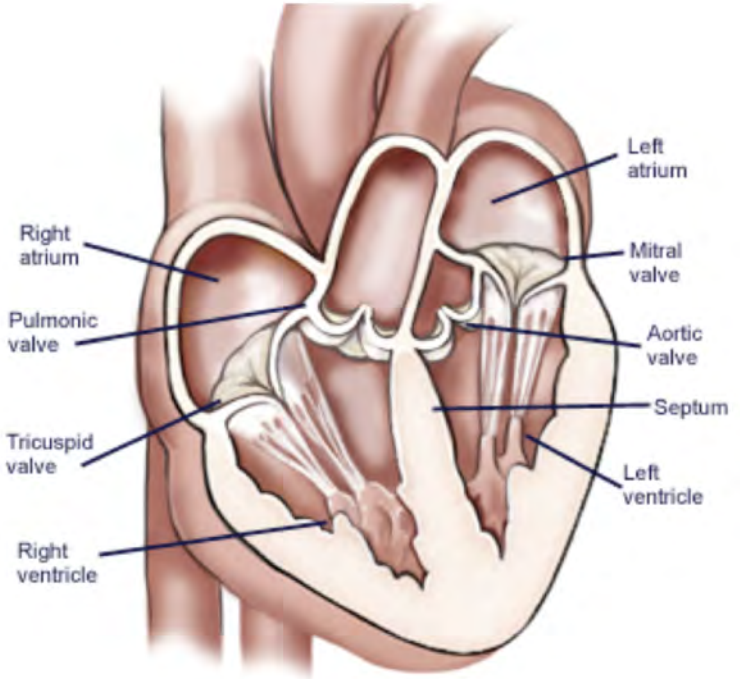
URI



The world is not made of strings, but is made of things

本体 (Ontology)

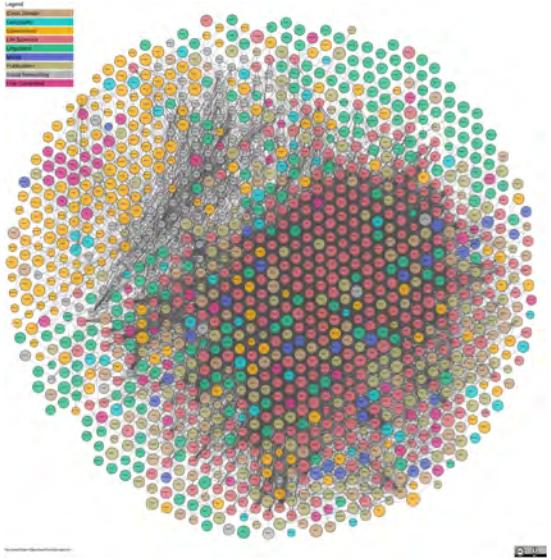
- 本体是领域知识规范的抽象和描述，是表达、共享、重用知识的方法
  - (部分) 真实世界的一个模型
    - 引入领域相关的术语集
      - » 概念、属性 .....
    - 指定术语的含义 (语义)
  - 使用合适的逻辑来形式化
    - 描述逻辑
      - » 一阶谓词逻辑的一个可判定子集
        - Heart is a muscular organ that is part of the circulatory system



$$\text{Heart} \sqsubseteq \text{MuscularOrgan} \sqcap \exists \text{isPartOf}.\text{CirculatorySystem}$$

链接数据 / 关联数据

- As a realization of Semantic Web
  - Linked Data refer to a collection of interrelated datasets
    - Used for large-scale integration of, reasoning on, data on the web
- Linked data principles
  1. Use URIs to name things
  2. Use HTTP URIs (can be “dereferenced”)
  3. Provide useful information using the open web standards (e.g. RDF)
  4. Include links to other related things



知识图谱

- Knowledge Graph is a **knowledge base** used by Google to enhance its search engine's search results with semantic-search information gathered from a wide variety of sources
  - 知识图谱 2012 年 5 月由 Google 正式提出，其初衷是为了提高搜索引擎的能力，改善用户的搜索质量及体验
  - 可看作是一张巨大的图，节点表示实体或概念，边则由属性或关系构成

大规模知识图谱

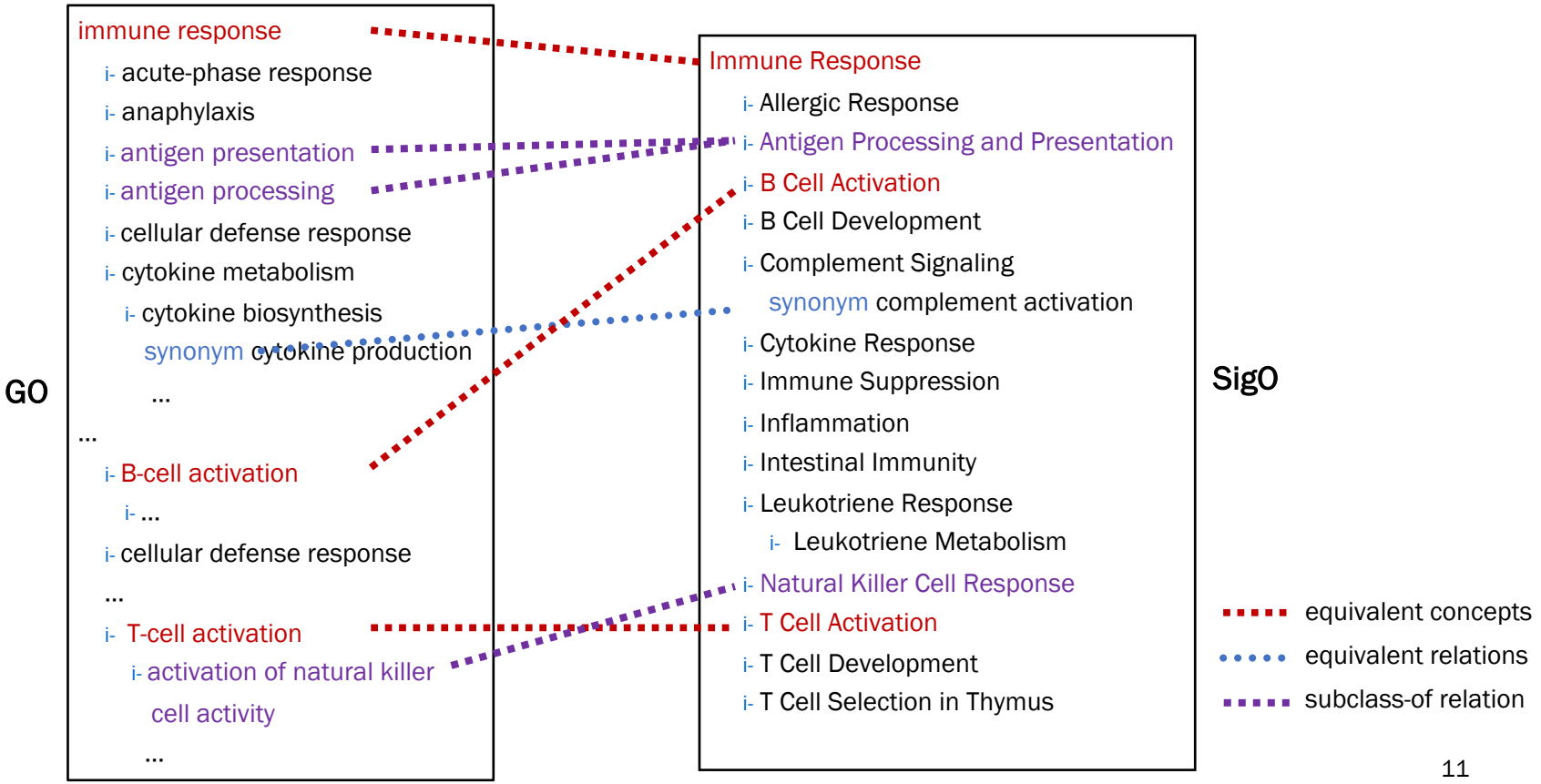
名称	规模
DBpedia	英文：4 百万个实体，5 亿个 RDF 三元组 685 个概念，2,795 个属性 125 种语言
YAGO	1 千万个实体，1.2 亿个 RDF 三元组 35 万个概念，100 个属性
Freebase	4 千万个实体，10 亿个 RDF 三元组 1.5 万个概念，4,000 个属性
Google	6 亿个实体，35 亿条 RDF 三元组
Wikidata、Wolfram Alpha、CMU NELL、阿里巴巴藏金阁、百度知心、搜狗知立方 …	

Adoption of Knowledge Graphs

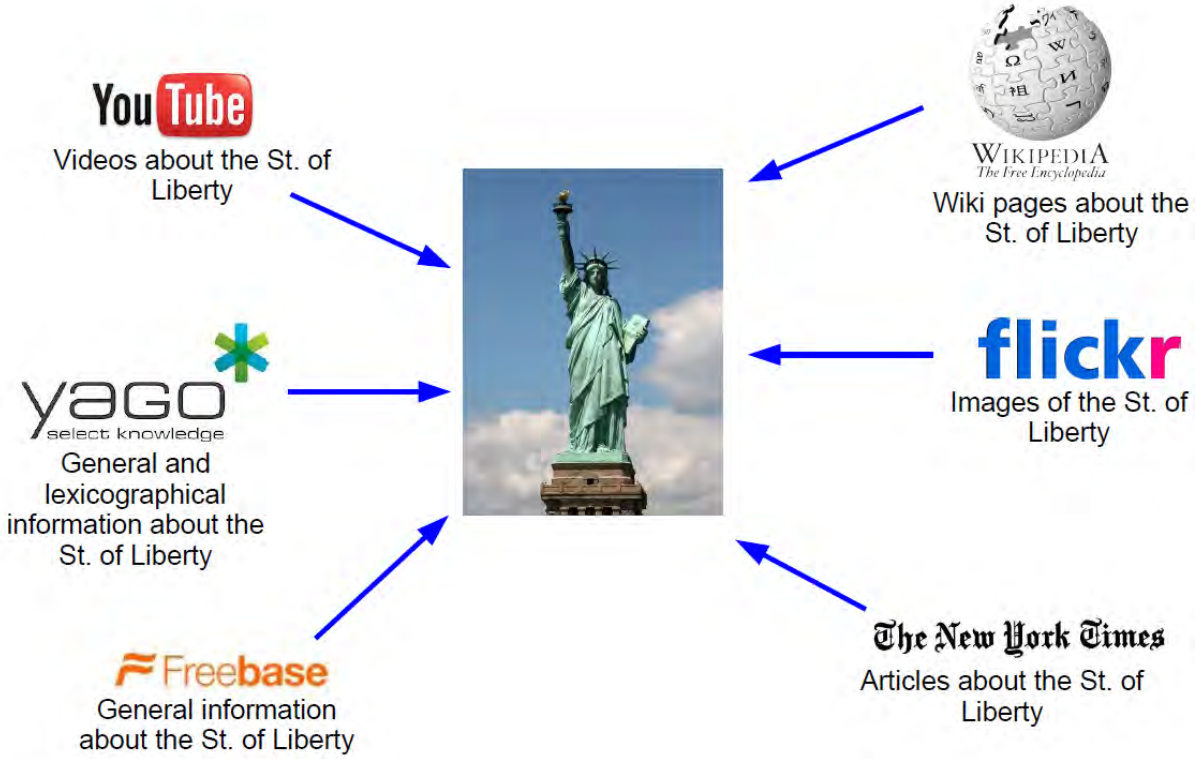
- Late 2019
  - By Frank van Harmelen

来源：https://www.slideshare.net/Frank.van.Harmelen

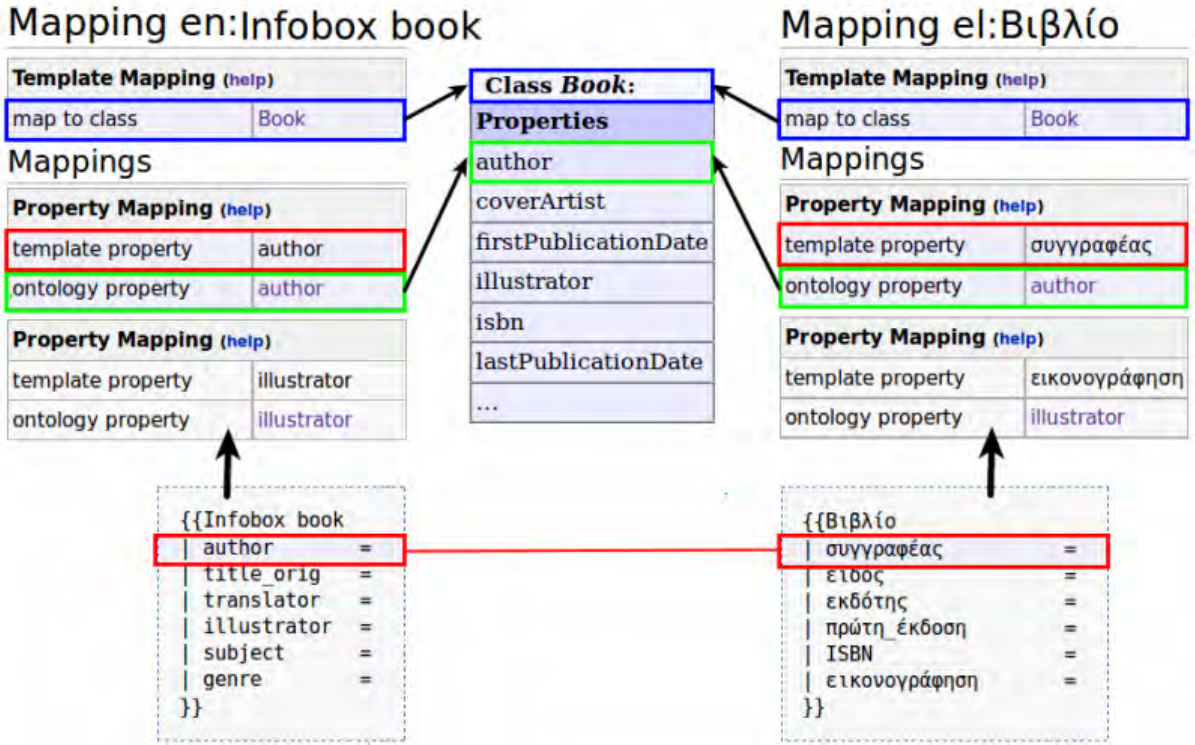
不同本体


GO
immune response
i- acute-phase response
i- anaphylaxis
i- antigen presentation
i- antigen processing
i- cellular defense response
i- cytokine metabolism
i- cytokine biosynthesis
synonym cytokine production
...
...
i- B-cell activation
i- ...
i- cellular defense response
...
i- T-cell activation
i- activation of natural killer cell activity
...
SigO
Immune Response
i- Allergic Response
i- Antigen Processing and Presentation
i- B Cell Activation
i- B Cell Development
i- Complement Signaling
synonym complement activation
i- Cytokine Response
i- Immune Suppression
i- Inflammation
i- Intestinal Immunity
i- Leukotriene Response
i- Leukotriene Metabolism
i- Natural Killer Cell Response
i- T Cell Activation
i- T Cell Development
i- T Cell Selection in Thymus
equivalent concepts
equivalent relations
subclass-of relation

不同实例


YouTube
Videos about the St. of Liberty
WIKIPEDIA
The Free Encyclopedia
Wiki pages about the St. of Liberty
yago
select knowledge
General and lexicographical information about the St. of Liberty
flickr
Images of the St. of Liberty
Freebase
General information about the St. of Liberty
The New York Times
Articles about the St. of Liberty

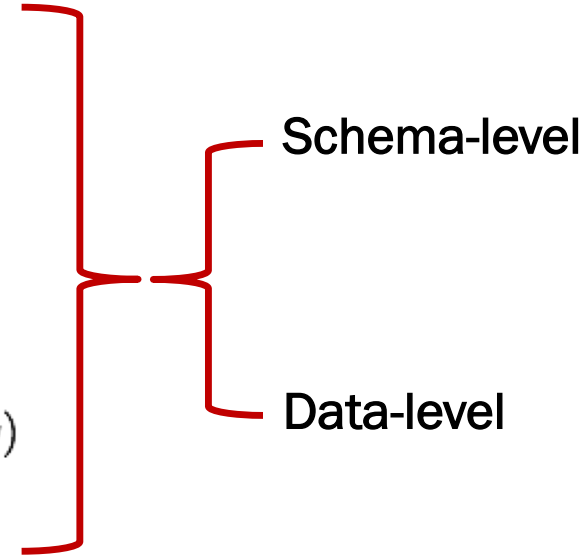
跨语言



http://mappings.dbpedia.org/

异构性

- Since long long time ago …
  - Syntactic
    - e.g., “Wei Hu” vs. “HU, Wei”
  - Terminological
    - e.g., “notebook” vs. “laptop”
  - Semantic
    - e.g., $\mathrm{hasSon}(x, y)$ vs. $\mathrm{hasChild}(x, y) \sqcap \mathrm{Male}(y)$
  - Pragmatic

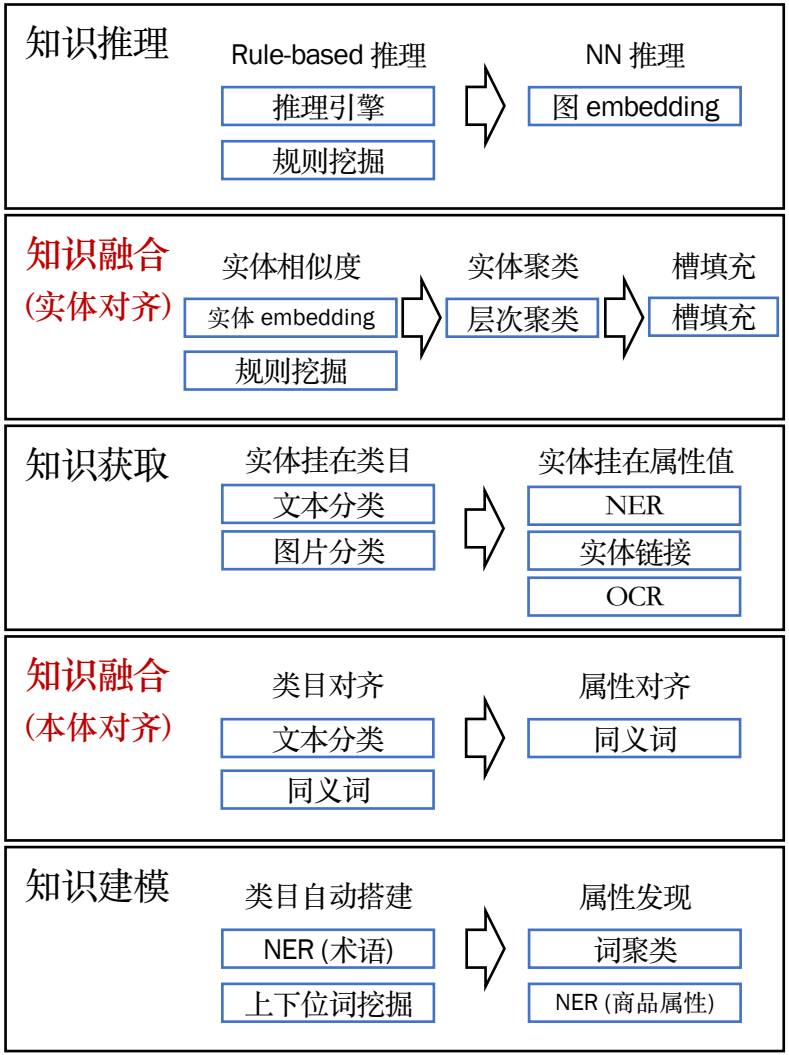


Knowledge graphs have reached a scale in billions of triples

阿里巴巴知识引擎

- 定义五大技术模块，并开发落地
  - 知识获取、知识建模、知识推理
  - 知识融合
    - 对异构和碎片化知识进行语义集成，通过发现知识之间的关联，获得更完整的知识描述和关联关系，实现知识互补和融合
  - 知识服务
    - 已在淘宝、天猫、盒马、飞猪等几十种产品上取得了成功应用
    - 每天有 8000 多万次在线调用，日均离线输出 9 亿条知识

来源：http://www.elecfans.com/d/908682.html



百度知识图谱

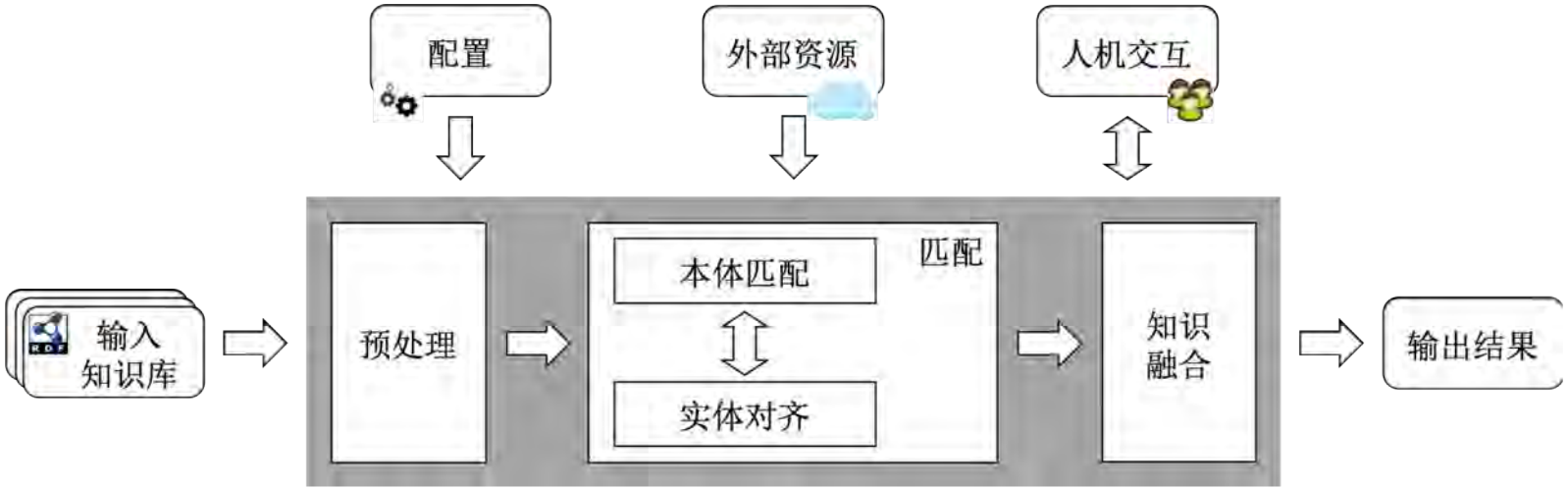
- 一个比较完整的百度语言和知识技术的布局
  - 底层基础是知识图谱，通过知识挖掘、知识整合和补全、分布式图索引及存储计算等，构建包括实体、关注点、事件、行业知识、多媒体等多元异构知识图谱
    - **多源数据知识的整合：**通过语义空间变换技术实现实体消歧、实体归一等等，解决知识表示形式多样，关联融合困难的问题

来源：https://www.chainnews.com/articles/353432152817.htm

2. 预备知识

知识图谱融合

- 目标：将不同知识图谱融合为一个统一、一致、简洁的形式，为使用不同知识图谱的应用程序之间的交互建立互操作性
- 常见流程：输入、预处理、匹配、知识融合和输出



输入

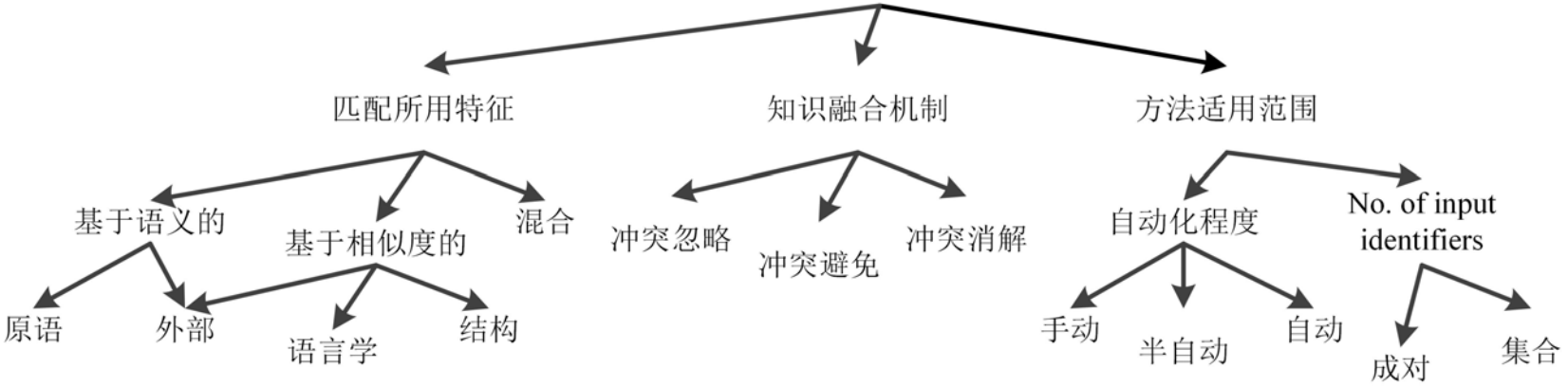
- 待集成的若干个知识图谱
  - 常见为两个，但也有一些工作支持输入更多的知识源
  - 一般为 RDF/OWL 数据文件或 SPARQL endpoint
- 配置
  - 需要预设的参数、阈值、规则
- 外部资源：知识集成过程中使用到的背景知识
  - 字/辞典 (WordNet)、常识知识 (Cyc)、实时 (Google)
  - 可能涉及人机交互，例如雇佣人来对部分数据或结果进行标注
    - 众包

预处理

- 主要包括预先对输入知识图谱进行清洗和后续步骤的准备
  - 清洗主要是为了解决输入质量问题
  - 后续步骤的准备分为配置和数据两方面
    - 配置方面
      - » 生成适合输入知识图谱的集成规则
      - » 计算出合适的模型 (超) 参数
    - 数据方面
      - » 通常使用索引技术以提高后续环节的处理速度和规模
      - » Blocking 是一项被广泛使用的技术

方法分类

- 匹配所用特征
- 知识融合机制
- 方法适用范围



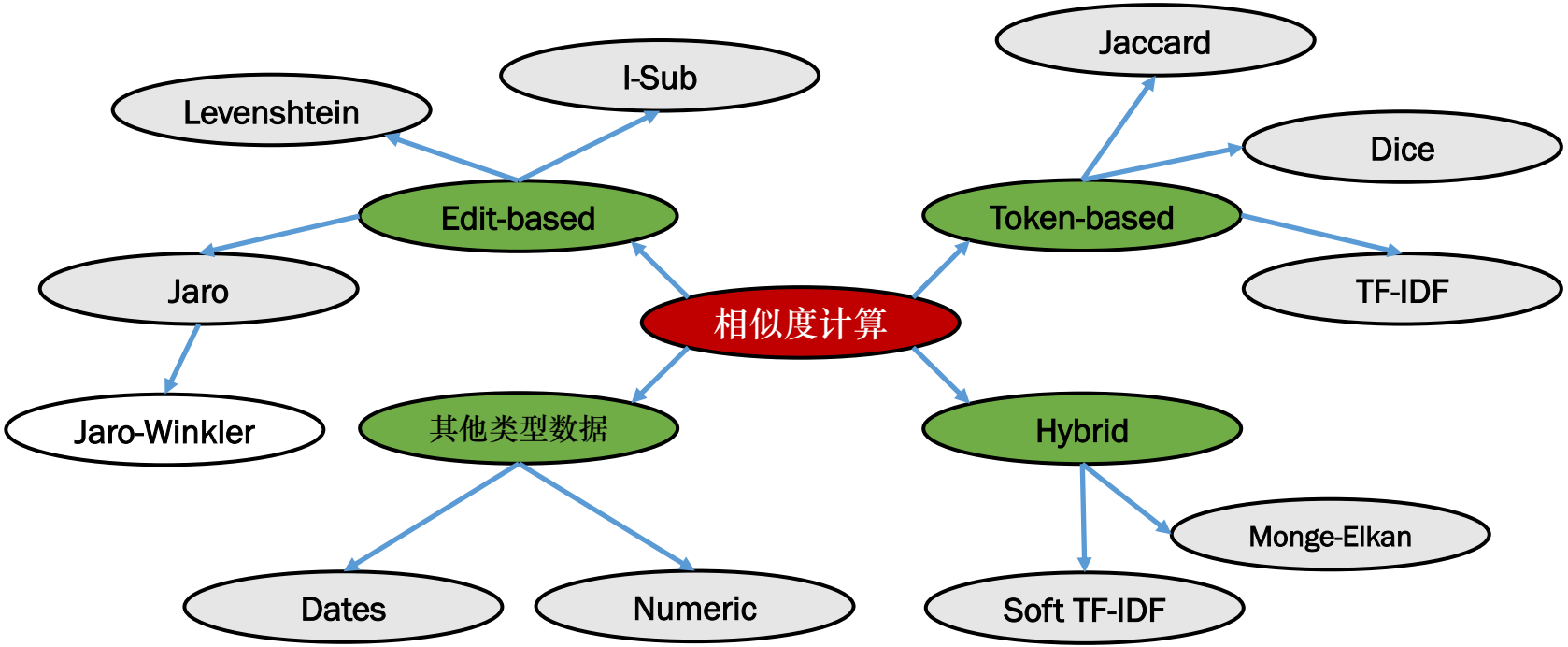
匹配 & 融合

- 本体匹配 (ontology matching)
  - 侧重发现（模式层）等价或相似的类、属性或关系
    - 本体映射 (mapping)、本体对齐 (alignment) ...
- 实体对齐 (entity alignment)
  - 侧重发现指称真实世界相同对象的不同实例
    - 实体消解 (resolution)、实例匹配 (instance matching) ...

相互作用

- 知识融合：一般通过冲突检测、真值发现等技术消解知识图谱融合过程中的冲突，再对知识进行关联与合并，最终形成一个一致的结果

相似度计算



来源：Similarity measures. 2013

Edit-based

- Levenshtein 编辑距离
  - 用最少的编辑操作将一个字符串转成另一个

$$'Lvensshtain' \xrightarrow{\text{insert } 'e'} 'Levensshtain'$$

$$'Levensshtain' \xrightarrow{\text{delete } 's'} 'Levenshtain'$$

$$'Levenshtain' \xrightarrow{\text{substitute } 'a' \to 'e'} 'Levenshtein'$$

    - 上述转换的编辑距离是 3
    - 动态规划算法

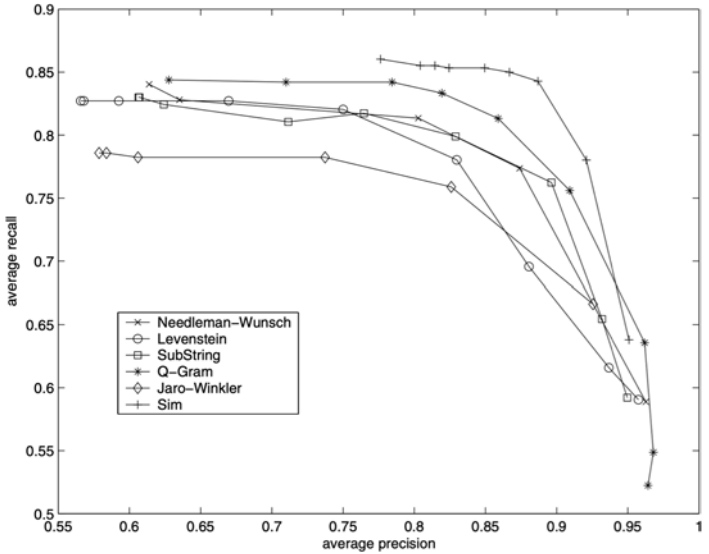
- Jaro-Winkler 编辑距离
  - Jaro 距离
    - 例如，CRATE vs. TRACE
      - $matching\ characters = 3$
      - $transportations = 0$
  - Jaro-Winkler 算法给予起始部分就相同的字符串更高的分数
    - 适合比如名字这样较短的字符串之间计算相似度
      - 不满足三角不等式

Edit-based

- I-Sub：一种改良的字符串比较算法 $\in [-1,1]$

$$Sim(s_1, s_2) = Commonality(s_1, s_2) - Difference(s_1, s_2) + Winkler(s_1, s_2)$$

  - Commonality
    - Biggest common substring between two strings
  - Difference
    - Length of unmatched strings resulted from initial matching step



来源：A string metric for ontology alignment. ISWC, 2005

Token-based

- TF-IDF
  - 预处理步骤：切词

词频 (term frequency)

一个词在某文档中的出现频率

逆文档频率 (inverse doc. freq.)

一个词的普遍重要性的度量

单词得分

$$Token\_score = TF \times IDF$$

$$TF = \frac{w}{W} \quad IDF = 1 + \log_2 \frac{N}{1+n}$$

余弦相似度

$$\cos(\overrightarrow{N_i}, \overrightarrow{N_j}) = \frac{\sum_{k=1}^{D} n_{ik} n_{jk}}{\sqrt{\sum_{k=1}^{D} n_{ik}^2 \sum_{k=1}^{D} n_{jk}^2}}$$

其他

- Hybrid: soft TF-IDF
  - Soft: Tokens are considered a partial match if they get a good score using an internal similarity measure
    - Scalability & efficiency issues
- 其他数据类型
  - Numerical comparison
    - Maximum percentage differences in absolute values
  - Time and space comparison

...

实验比较

- Precision
  - Less than two words per label: Jaro-Winkler 1, 1
  - Two or more words per label
    - Synonyms: Soft Jaccard .2, .5 with Levenstein .9 base metric
    - No synonyms: Soft Jaccard 1, 1 with Levenstein .8 base metric
- Recall
  - Less than two words per label: TF-IDF .8, .8
  - Two or more words per label
    - Synonyms: Soft TF-IDF .5, .8 with Jaro-Winkler .8 base metric
    - Different Languages: Soft TF-IDF 0, .7 with Jaro-Winkler .9 base metric
    - Other: Soft TF-IDF .8, .8 with Jaro-Winkler .8 base metric

Exact
Jaccard
Jaro-Winkler
LCS
Levenstein
Monge Elkan
N-gram
Soft Jaccard
Soft TF-IDF
I-Sub
TF-IDF

None
Tokenization
Stemming
Stopwords
Normalization
Syns/Trans

来源：String similarity metrics for ontology alignment. ISWC, 2013

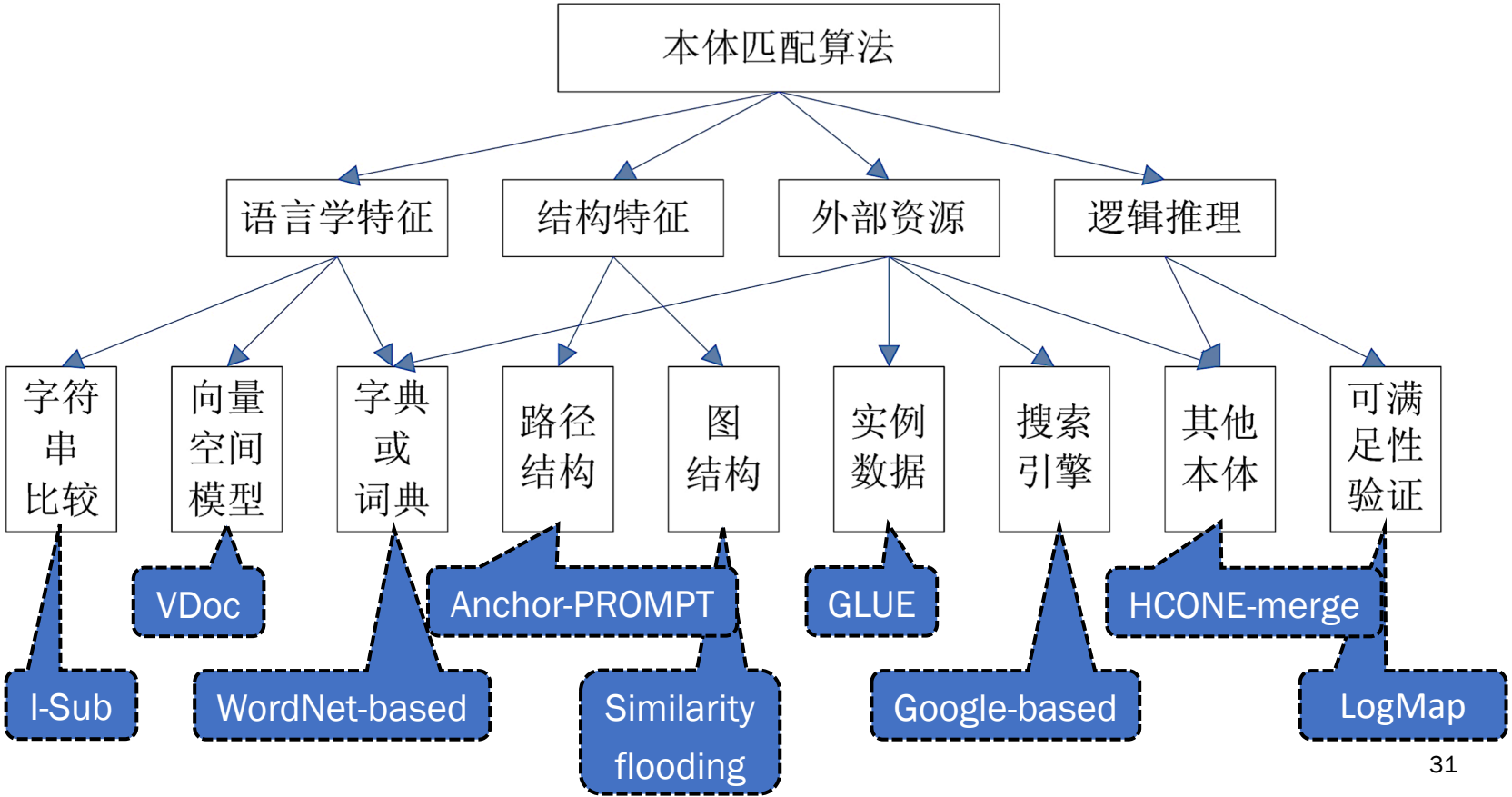
3. 本体匹配

问题定义

- 本体匹配发现一个三元组 $\mathcal{M} = \langle O, O', M \rangle$，包括一个源本体 O，一个目标本体 O'，以及一个映射单元集合 $M = \{m_1, m_2, \ldots, m_n\}$。其中 m_i 表示一个基本映射单元，可以写成 $m_i = \langle id, c, c', s \rangle$ 的四元组形式：
  - id 为映射单元的标识符，用于唯一标识该四元组
  - c 和 c' 分别为 O 和 O' 中的概念
  - s 表示 c 和 c' 之间的相似度，满足 $s \in [0, 1]$

// 另外，可有 r 表示 c, c' 之间的关系，常见的关系有等价、包含等

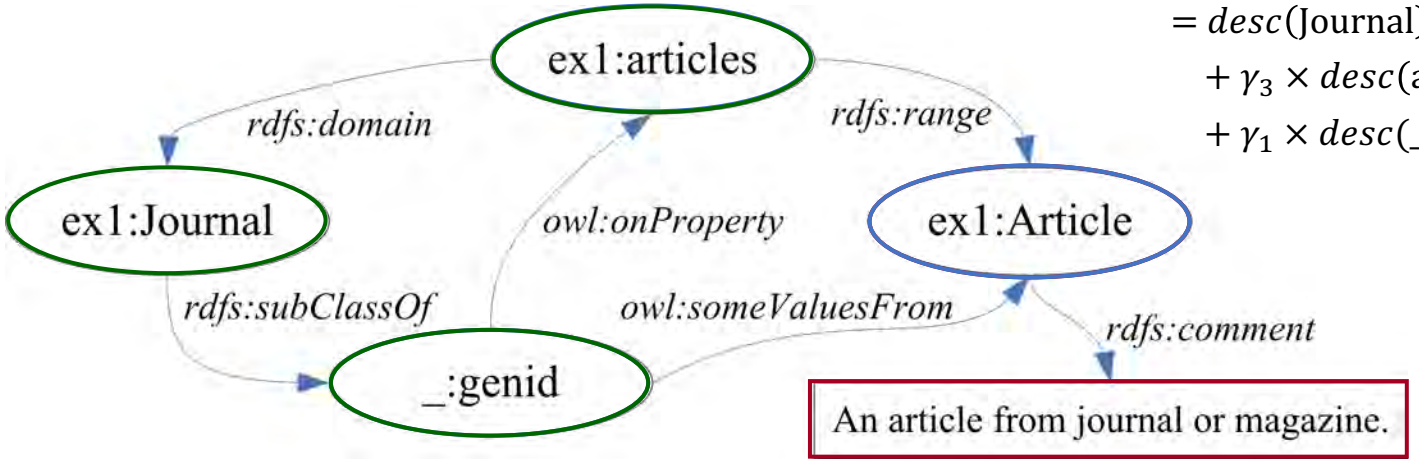
分类


本体匹配算法
语言学特征
结构特征
外部资源
逻辑推理
字符串比较
向量空间模型
字典或词典
路径结构
图结构
实例数据
搜索引擎
其他本体
可满足性验证
I-Sub
VDoc
WordNet-based
Anchor-PROMPT
Similarity flooding
GLUE
Google-based
HCONE-merge
LogMap

向量空间模型：VDoc

- 虚拟文档的构建
  - 概念的语言学描述：本地名、标签、注释
  - 匿名结点的语言学描述：前向邻居的语言学描述
  - 概念的邻居：主语邻居、谓语邻居、宾语邻居
  - 概念的虚拟文档：自身 + 邻居结点

$$vdoc(\text{Journal}) = desc(\text{Journal}) + \gamma_3 \times desc(\text{articles}) + \gamma_1 \times desc(\text{\_:genid})$$



字典或词典：WordNet-based

- 名词、动词、(形容词和副词) 各自被组织成一个同义词网络，每个同义词集合都代表一个基本的语义概念，并且由各种关系连接
- 基于 WordNet 的语义距离
  - 基于边的方法：距离越近，越相似

$$sim(C_1, C_2) = \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3}$$ distance to the least common super-concept

  - 基于信息量的统计方法：共有信息越多，它们越相似

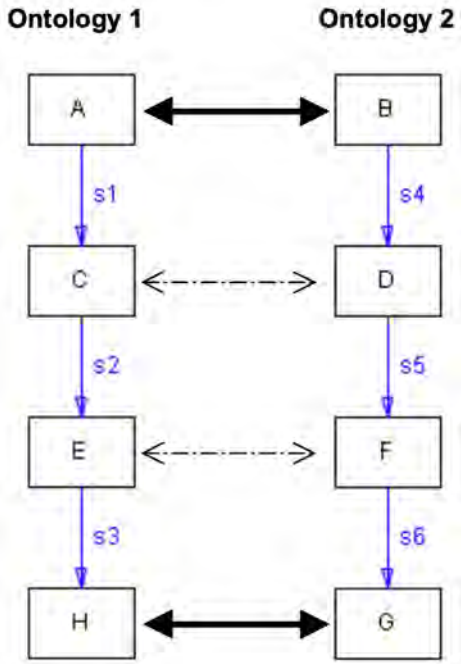
$$sim(x_1, x_2) = \frac{2 \times \log p(c_0)}{\log p(c_1) + \log p(c_2)}$$ c_0 is the most specific class that subsumes both c_1 and c_2

  - 混合方法

来源：A survey of exploiting WordNet in ontology matching. IFIP AI, 2008

路径结构：Anchor-PROMPT

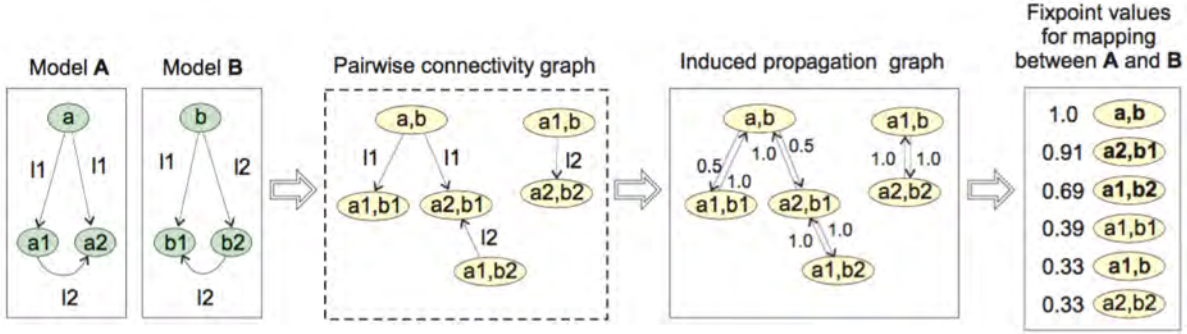
- The PROMPT plug-in allows you to compare, map, move, merge and extract multiple ontologies in Protégé
- Anchor-PROMPT
  - 如果两对术语相似且有连接它们的路径，那么路径中的元素也通常相似
    1. 生成一组长度小于 L 的路径来连接两个本体中的 anchor
    2. 生成一组长度相等的路径对
    3. 对于路径对中相同位置的节点，增加它们的相似度得分



来源：Anchor-PROMPT: Using non-local context for semantic matching. IJCAI, 2001

图结构：Similarity Flooding

- 基本思想：terms of two distinct ontologies are similar when their adjacent terms are similar



$$\sigma^{i+1}(x,y) = \sigma^{i}(x,y) + \sum_{(a_u,p,x)\in A,(b_u,p,y)\in B} \sigma^{i}(a_u,b_u)\cdot w((a_u,b_u),(x,y))$$

$$+ \sum_{(x,p,a_v)\in A,(y,p,b_v)\in B} \sigma^{i}(a_v,b_v)\cdot w((a_v,b_v),(x,y))$$

来源：Similarity flooding: a versatile graph matching algorithm and its application to schema matching. ICDE, 2002

搜索引擎：Google-based

- Approximate mappings between concepts
- Google-based similarity measure

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

where

- $f(x)$ is the number of Google hits for the search term x
- $f(y)$ is the number of Google hits for the search term y
- $f(x,y)$ is the number of Google hits for the tuple of search terms $x\ y$
- M is the number of web pages indexed by Google ($M \approx 10^{10}$)

来源：Using Google distance to weight approximate ontology matches. WWW, 2007

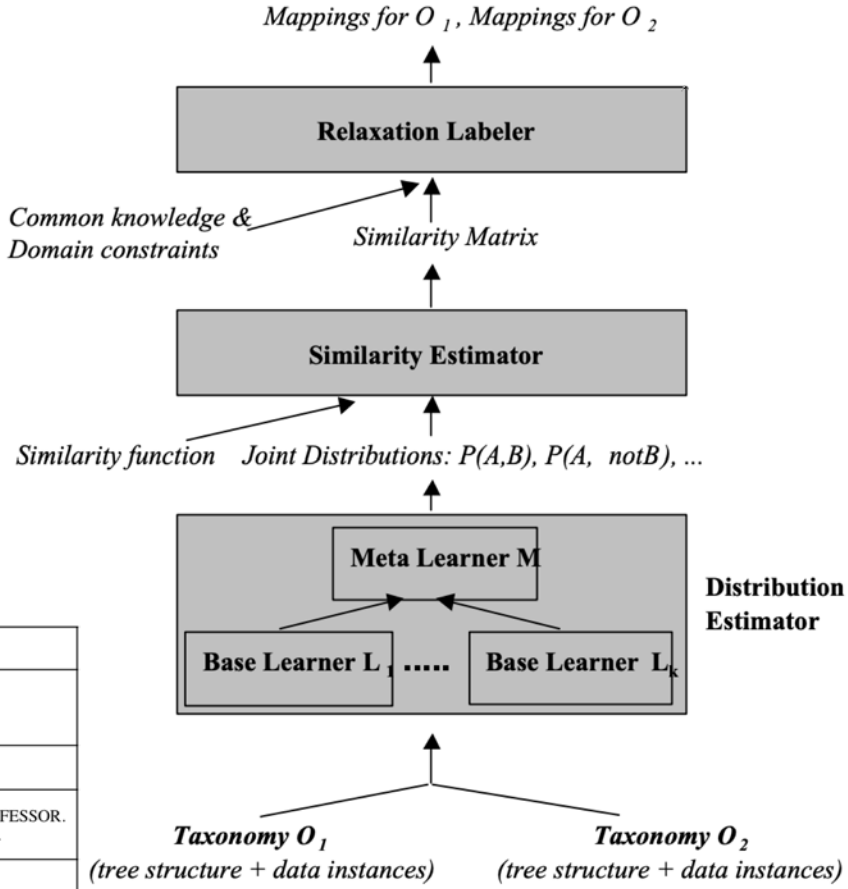
来源：Learning to map between ontologies on the semantic web. WWW, 2002

实例数据：GLUE

- 基于实例数据的机器学习
  - 联合概率分布
  - Learners
    - Content learner
    - Name learner
    - Meta learner

  } Naïve Bayes

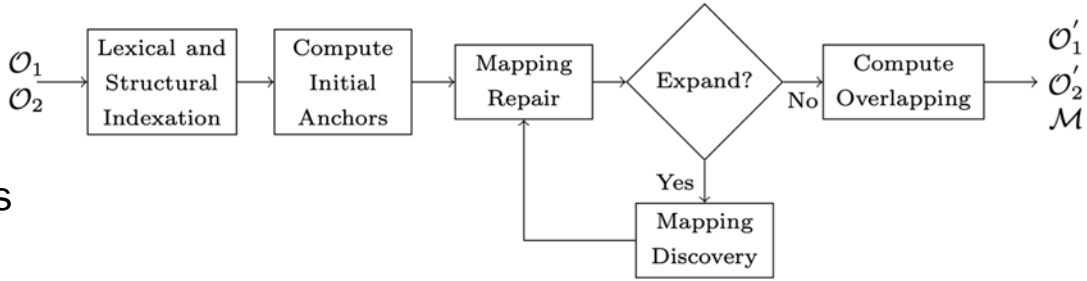
  - Relaxation labeling



Constraint Types		Examples
Domain-Independent	*Neighborhood*	Two nodes match if their children also match. Two nodes match if their parents match and at least x% of their children also match. Two nodes match if their parents match and some of their descendants also match.
	Union	If all children of node X match node Y, then X also matches Y.
Domain-Dependent	*Subsumption*	If node Y is a descendant of node X, and Y matches PROFESSOR, then it is unlikely that X matches ASSISTANT-PROFESSOR. If node Y is NOT a descendant of node X, and Y matches PROFESSOR, then it is unlikely that X matches FACULTY.
	Frequency	There can be at most one node that matches DEPARTMENT-CHAIR.
	Nearby	If a node in the neighborhood of node X matches ASSOCIATE-PROFESSOR, then the chance that X matches PROFESSOR is increased.

可满足性验证：LogMap

- 优化的数据结构用于词法索引和结构索引
  - 用来计算 anchor mappings
- 迭代过程
  - Start from initial anchors

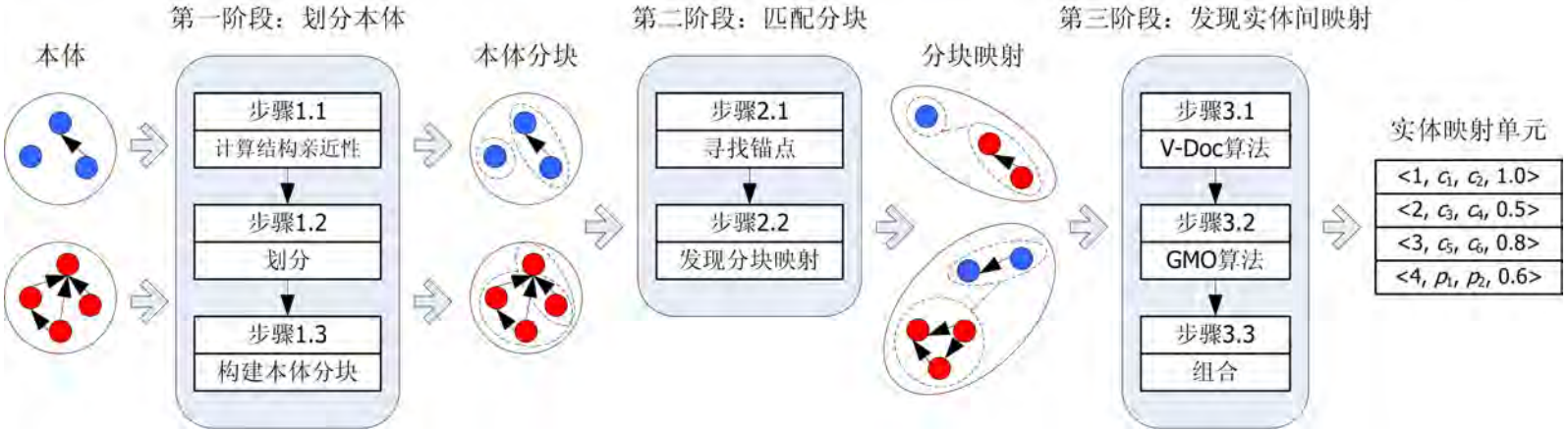


1. Mapping repair
   - Satisfiability checking w.r.t. (the merge of) both ontologies and the mappings got so far
     - » A sound and scalable (but incomplete) ontology reasoner
2. Mapping discovery
   - Use the ontologies' extended class hierarchy

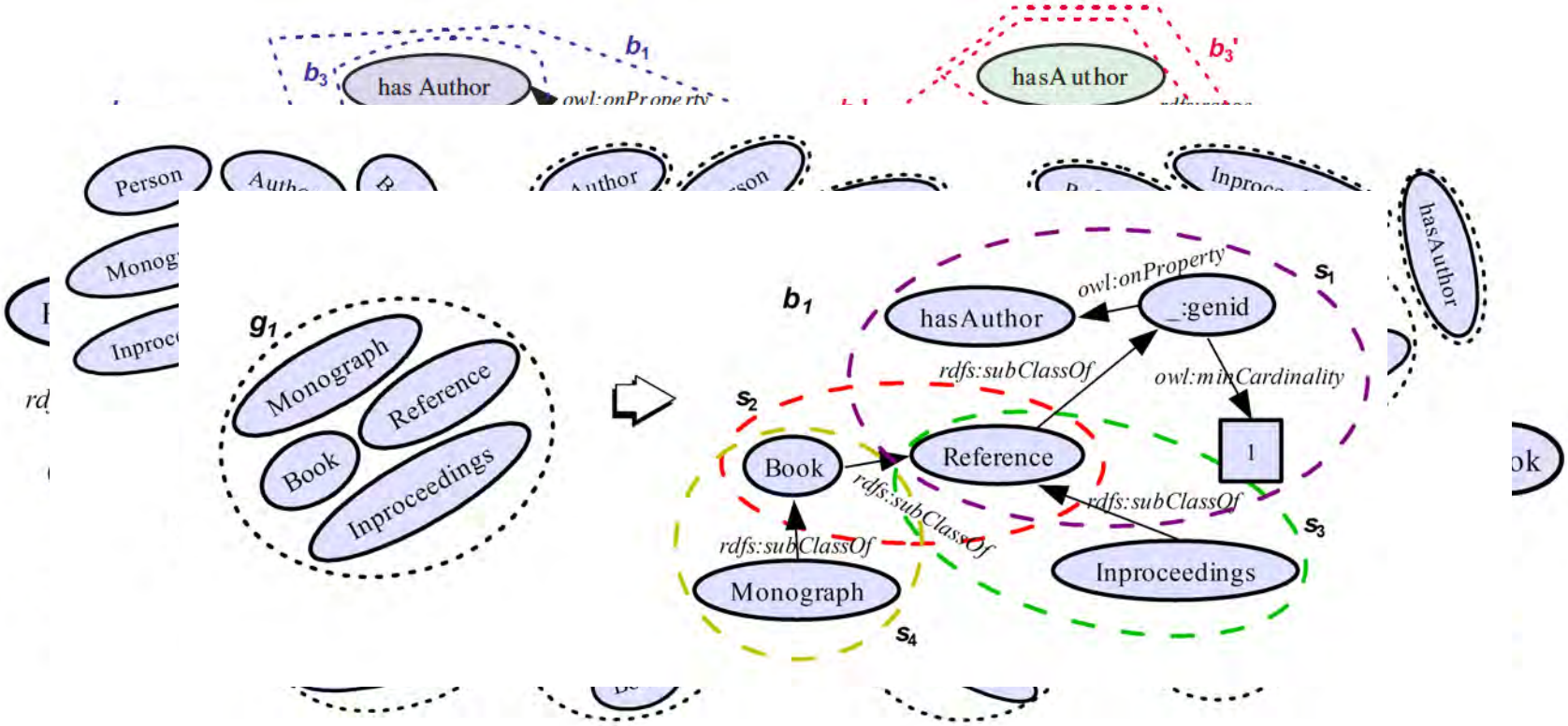
来源：LogMap: Logic-based and scalable ontology matching. ISWC, 2011

大型本体匹配

- 一方面，大多数本体匹配方法或工具仅适用于小型本体
  - 两两比较: $O(n^2)$ 复杂度
- 另一方面，许多应用需要匹配大型本体
  - 图书分类目录、生命科学本体 …



来源：Matching large ontologies: A divide-and-conquer approach. DKE, 2008

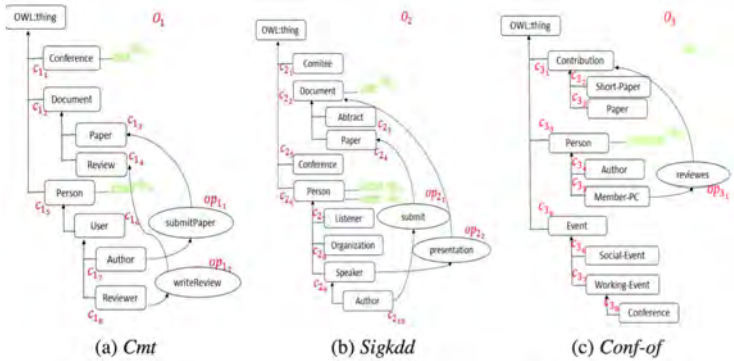


b_3	t_8: < hasAuthor, *rdfs:domain*, Reference > t_9: < hasAuthor, *rdfs:range* , Author >

b_3'	t_6': < hasAuthor, *rdfs:domain*, Entry > t_7': < hasAuthor, *rdfs:range* , Author >

全体 (holistic) 本体匹配

- 现有大多数方法处理的是成对的本体
  - 在同时匹配多个本体时会产生冲突



(a) *Cmt* (b) *Sigkdd* (c) *Conf-of*

- 建模成基于最大权图匹配的线性规划问题
  - 增加了 4 种针对本体匹配问题的一般性约束

$$\max \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sum_{k=1}^{nbC_i}\sum_{l=1}^{nbC_j} sim_{i_k,j_l}\, x_{i_k,j_l} + \sum_{m=1}^{nbOP_i}\sum_{n=1}^{nbOP_j} sim_{i_m,j_n}\, y_{i_m,j_n} + \sum_{q=1}^{nbDP_i}\sum_{r=1}^{nbDP_j} sim_{i_q,j_r}\, z_{i_q,j_r}$$

$$s.t. \sum_{l=1}^{nbC_j} x_{i_k,j_l} \le 1,\ \forall k \in [1,nbC_i] \quad (C1\ Classes)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$

$$\sum_{n=1}^{nbOP_j} y_{i_m,j_n} \le 1,\ \forall m \in [1,nbOP_i] \quad (C1\ Object\ Properties)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$

$$\sum_{r=1}^{nbDP_j} z_{i_q,j_r} \le 1,\ \forall q \in [1,nbDP_i] \quad (C1\ Data\ Properties)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$

$$x_{i_k,j_l} + x_{i_{k'},j_l} \le 1 \quad (C2\ Classes)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$
$$\forall k,k' \in [1,nbC_i],\ \forall l \in [1,nbC_j]$$

$$y_{i_m,j_n} + x_{i_{m'},j_n} \le 1 \quad (C2\ Object\ Properties)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$
$$\forall m,m' \in [1,nbDP_i],\ \forall n \in [1,nbDP_j]$$

$$z_{i_q,j_r} + x_{i_{q'},j_r} \le 1 \quad (C2\ Data\ Properties)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$
$$\forall q,q' \in [1,nbDP_i],\ \forall r \in [1,nbDP_j]$$

$$y_{i_m,j_n} \le x_{i_{k'},j_{l'}} + x_{i_{k''},j_{l''}} \quad (C3)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$
$$\forall m \in [1,nbOP_i],\ \forall n \in [1,nbOP_j]$$
$$\forall k',k'' \in [1,nbC_i],\ \forall l',l'' \in [1,nbC_j]$$

$$z_{i_q,j_r} \le x_{i_{k'},j_{l'}} \quad (C4)$$
$$\forall i \in [1,N-1],\ j \in [i+1,N]$$
$$\forall q \in [1,nbDP_i],\ \forall r \in [1,nbDP_j]$$
$$\forall k' \in [1,nbC_i],\ \forall l' \in [1,nbC_j]$$

$$x_{i_k,j_l} \in \{0,1\}\ \forall i \in [1,N-1],\ j \in [i+1,N]$$
$$\forall k \in [1,nbC_i],\ \forall l \in [1,nbC_j]$$

$$y_{i_m,j_n} \in \{0,1\}\ \forall i \in [1,N-1],\ j \in [i+1,N]$$
$$\forall m \in [1,nbOP_i],\ \forall n \in [1,nbOP_j]$$

$$z_{i_q,j_r} \in \{0,1\}\ \forall i \in [1,N-1],\ j \in [i+1,N]$$
$$\forall q \in [1,nbDP_i],\ \forall r \in [1,nbDP_j]$$

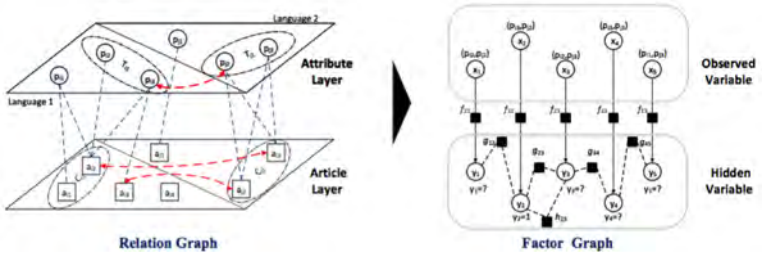
来源：An extensible linear approach for holistic ontology matching. ISWC, 2016

来源：SMap: 基于语义的关系数据库模式与 OWL 本体间映射方法. 计算机研究与发展, 2012

其他

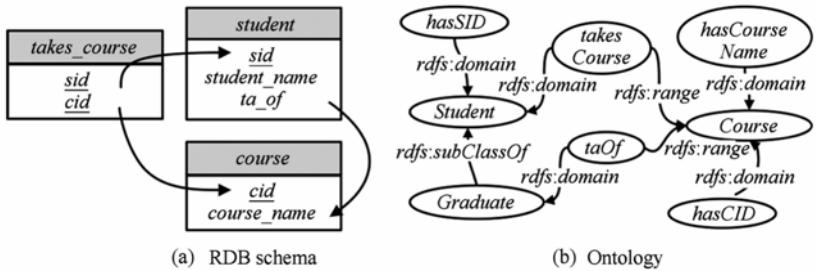
跨语言本体匹配

- EAFG
  - 用于解决跨语言属性匹配问题的因子图模型
    - 同时考虑了属性对自身的特征和属性对之间的相关性



本体与表格数据匹配

- SMap



- 复杂映射

T:student(sid, _, ta_of) [NOT NULL | ta_of]
← O:Graduate(x), O:hasSID(x,sid)

来源：Cross-lingual infobox alignment in Wikipedia using entity-attribute factor graph. ISWC, 2017

Ontology Alignment Evaluation Initiative (OAEI)

- 本体对齐竞赛，目的是评估、比较、交流及促进本体对齐工作
- 每年举办一次，结果公布在官网上
  - http://oaei.ontologymatching.org/
- 评测指标：精度、召回率、F1-score

$$\text{Precision} = \frac{\#\ correctly\_found\_matched\_pairs}{\#\ found\_matched\_pairs}$$

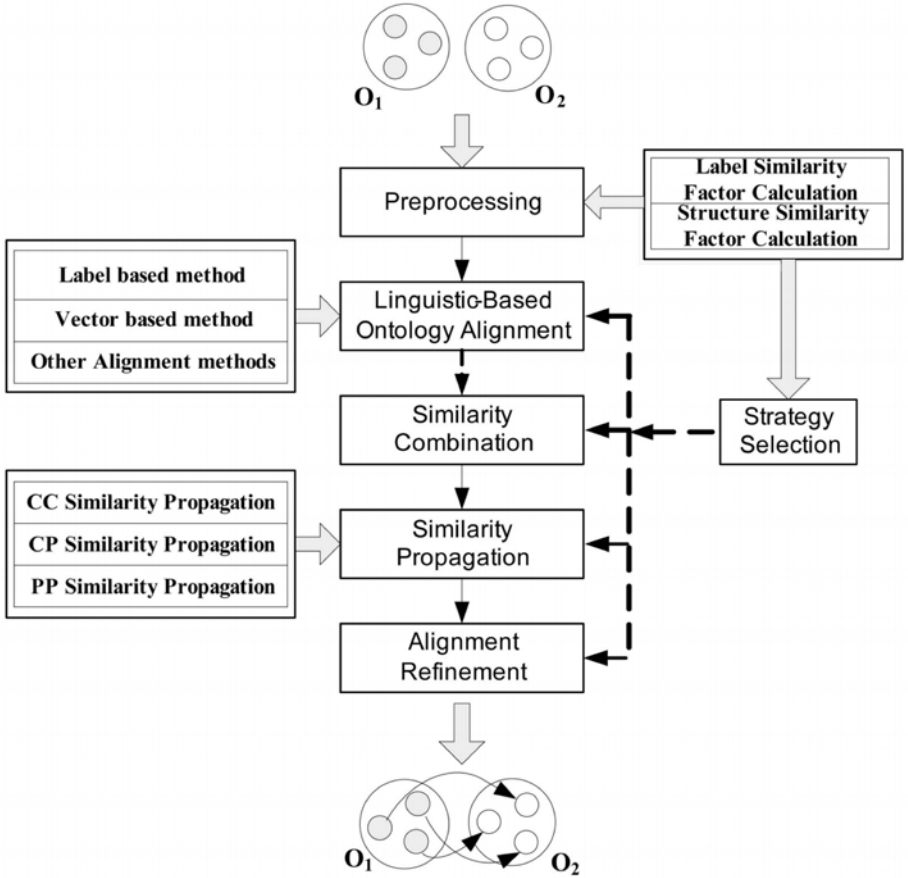
$$\text{Recall} = \frac{\#\ correctly\_found\_matched\_pairs}{\#\ existing\_matched\_pairs}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

本体匹配	关注点
anatomy	解剖学
conference	会议
Multifarm	不同语言会议数据
Complex	复杂关系
Interactive matching eval.	含交互
Large Biomedical Ontologies	大型生物本体
Disease and Phenotype	疾病及症状
Biodiversity and Ecology	环境

本体匹配系统：RiMOM

- 基于编辑距离的方法
- 基于 WordNet 的方法
- 基于 KNN 的方法
- 基于本体结构的方法
- 基于数据场和高斯函数的方法
  - 针对大规模、不平衡本体匹配问题
- 基于已有匹配结果的方法
  - 本体外部信息 (背景知识)

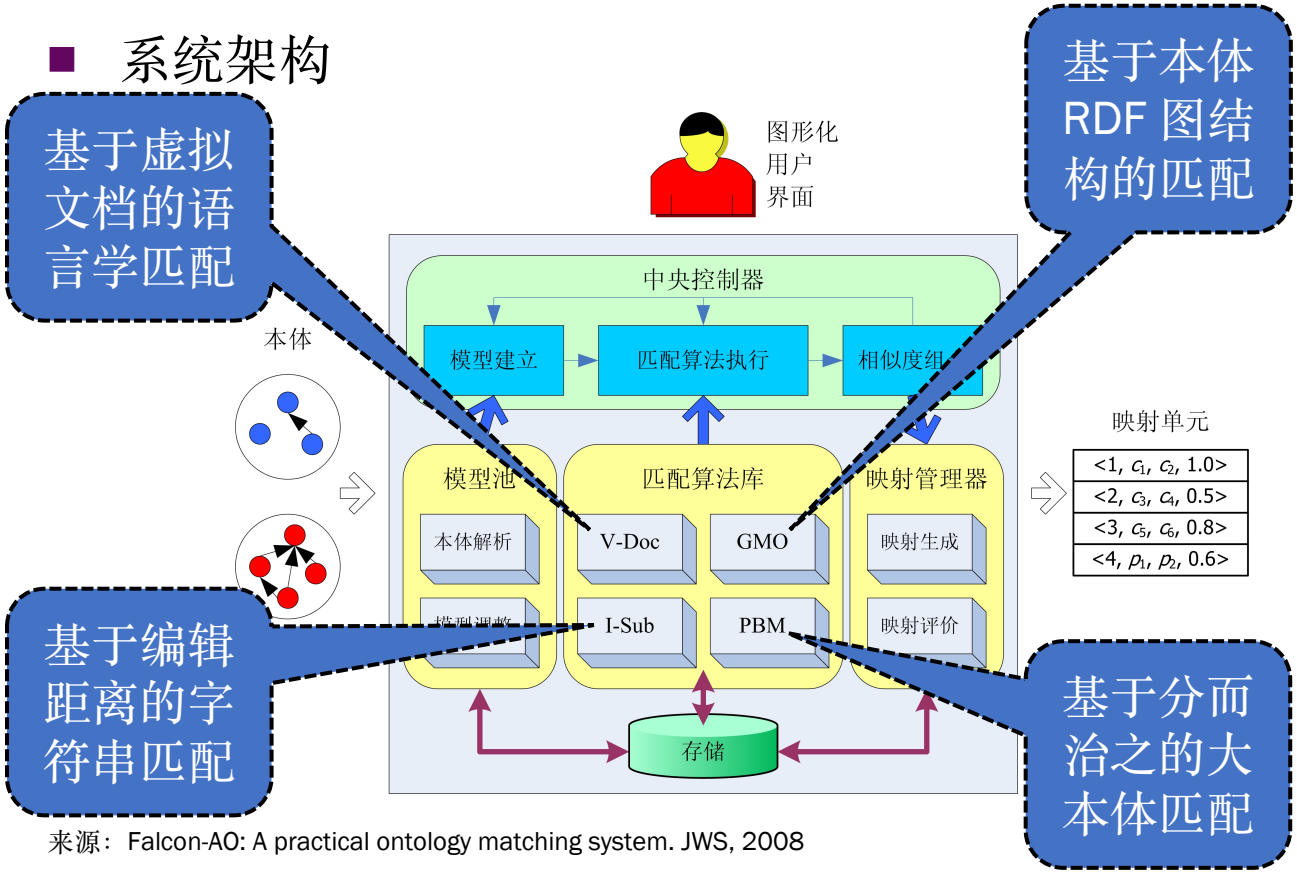


- 此外，RiMOM-IM 是一个面向实例匹配的迭代框架

来源：RiMOM: A dynamic multistrategy ontology alignment framework. TKDE, 2009

本体匹配系统：Falcon-AO

- 系统架构



来源：Falcon-AO: A practical ontology matching system. JWS, 2008

语言学可比性

- 语言学算法找到的映射单元数目对比本体概念数目

结构可比性

- 本体间使用的原语的数目对比

映射单元集成

- 语言学可比性和结构可比性分别分为高、中、低档

映射单元选取算法

- 贪心选取

4. 实体对齐

实体对齐

- 侧重发现指称真实世界相同对象的不同实例
- 问题定义与本体匹配类似
  - 规模更大、关系简单 (对齐或不对齐)
- 现有方法分类
  - 传统方法
    - 等价关系推理
      - owl:sameAs、反函数属性 ...
    - 相似度计算
      - 比较实体的属性和取值
  - 基于表示学习的方法
    - Embedding-based

成对实体对齐
集体实体对齐
大规模集体实体对齐

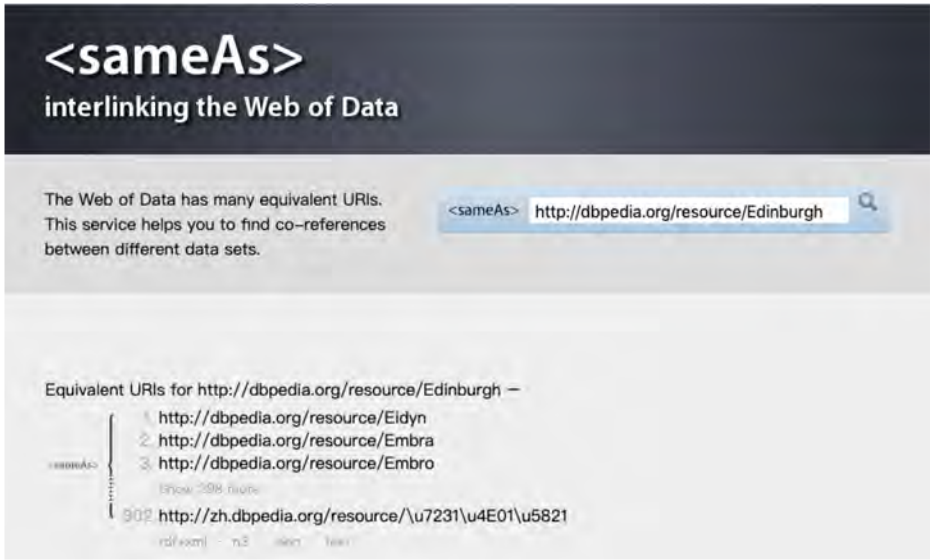
4.1 传统实体对齐方法

等价关系推理

- Same-as relation: S
  - $\langle s, \text{owl:sameAs}, o \rangle$ → $\langle s, o \rangle \in S$ and $\langle o, s \rangle \in S$
- Inverse functional property (IFP) relation: I
  - IFP: a value can only be the value of this property for a single object
    - e.g., $\langle s1, \text{foaf: mbox}, o \rangle, \langle s2, \text{foaf: mbox}, o \rangle$ → $\langle s1, s2 \rangle \in I$ and $\langle s2, s1 \rangle \in I$
- Functional property (FP) relation: F
- Cardinality relation: C
  - owl:cardinality / owl:maxCardinality = 1
- $K = (S \cup I \cup F \cup C)^+$, K is an equivalence relation

sameAs.org

- Currently serving 203,953,936 URIs which relate to over 53,054,359 apparently distinct entities

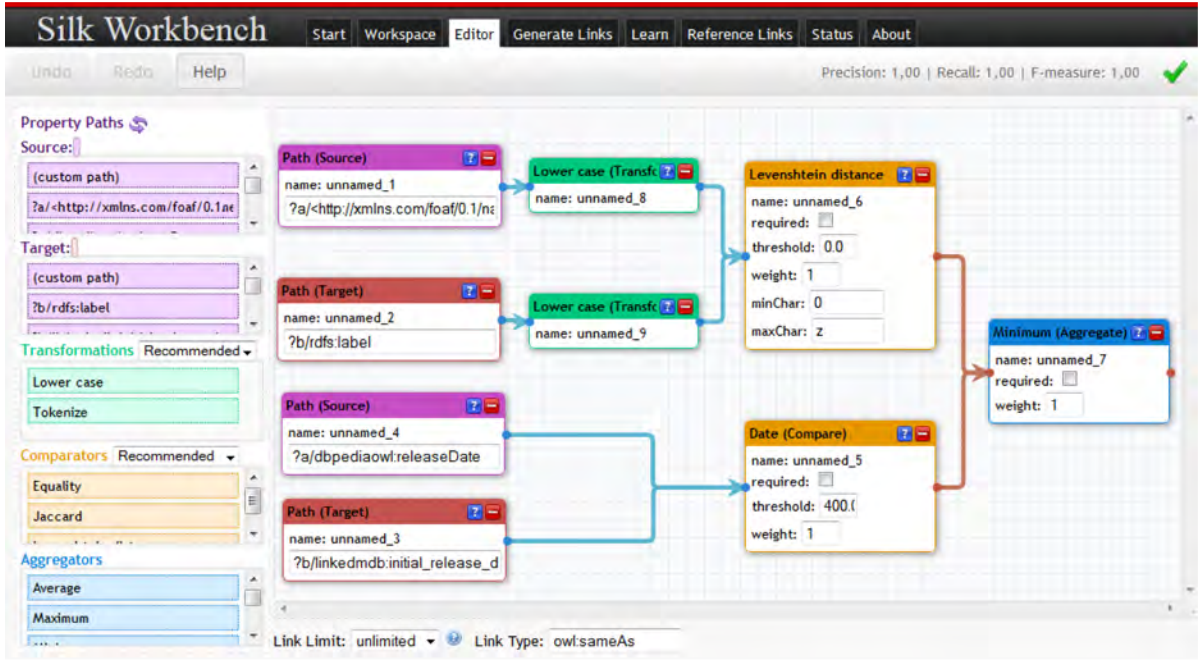


http://go.bio2rdf.org/ http://purl.org/hcls/
http://moustaki.org/ http://rdf.dmoz.org/
http://doapstore.org/ http://dbpedia.org/
http://rdf.geospecies.org/ http://www.yr-bcn.es/pmika/
http://umbel.org/ http://downloads.dbpedia.org/
http://www.opencyc.org/ http://hcls.deri.org/
http://lingvoj.org/ http://www.cs.vu.nl/STITCH/rameau/
http://rkbexplorer.com/ http://airports.dataincubator.org/
http://telegraphis.net/ http://ontologi.es/rail/stations
http://data.linkedct.org/ http://discogs.dataincubator.org/
http://www.bbc.co.uk/music/ http://linkedgeodata.org/
http://data.nytimes.com/ http://bnb.data.bl.uk
http://d-nb.info http://data.bibsys.no
http://nektar.oszk.hu http://dbpedia.org
http://id.loc.gov http://id.ndl.go.jp
http://stitch.cs.vu.nl

来源：http://sameas.org/

相似度计算：Silk

- Indexing
- Similarity metrics
  - String
  - Numeric
  - Geographic
  - Aggregation
- Transformation functions
- Link specification



来源：http://silkframework.org/

混合方法：ObjectCoref

- 问题定义：query-driven

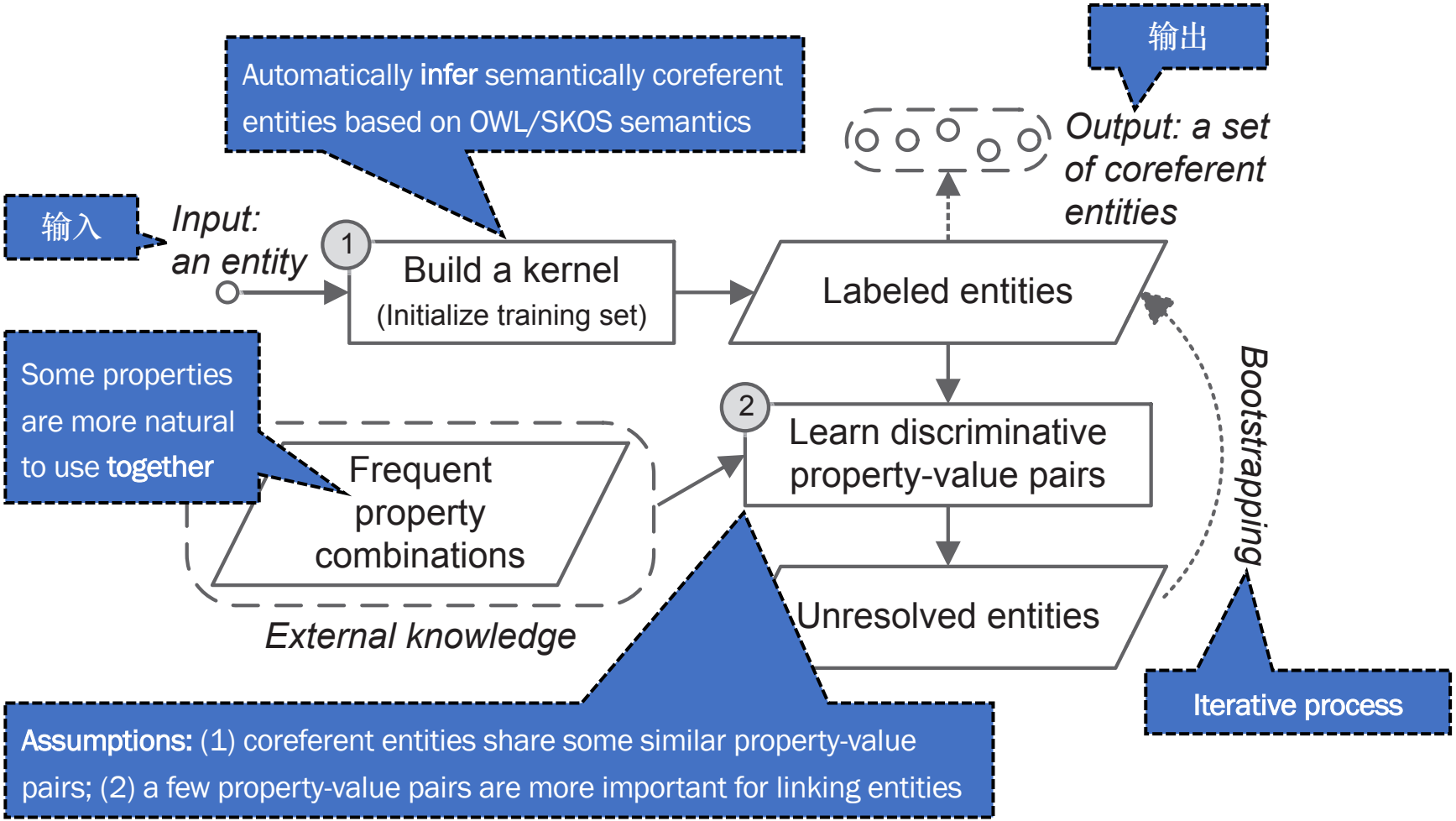
Let $\mathbf{U}$ be the set of entities in a set $\mathbf{D}$ of data sources. Given an entity $u \in \mathbf{U}$, the entity alignment for u is to query a subset $\mathbf{E}(u) \subseteq \mathbf{U}$ of entities for which a relation ε holds:

$$\mathbf{E}(u) = \{v \in \mathbf{U} \mid (u, v) \in \mathcal{E}\}$$

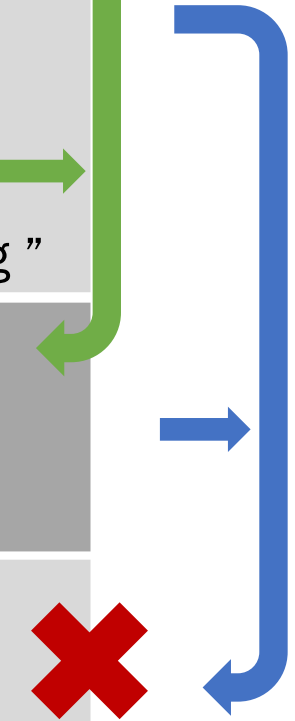
where ε links all the entities in $\mathbf{U}$ that refer to the same object as u does.

- 基本思想：bootstrapping
- 应用场景
  1. Search / browsing – a system knows “what to link” only at query time
  2. Analyze small portions of a very large dataset to answer on-demand queries

来源：A self-training approach for resolving object coreference on the semantic web. WWW, 2011


输出
Automatically infer semantically coreferent entities based on OWL/SKOS semantics
Output: a set of coreferent entities
输入
Input: an entity
1
Build a kernel
(Initialize training set)
Labeled entities
Some properties are more natural to use together
2
Learn discriminative property-value pairs
Frequent property combinations
External knowledge
Unresolved entities
Bootstrapping
Iterative process
Assumptions: (1) coreferent entities share some similar property-value pairs; (2) a few property-value pairs are more important for linking entities

dbpedia:Nanjing (DBpedia)	rdfs:label owl:sameAs	“ **Nanjing** ” *geo:1799962*
geo:1799962 (GeoNames)	geo:lat geo:long geo:alternateName	“ 32 N ” “ 118 E ” “ **Nanjing** ” “ Nan-ching ”
fb:m.05gqy (Freebase)	rdfs:label geo:lat geo:long	“ **Nanjing** ” “ 32 N ” “ 118 E ”
ex:NationalCity	geo:long geo:lat	“ 117 W ” “ 32 N ”



成对 (pairwise) 实体对齐

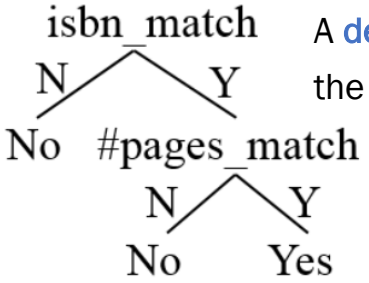
成对实体对齐

- 实体对齐 → 分类问题
  - Features for entity pair (x, y): $\vec{v} = (\text{sim}_1(x, y), \text{sim}_2(x, y), \ldots, \text{sim}_k(x, y))$
  - Label for entity pair: 0 - nonmatch, 1 - match
  - Use features and labels to learn a binary classifier

  - 例如，the features of R_1 and R_2 is $\vec{v} = (0.50, 0.67, 0.67, 0.50)$, where all similarity functions are Jaccard

Identifiers	Givennames	Surnames	Postcodes	Suburb names
R1	Peter	Christen	2010	North Sydney
R2	Pedro	Kristen	2000	Sydney
R3	Paul	Smith	2600	Canberra
R4	Pablo	Smyth	2700	Canberra Sth

成对实体对齐：Magellan



A decision tree in the random forest

- Random forest + various literal similarities
  - Use various similarity functions to generate features
  - Use random forest to learn several decision trees

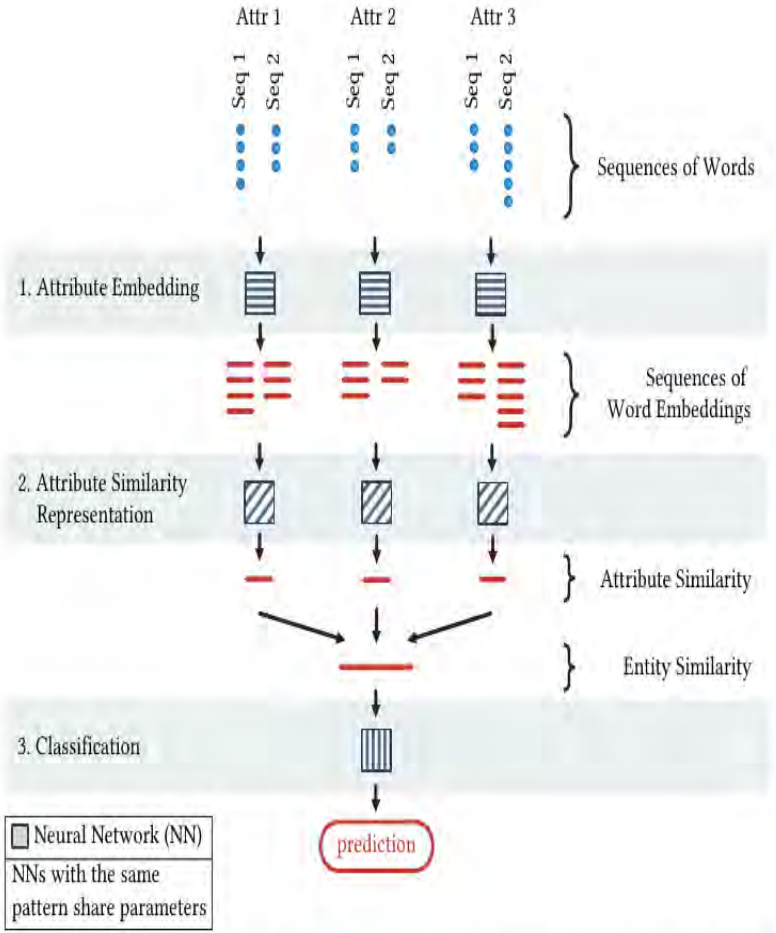
Attribute Type & Characteristic	Similarity Function	Intuition
Single word string	Exact Match, Jaccard_3gram, Overlap_3gram, Dice_3gram, Levenshtein, Jaro*, Jaro-Winkler*	first names, last names, zip codes, etc.
Multi-word short string ($\#words \leq 5$)	Jaccard_3gram, Overlap_3gram, Dice_3gram, Jaccard_word, Overlap_word, Dice_word, Cosine_word, Monge-Elkan*, Needleman-Wunsch*, Smith-Waterman*, Smith-Waterman-Gotoh*	product brand names, full names of people, etc.
Multi-word medium string ($6 \leq \#words \leq 10$)	Jaccard_word, Overlap_word, Dice_word, Cosine_word, Monge-Elkan*	street addresses, short product descriptions, etc.
Multi-word long string ($\#words \geq 11$)	Jaccard_word, Overlap_word, Dice_word, Cosine_word, TF/IDF*, Soft TF/IDF*	long product descriptions, product reviews, etc.
Numeric	Exact Match, Absolute Difference, Relative Difference, Levenshtein	age, size, weight, height, price, etc.

来源：Falcon: Scaling up hands-off crowdsourced entity matching to build cloud services. SIGMOD, 2017

成对实体对齐：DeepMatcher

- Word embedding for literal values
- Attr. summarization combines tokens to one vector
- Attr. comparison compares summarized vectors

Architecture module		Options	
Attribute embedding		Granularity: (1) Word-based (2) Character-based	Training: (3) Pre-trained (4) Learned
Attribute Similarity representation	(1) Attribute summarization	(1) Heuristic-based (2) RNN-based (3) Attention-based (4) Hybrid	
	(2) Attribute comparison	(1) Fixed distance (cosine, Euclidean) (2) Learnable distance (concatenation, element-wise absolute difference, element-wise multiplication)	
Classifier		NN (multi-layer perceptron)	



来源：Deep learning for entity matching: A design space exploration. SIGMOD, 2018

Magellan 对比 DeepMatcher

- For structured data, they achieve similar performance in current empirical evaluation
- For textual and dirty data, complex DL models offer significant accuracy improvements but often require far longer training time
- DL models requires large amounts of training data to achieve good performance

集体 (collective) 实体对齐

集体实体对齐

- 利用实体间的关系来提高精度和召回率
  - $Jaccard_{\text{Name}}(u_1, u_3) = 0.36$
  - $Jaccard_{\text{Title}}(p_1, p_3) = 1 \Rightarrow Jaccard_{\text{Papers}}(u_1, u_3) = 1$

    $\Rightarrow u_1, u_3$ are likely to be a match

Block	P_Id	Title	Abstract	Keywords	Authors	Venue
P_1	p_1	Transaction Support in Read Optimized and ...	...	{File System, Transactions}	$\{a_1, a_2\}$	u_1
	p_3	Transaction Support in Read Optimized and ...	...	{File System, Transactions}	$\{a_3, a_4\}$	u_3
P_2	p_2	Read Optimized File System Designs: A performance ...	...	{File System, Database}	$\{a_1\}$	u_2
	p_4	Berkeley DB: A Retrospective	...	{File System, Database}	$\{a_3\}$	u_4

(a) Entity-set Papers.

Block	A_Id	Name	Email	Papers
A_1	a_1	Margo Seltzer	margo@harvard.edu	$\{p_1, p_2\}$
	a_3	Margo I. Seltzer	seltzer@gmail.com	$\{p_3, p_4\}$
A_2	a_2	Michael Stonebraker	stonebraker@mit.edu	$\{p_1\}$
	a_4	M. Stonebraker	stonebraker@ucb.edu	$\{p_3\}$

(b) Entity-set Authors.

Block	V_Id	Name	Papers
U_1	u_1	Very Large Data Bases	$\{p_1\}$
	u_3	VLDB	$\{p_3\}$
U_2	u_2	ICDE Conference	$\{p_2\}$
	u_4	IEEE Data Eng. Bull	$\{p_4\}$

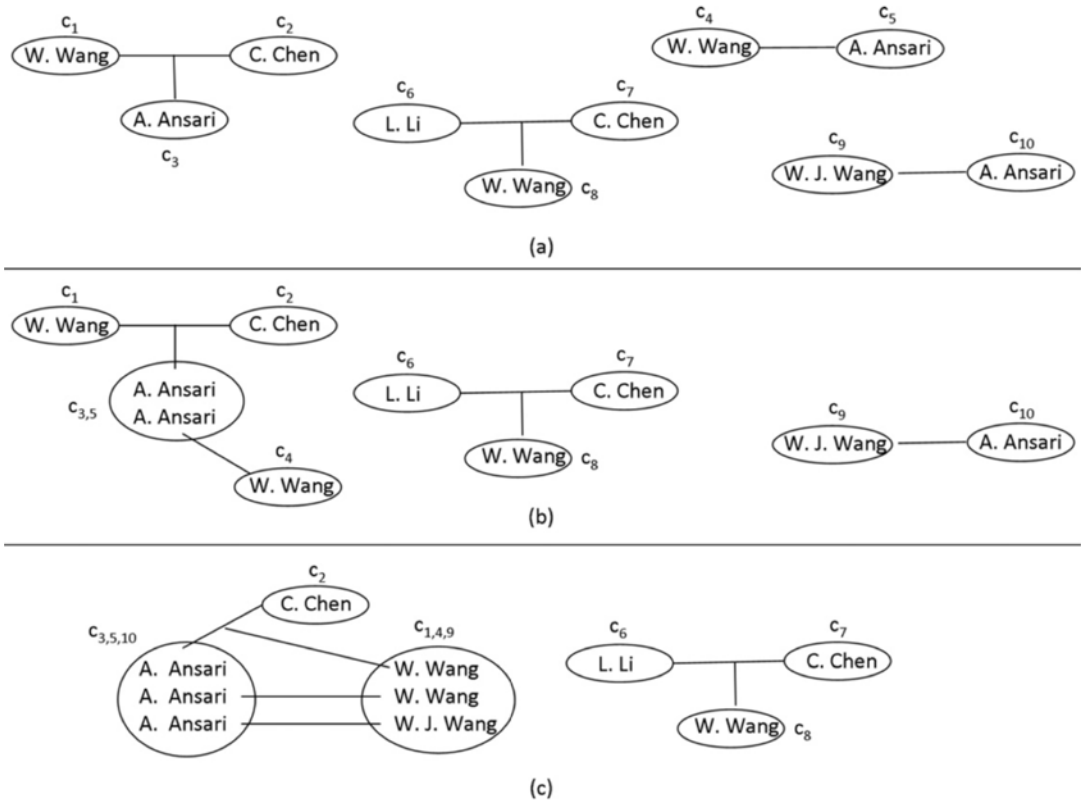
(c) Entity-set Venues.

Treat related references as additional attributes for matching

来源：Progressive approach to relational entity resolution. VLDB, 2014

基于聚类的集体实体对齐

- 每条边表示 “coauthor” 关系
- 每次迭代计算聚类相似度并将两个最相似的聚类合并
  - 邻居相似度度量
    - 公共邻居
    - Jaccard coefficient
    - Adar 相似度



来源：Data matching. Principles of data integration, 2012

基于马尔科夫逻辑网络的集体实体对齐

- Markov Logic Network (MLN)
  - 定义：a set of pairs (F_i, w_i), where F_i is a formula in first-order logic and w_i is a real number
    - 例如，$\forall x, y : x = y \wedge y = x$ has a weight of 1
      - » $x = y$ denotes that (x, y) is a match
- Ground truth from KG
  - Known triples: $(h, r, t) \in KG \Rightarrow R(h, t)$
  - Training data: (e_1, e_2) is a match $\Rightarrow e_1 = e_2$; (e_1, e_2) is a non-match$\Rightarrow$ $e_1 \neq e_2$

来源：Entity resolution with Markov logic. ICDM, 2006

基于马尔科夫逻辑网络的集体实体对齐

Literal comparison

$$\forall x_1, x_2, y_1, y_2\ HasWord(x_1, y_1) \wedge HasWord(x_2, y_2) \wedge y_1 = y_2 \Rightarrow x_1 = x_2$$

$$\forall x_1, x_2, y_1, y_2\ \neg HasWord(x_1, y_1) \wedge HasWord(x_2, y_2) \wedge y_1 = y_2 \Rightarrow x_1 \neq x_2$$

$$\forall x_1, x_2, y_1, y_2\ HasWord(x_1, y_1) \wedge \neg HasWord(x_2, y_2) \wedge y_1 = y_2 \Rightarrow x_1 \neq x_2$$

$$\forall x_1, x_2, y_1, y_2\ \neg HasWord(x_1, y_1) \wedge \neg HasWord(x_2, y_2) \wedge y_1 = y_2 \Rightarrow x_1 = x_2$$

Attribute comparison

$$\forall x_1, x_2, y_1, y_2\ HasWord(x_1, y_1) \wedge HasWord(x_2, y_2) \wedge y_1 = y_2 \wedge R(z_1, x_1) \wedge R(z_2, x_2) \Rightarrow z_1 = z_2$$

$$\forall x_1, x_2, y_1, y_2\ \neg HasWord(x_1, y_1) \wedge HasWord(x_2, y_2) \wedge y_1 = y_2 \wedge R(z_1, x_1) \wedge R(z_2, x_2) \Rightarrow z_1 \neq z_2$$

$$\forall x_1, x_2, y_1, y_2\ HasWord(x_1, y_1) \wedge \neg HasWord(x_2, y_2) \wedge y_1 = y_2 \wedge R(z_1, x_1) \wedge R(z_2, x_2) \Rightarrow z_1 \neq z_2$$

$$\forall x_1, x_2, y_1, y_2 \neg HasWord(x_1, y_1) \wedge \neg HasWord(x_2, y_2) \wedge y_1 = y_2 \wedge R(z_1, x_1) \wedge R(z_2, x_2) \Rightarrow z_1 = z_2$$

基于马尔科夫逻辑网络的集体实体对齐

- Relation comparison

$$\forall x_1, y_1, x_2, y_2\ R(x_1, y_1) \wedge R(x_2, y_2) \wedge (y_1 = y_2) \Rightarrow (x_1, x_2)$$

  - 例如

$$Author(bc_1, a_1) \wedge Author(bc_2, a_2) \wedge SameAuthor(a_1, a_2) \Rightarrow SameBib(bc_1, bc_2)$$

$$Author(bc_1, a_1) \wedge Author(bc_2, a_2) \wedge SameBib(bc_1, bc_2) \Rightarrow SameAuthor(a_1, a_2)$$

- 其他规则
  - 创建关系

$$\forall x, y_1, y_2\ HasAuthor(x, y_1) \wedge HasAuthor(x, y_2) \Rightarrow Coauthor(y_1, y_2)$$

MLN 实体对齐示例：http://alchemy.cs.washington.edu/mlns/er/

大规模集体实体对齐

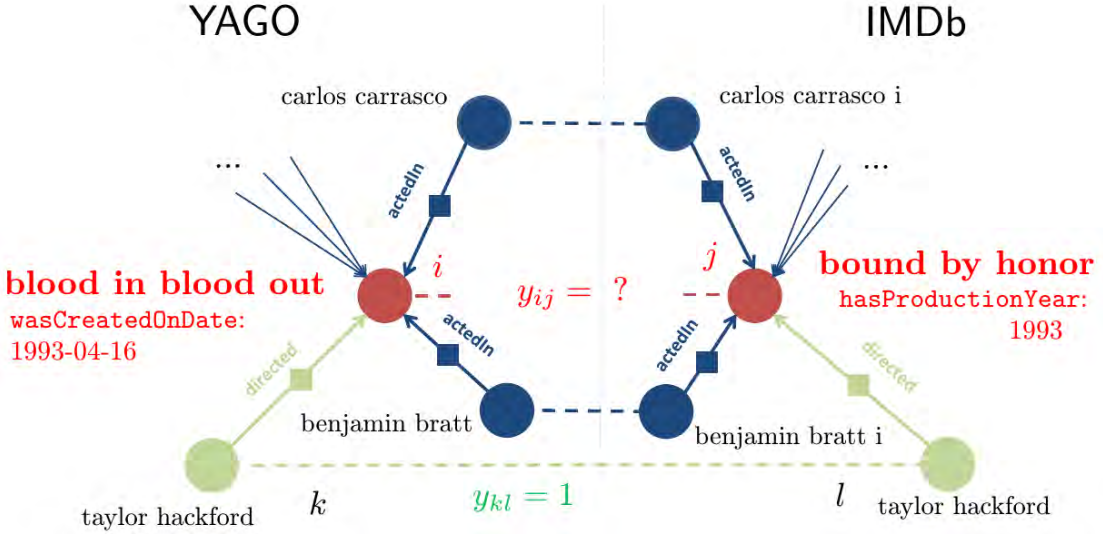
大规模集体实体对齐：SiGMa

- 依赖于 1:1 假设的贪心算法
  - 一旦识别出某个 match，就不再需要将其与其他实体进行比较
- 利用关系图为决策打分并提出候选
- 可使用定制的评分函数处理领域知识，并实现了一个简单的迭代算法
- 在精度和召回率之间以及计算量和召回率之间提供自然的折衷
  - Simplicity & greediness → high time efficiency
  - High effectiveness, as well

来源：SiGMa: Simple greedy matching for aligning large knowledge bases. KDD, 2013

大规模集体实体对齐：SiGMa

- SiGMa uses neighbors to
  - Score candidates
  - Suggest candidates (iterative blocking)



大规模集体实体对齐：SiGMa

- Pair scoring based on aligned pairs

$$score(i,j;y) = (1-\alpha)\ s_{ij} + \alpha\ \delta g_{ij}(y)$$

where $\delta g_{ij}(y) \doteq \sum_{(k,l)\in\mathcal{N}_{ij}} y_{kl}(w_{ij,kl} + w_{kl,ij})$

  - y denotes aligned entity pairs
  - s_{ij} denotes the static similarity of pair (i,j)

$$s_{ij} = (1-\beta)\ string(i,j) + \beta\ prop(i,j)$$

  - g_{ij} denotes the dynamic structural similarity of pair (i,j)
    g_{ij} increases when SiGMa finds more matches among neighbors of pair (i,j)

$$g_{ij}(y) = \sum_{(k,l)\in\mathcal{N}_{ij}} y_{kl}(\gamma_i w_{ik} + r_j w_{jl})$$

大规模集体实体对齐：SiGMa

- Dynamic structural similarity

$$g_{ij}(y) = \sum_{(k,l)\in\mathcal{N}_{ij}} y_{kl}(\gamma_i w_{ik} + r_j w_{jl})$$

$$\gamma_i = \frac{1}{2}\left(1 + \sum_{k\in\mathcal{N}_i} w_{ik}\right)^{-1} \quad \gamma_j = \frac{1}{2}\left(1 + \sum_{l\in\mathcal{N}_j} w_{jl}\right)^{-1}$$

  - By this definition, the dynamic structural similarity equals to the weighted Jaccard similarity of neighbor sets

大规模集体实体对齐：PARIS

- PARIS
  - 不仅适用于实例，还适用于关系和类的整体算法
  - 不需要人工输入，也不需要训练数据
  - 不需要调参
  - 需要每个 KG 中不存在冗余实体
- 方法概述
  - 实体对齐逻辑规则 → 概率模型
  - 寻找整体概率模型的不动点
    - Iterate the estimations for relations and entities until convergence (no proof)
    - Compute the class matching

来源：PARIS: Probabilistic alignment of relations, instances, and schema. VLDB, 2012

大规模集体实体对齐：PARIS

- Probabilistic modeling
  - Assume mutual independence of all distinct elements in the model

$$\Pr(A \wedge B) = \Pr(A) \times \Pr(B)$$

$$\Pr(A \vee B) = 1 - \big(1 - \Pr(A)\big)(1 - \Pr(B))$$

$$\Pr\big(\forall x : \varphi(x)\big) = \prod_x \Pr(\varphi(x))$$

- Inverse functionality of relation
  - If a property is declared to be inverse-functional, then the object of a property statement uniquely determines the subject (some individual)

$$fun(r) = \frac{\#x{:}\exists y{:}r(x,y)}{\#x,y,r{:}r(x,y)} \quad fun^{-1}(r) = fun(r^{-1}) = \frac{\#x{:}\exists y : r(x,y)}{\#x,y : r(x,y)}$$

大规模集体实体对齐：PARIS

- Entity matching by (inverse) functionality
  - Rule 1

$$\exists r, y, y' : r(x, y) \wedge r(x', y') \wedge y \equiv y' \wedge fun^{-1}(r) \text{ is high} \Rightarrow x \equiv x'$$

  - Rule 1 to probability

$$Pr_1(x \equiv x') \coloneqq 1 - \prod_{\substack{r(x,y) \\ r(x',y')}} (1 - fun^{-1}(r) \times \Pr(y \equiv y'))$$

  - Generate rule 1 with sub-relations r'

$$1 - \prod_{\substack{r(x,y) \\ r(x',y')}} (1 - fun^{-1}(r) \times \Pr(y \equiv y')) \times (1 - \Pr(r \subseteq r') \times fun^{-1}(r) \times \Pr(y \equiv y'))$$

大规模集体实体对齐：PARIS

- Entity matching by (inverse) functionality
  - Rule 2

$$\exists r, y, y' : r(x,y) \wedge (\forall y' : r(x',y') \Rightarrow y \not\equiv y') \wedge fun(r) \text{ is high} \Rightarrow x \not\equiv x'$$

  - Combine rule 1 and rule 2

$$\left(1 - \prod_{\substack{r(x,y) \\ r(x',y')}} (1 - \Pr(r' \subseteq r) \times fun^{-1}(r) \times \Pr(y \equiv y')) \times (1 - \Pr(r \subseteq r') \times fun^{-1}(r') \times \Pr(y \equiv y'))\right)$$

$$\times \prod_{\substack{r(x,y) \\ r'}} \left(1 - fun(r) \times \Pr(r' \subseteq r) \times \prod_{r'(x',y')} (1 - \Pr(x \equiv x'))\right) \times \left(1 - fun(r') \times \Pr(r \subseteq r') \times \prod_{r'(x',y')} (1 - \Pr(x \equiv x'))\right)$$

大规模集体实体对齐：PARIS

- 跨 KG 子关系

$$\Pr(r \subseteq r') := \frac{\#x,y : r(x,y) \wedge r'(x,y)}{\#x,y : r(x,y)}$$

  - 考虑匹配后，分子变为

$$\#x,y : r(x,y) \wedge (\exists x', y' : x \equiv x' \wedge y \equiv y' \wedge r'(x',y'))$$

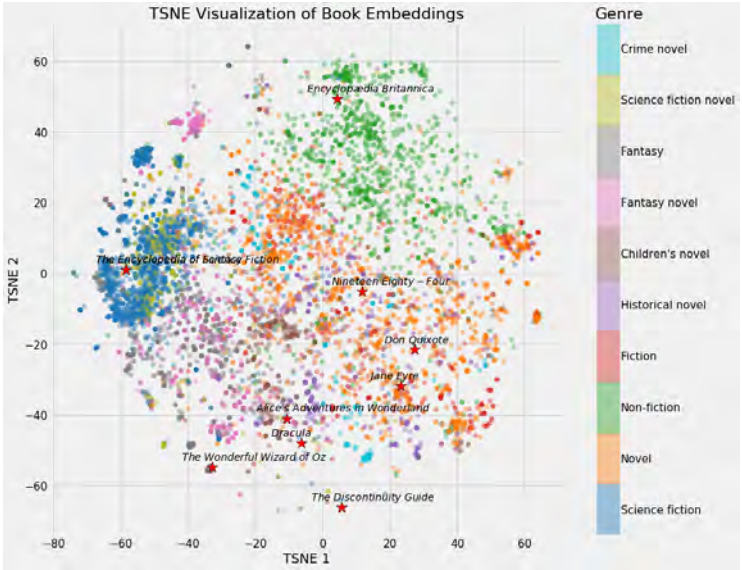
  - 进一步，通过 r 中与另一个 KG 对应的对的数目来限制分母

$$\Pr(r \subseteq r') := \frac{\sum_{r(x,y)}\left(1 - \prod_{r'(x',y')}(1 - (\Pr(x \equiv x') \times \Pr(y \equiv y')))\right)}{\sum_{r(x,y)}\left(1 - \prod_{x',y'}(1 - \Pr(x \equiv x') \times \Pr(y \equiv y'))\right)}$$

4.2 基于 embedding 的方法

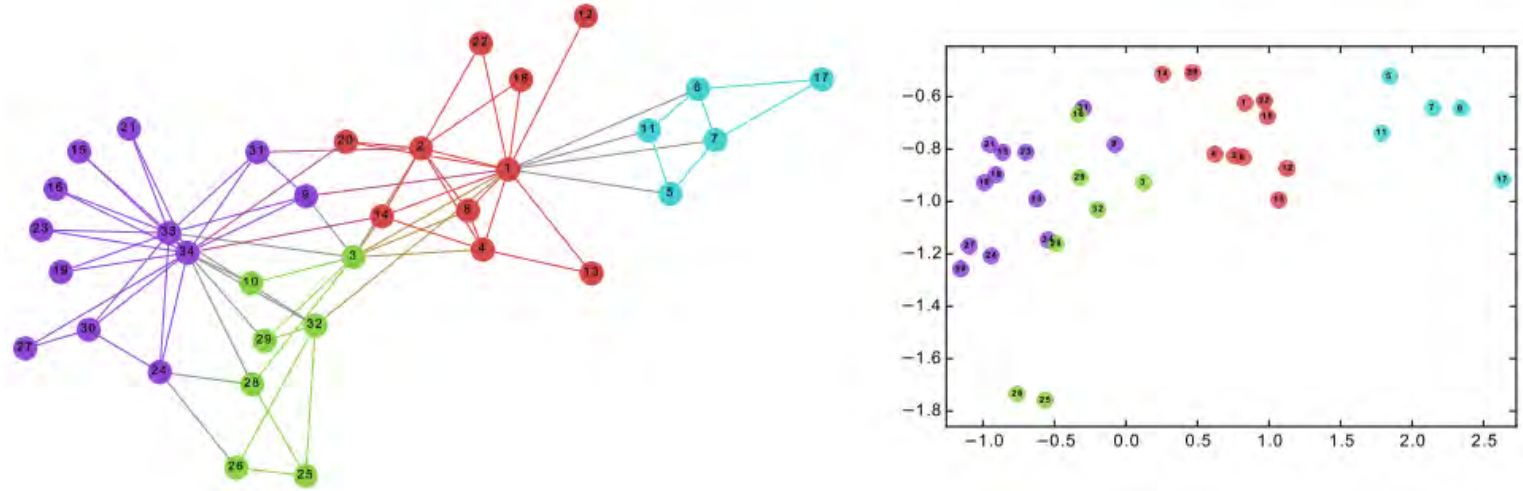
表示学习

- 一个 embedding 是一个离散变量到一个连续数字向量的映射
- 近些年，表示学习技术在诸如图像、视频、语音、自然语言处理等领域取得实质性进展



知识图谱表示学习

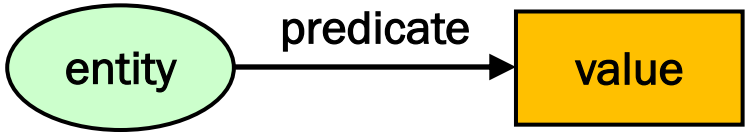
- 知识图谱表示学习的目标是将 KG 的离散符号表示嵌入到连续向量空间中



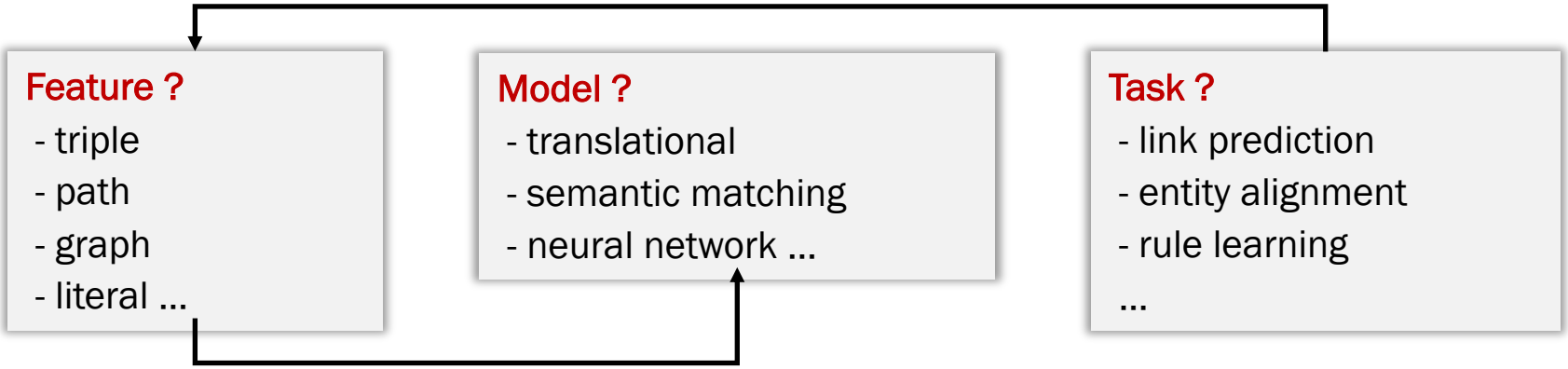
来源：Knowledge graph embedding: A survey of approaches and applications. TKDE, 2017

知识图谱表示学习

- In a KG, a fact is organized as a triple: $(subject, predicate, value)$
  - $Subjects$ are entities
  - $Predicates$ are relations and attributes

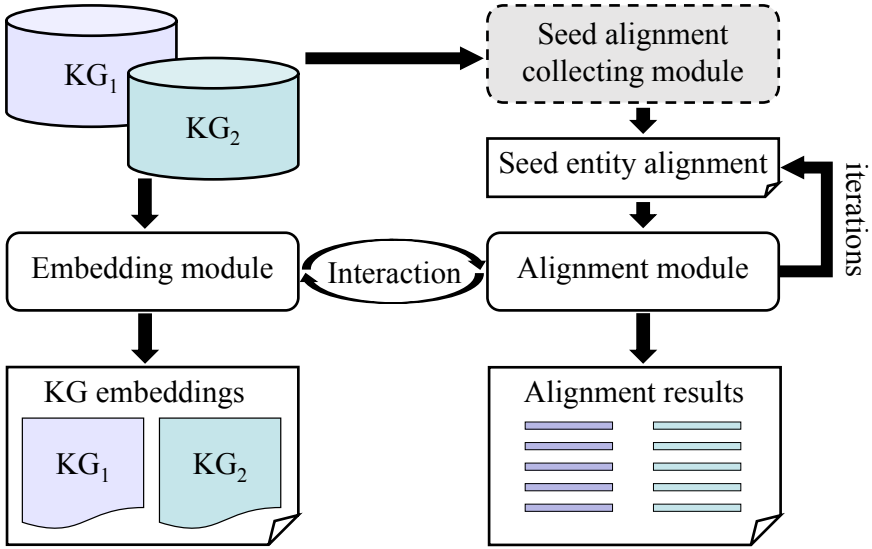


- Knowledge graph embedding



基于 embedding 的实体对齐方法的架构

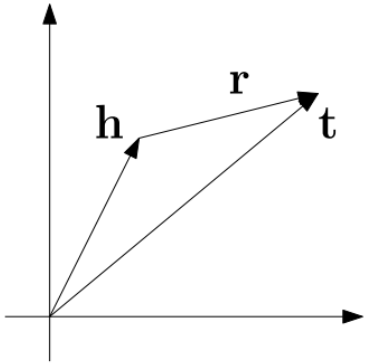
- *Input:* KG_1, KG_2, seed alignment
- Embedding module
  - Relation embedding: triple, path, graph
  - Attribute embedding: attribute, value
- Alignment module
  - Distance metrics: cosine, Euclidean …
  - Result finding algorithms: greedy
- Interaction module
  - Learning strategies: supervised, semi-supervised
  - Combination modes: transition, parameter sharing/swapping …
- *Output:* KG embeddings + alignment results

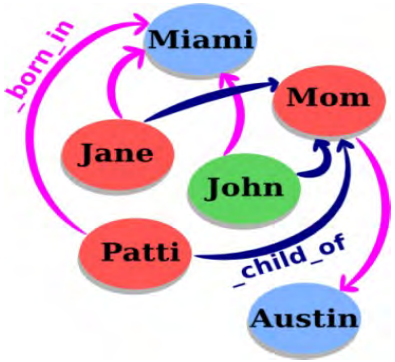


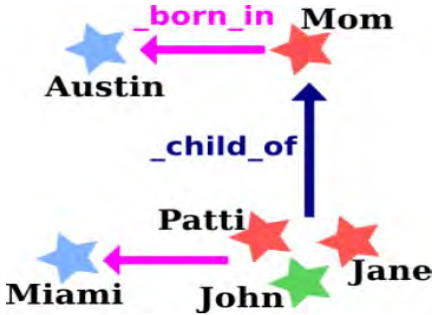
基于翻译模型的方法

翻译模型

- TransE 将关系解释为从其头实体到尾实体的翻译
  - 对于 (h, r, t)，TransE 希望 $\mathbf{h} + \mathbf{r} = \mathbf{t}$







来源：Translating embeddings for modeling multi-relational data. NIPS, 2013

翻译模型：MTransE

- Knowledge model
  - 使用 TransE 来 embedding
    - $S_K = \sum_{L \in \{L_i, L_j\}} \sum_{(h,r,t) \in G_L} \| \mathbf{h} + \mathbf{r} - \mathbf{t} \|$
- Alignment model
  - $S_A = \sum_{(T,T') \in \delta(L_i, L_j)} S_a(T, T')$
- 损失函数
  - $J = S_K + \alpha S_A$

➢ Distance-based axis calibration

$S_{a1} = \| h - h' \| + \| t - t' \|$

$S_{a2} = \| h - h' \| + \| r - r' \| + \| t - t' \|$

➢ Translation vector

$S_{a3} = \| h + v_{ij}^e - h' \| + \| r + v_{ij}^r - r' \| + \| t + v_{ij}^e - t' \|$

➢ Linear transformations

$S_{a4} = \| M_{ij}^e h - h' \| + \| M_{ij}^e t - t' \|$

$S_{a5} = \| M_{ij}^e h - h' \| + \| M_{ij}^r r - r' \| + \| M_{ij}^e t - t' \|$

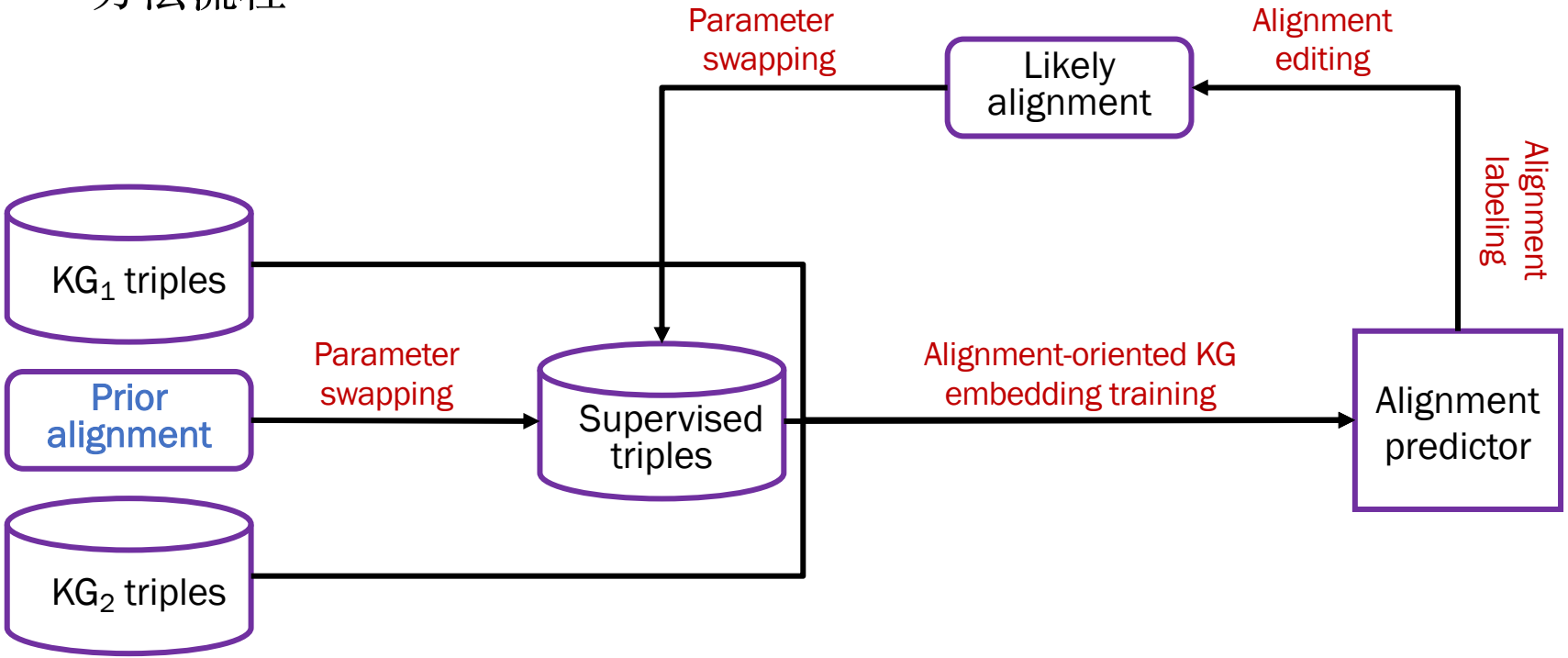
来源：Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. IJCAI, 2017

自训练：BootEA

Parameter swapping

$$T^{s}_{(x,y)} = \{(y,r,t) \mid (x,r,t) \in T_1^+\} \cup \{(h,r,y) \mid (h,r,x) \in T_1^+\}$$
$$\cup \{(x,r,t) \mid (y,r,t) \in T_2^+\} \cup \{(h,r,x) \mid (h,r,y) \in T_2^+\}$$

- 方法流程



来源：Bootstrapping entity alignment with knowledge graph embedding. IJCAI, 2018

自训练：BootEA

- Translational score function: $f(\tau) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2^2$
- Margin-based ranking loss

$$\mathcal{O}_m = \sum_{\tau \in \mathbb{T}^+} \sum_{\tau' \in \mathbb{T}_\tau^-} [\gamma + f(\tau) - f(\tau')]_+$$

$f(\tau') - f(\tau) > \gamma$

not controlled　not controlled

- **Limited loss function**

$$O_e = \sum_{\tau \in T^+} [f(\tau) - \gamma_1]_+ + \sum_{\tau' \in T^-} [\gamma_2 - f(\tau')]_+$$

$f(\tau') - f(\tau) \geq \gamma_2 - \gamma_1$

$f(\tau') \geq \gamma_2$　$f(\tau) \leq \gamma_1$

- Conventional uniform negative sampling

(*Washington DC*, *capitalOf*, *USA*) → (*Tim Berners-Lee*, *capitalOf*, *USA*)

- ϵ-Truncated negative sampling

(*Washington DC*, *capitalOf*, *USA*) → (*New York*, *capitalOf*, *USA*)

自训练：BootEA

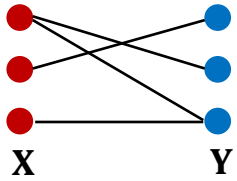
- 候选对齐标记和编辑

$$\max \sum_{x \in \mathbf{X}} \sum_{y \in \mathbf{Y}_x} \pi(y|x;\Theta^{(t)}) \cdot \psi^{(t)}(x,y),$$

$$\text{s.t. } \sum_{x' \in \mathbf{X}} \psi^{(t)}(x',y) \le 1,$$

$$\sum_{y' \in \mathbf{Y}_x} \psi^{(t)}(x,y') \le 1, \forall x,y$$

1-to-1 labeling ⇒ **max-weighted matching** on bipartite graphs

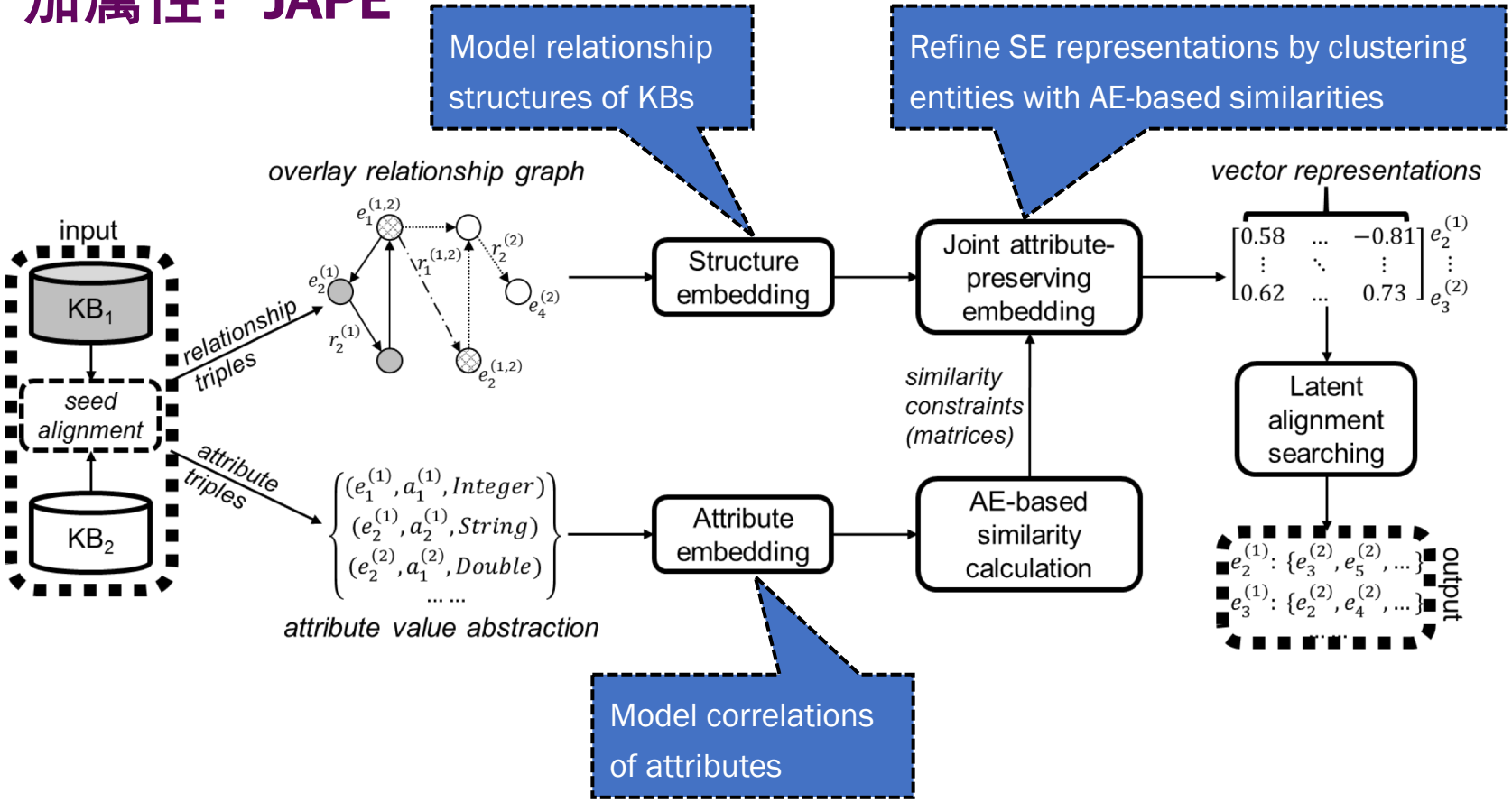


  - 当累积不同迭代轮产生的新标记对齐时，可能存在标记冲突
    - x is labeled as y at the t^{th} iteration while as y' at the $(t+1)^{\text{th}}$ iteration
    - We calculate the likelihood difference

$$\Delta^{(t)}_{(x,y,y')} = \pi(y|x;\Theta^{(t)}) - \pi(y'|x;\Theta^{(t)})$$

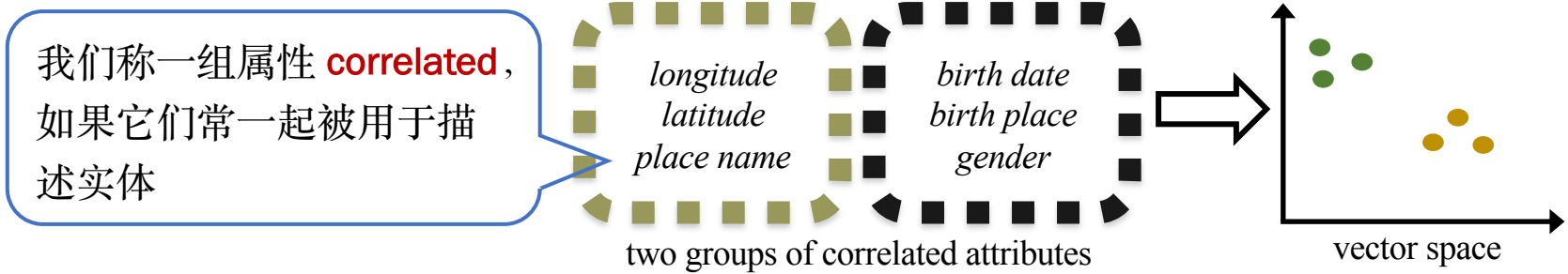
$$\phi_x(y) = \begin{cases} 1_{[y=\hat{y}]} & \text{if } x \text{ is labeled as } \hat{y} \\ \dfrac{1}{|\mathbf{Y}'|} & \text{if } x \text{ is unlabeled} \end{cases}$$

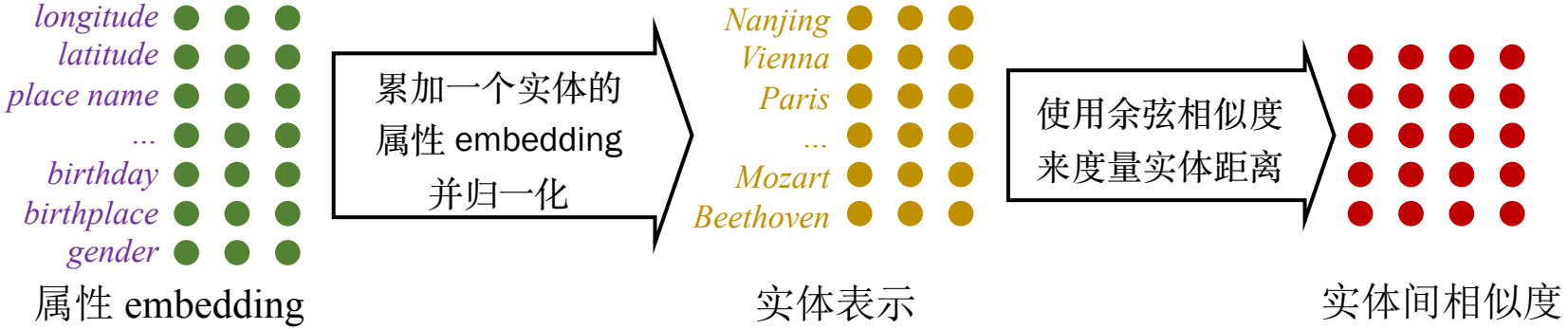
加属性：JAPE



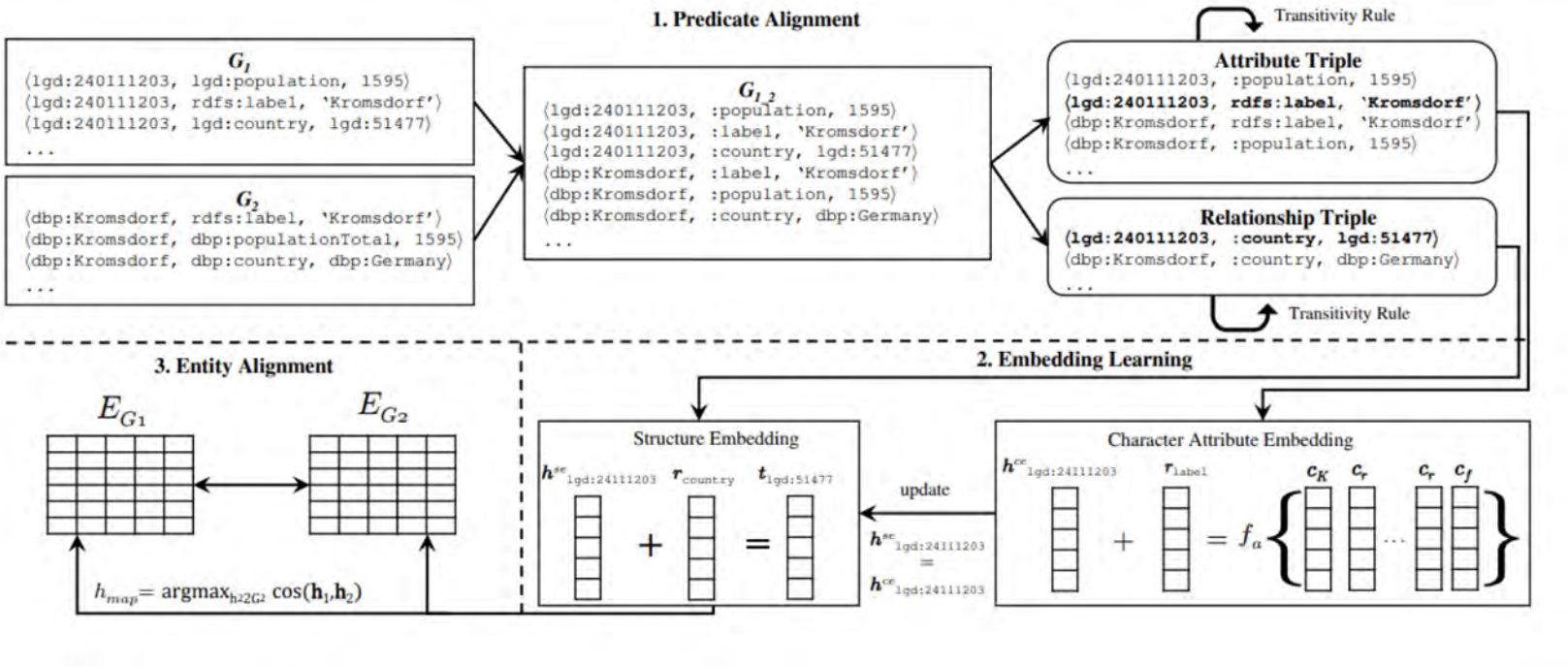
来源：Cross-lingual entity alignment via joint knowledge graph embedding. ISWC, 2017

加属性：JAPE


我们称一组属性 correlated，如果它们常一起被用于描述实体
longitude
latitude
place name
birth date
birth place
gender
two groups of correlated attributes
vector space


longitude
latitude
place name
...
birthday
birthplace
gender
属性 embedding
累加一个实体的属性 embedding 并归一化
Nanjing
Vienna
Paris
...
Mozart
Beethoven
实体表示
使用余弦相似度来度量实体距离
实体间相似度

加取值：AttrE



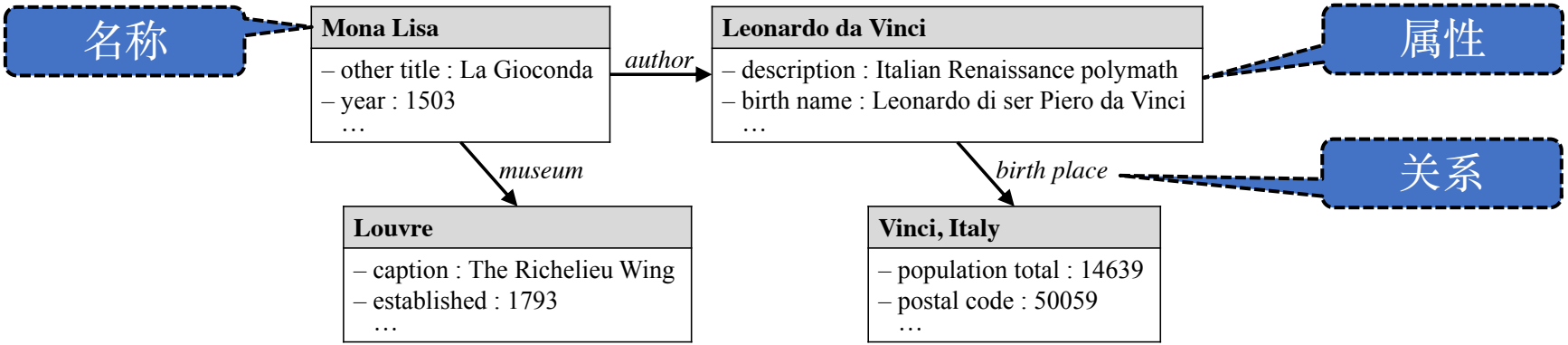
来源：Entity alignment between knowledge graph using attribute embeddings. AAAI, 2018

加取值：AttrE

- Embedding learning
  - The embedding learning module jointly learns the entity embeddings of two KGs using structure embedding and attribute embedding
    - Structure embedding
    - Attribute character embedding
      - Compositional function $f_a(a) \approx h + r$
        - Sum compositional function
        - LSTM-based compositional function
        - N-gram-based compositional function
    - Joint learning of structure embedding and attribute character embedding

多视图：MultiKE

- KG 中的实体常拥有多种特征



  - 当前基于 embedding 的实体对齐方法仅利用了其中一种或两种特征

来源：Multi-view knowledge graph embedding for entity alignment. IJCAI, 2019

多视图：MultiKE

- Literal embedding
  - auto-encoder

$$\varphi(l) = \text{encode}([\text{LP}(o_1); \text{LP}(o_2); \ldots; \text{LP}(on)])$$

- 名称视图 embedding

$$\mathbf{h}^{(1)} = \varphi(\text{name}(h))$$

这里，仅使用基础 embedding 模型进行展示

- 关系视图 embedding
  - TransE

$$f_{rel}(\mathbf{h}^{(2)}, \mathbf{r}, \mathbf{t}^{(2)}) = -\|\mathbf{h}^{(2)} + \mathbf{r} - \mathbf{t}^{(2)}\|$$

- 属性视图 embedding
  - CNN

$$f_{attr}(\mathbf{h}^{(3)}, \mathbf{a}, \mathbf{v}) = -\|\mathbf{h}^{(3)} - \text{CNN}(\langle \mathbf{a}; \mathbf{v} \rangle)\|$$

多视图：MultiKE

- Weighted view averaging
  - 平均不同视图的嵌入

$$w_i = \frac{\cos(h^i, \bar{h})}{\sum_{j=1}^{D} \cos(h^j, \bar{h})}$$

- Sharing space learning
  - 导出一个从每个特定视图嵌入空间到共享空间的正交映射矩阵

$$L_{SSL}(\tilde{H}, Z) = \sum_{i=1}^{D} (\| \tilde{H} - H^i Z^i \|_F^2 + \| I - Z^{iT} Z^i \|_F^2)$$

- In-training combination
  - 参与多视图嵌入的联合训练，从而使多视图彼此受益

$$L_{ITC}(\tilde{H}, H) = \sum_{i=1}^{D} \| \tilde{H} - H^i \|_F^2$$

Input: $\mathcal{G}_a, \mathcal{G}_b$, word embeddings, max epochs Q
Train literal embeddings and get the name embeddings;
for $q = 1, 2, \ldots, Q$ **do**
  Minimize $\mathcal{L}(\Theta^{(2)})$ under the relation view;
  Minimize $\mathcal{L}(\Theta^{(3)})$ under the attribute view;
  if *in-training combination is used* **then**
    Minimize $\mathcal{L}_{\text{ITC}}(\tilde{\mathbf{H}}, \mathbf{H})$;
  Minimize $\mathcal{L}_{\text{CE}}(\Theta^{(2)})$ and $\mathcal{L}_{\text{CE}}(\Theta^{(3)})$;
  Update soft alignment $\mathcal{S}_{\text{rel}}$ and $\mathcal{S}_{\text{attr}}$;
  Minimize $\mathcal{L}_{\text{CRA}}(\Theta^{(2)})$ and $\mathcal{L}_{\text{CRA}}(\Theta^{(3)})$;
if *in-training combination is not used* **then**
  Run weighted view averaging or shared space learning;
Find entity alignment in $\tilde{\mathbf{H}}$ by nearest-neighbor search;

路径：RSN

- 对于 KG 的表示学习，现有方法主要侧重于从实体的关系三元组中学习

- 三元组级别的学习存在两个主要不足
  - 表达能力低
    - 从一个相对局部的角度学习实体表示 (即 1-跳邻居)
  - 信息传播不充分
    - 在 KG 内 / 跨 KG 间，仅使用三元组传递语义信息

来源：Learning to exploit long-term relational dependencies in knowledge graphs. ICML, 2019

路径：RSN

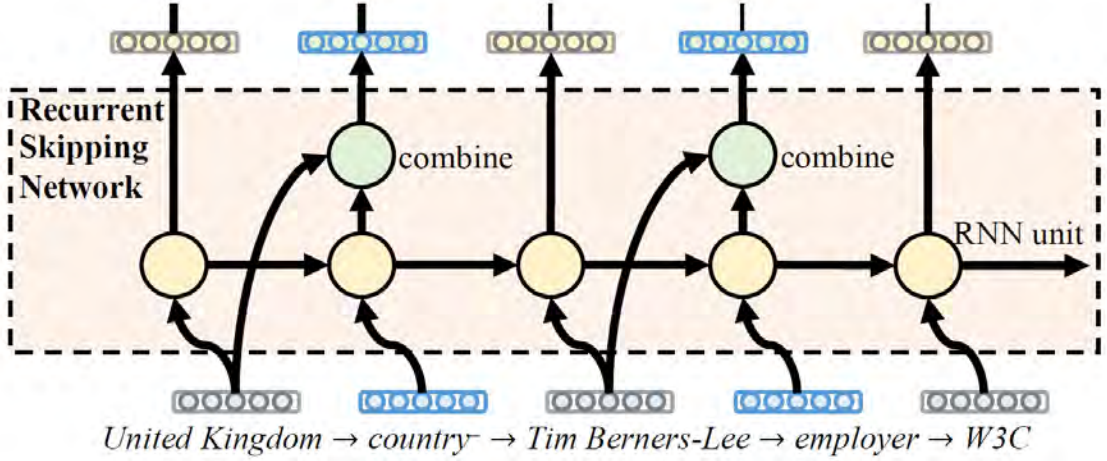
- 一条关系路径是实体-关系链，其中实体和关系交替出现

  United Kingdom → country⁻ → Tim Berners-Lee → employer → W3C

- RNN 在序列数据上表现很好
  - 但是，利用 RNN 来建模关系路径依然存在不足
    1. 关系路径中存在两种不同的元素: “entity” 和 “relation”
       - » 并且它们总是交替出现
    2. 关系路径由三元组组成，但是 RNN 忽略了这种基本结构单元

路径：RSN

- 循环跳跃网络 (recurrent skipping network)
  - 一种条件跳跃机制允许 RSN 短路当前输入实体，使其直接参与预测其宾语实体

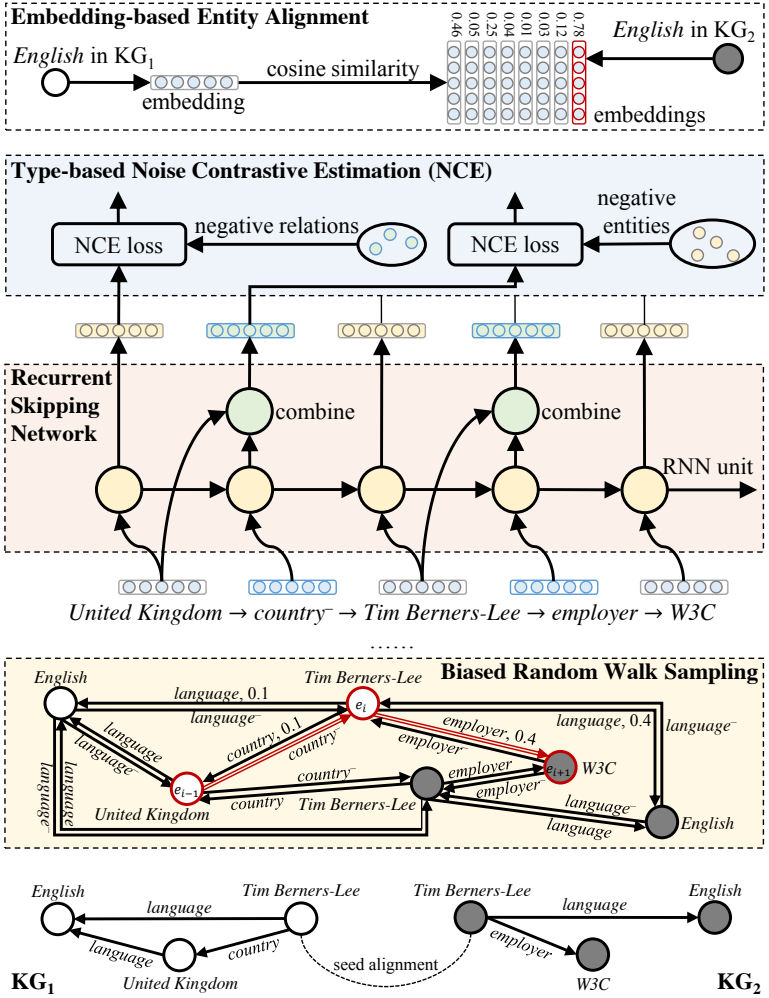


三元残差学习

- Compared with directly learning to predict *W3C* by *employer* and its mixed context, it's easier to learn the residual part between *W3C* and *Tim Berners-Lee*

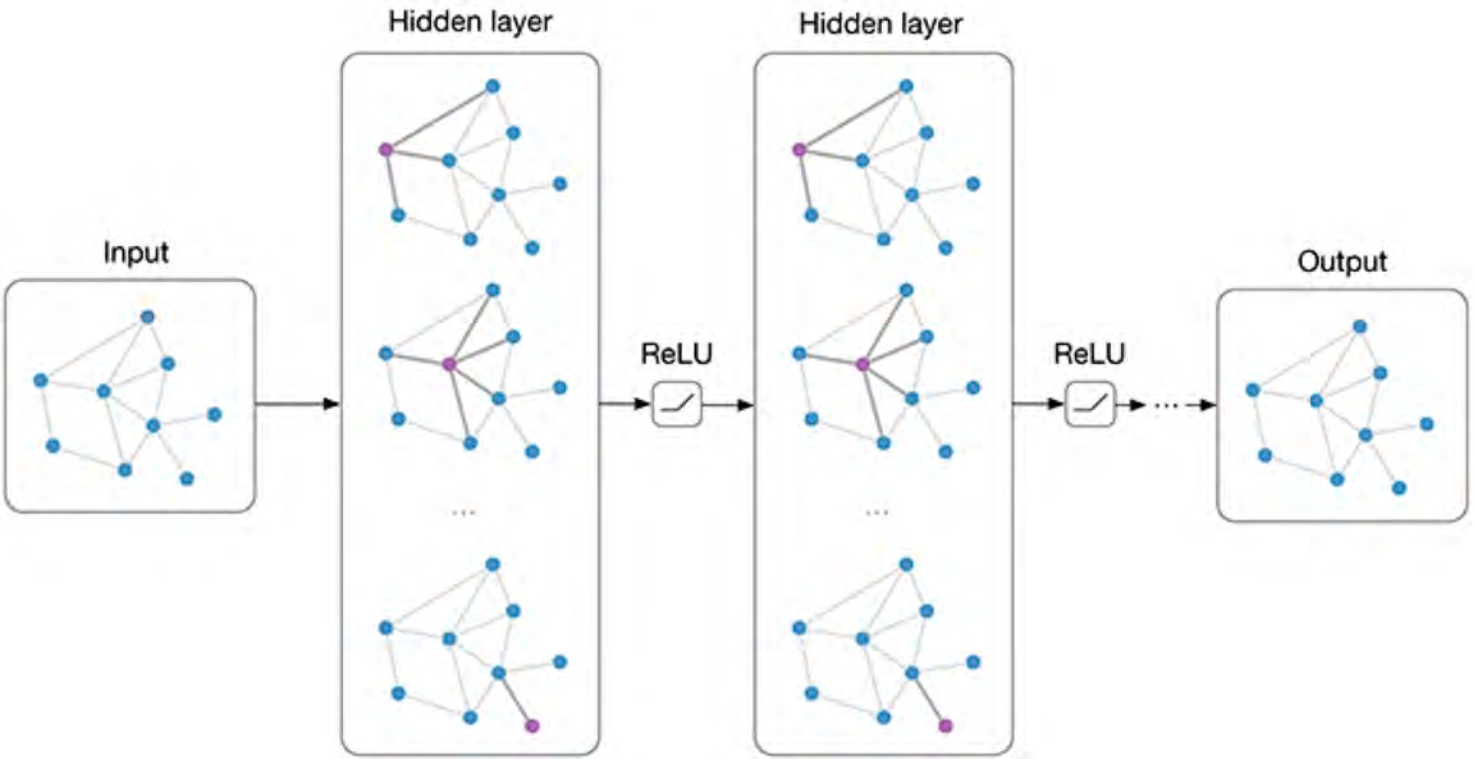
路径：RSN

- 端到端架构
  - 有偏随机游走采样
    - 深的路径比三元组能携带更多的关系依赖
    - 跨 KG 路径能传递 KG 之间的对齐信息
  - 循环跳跃网络
  - 基于类型的噪声对比估计
    - 以一种优化的方式评价损失



基于图神经网络的方法

图神经网络



图神经网络

- 图神经网络 (GNN) 使用节点特征和图结构来学习节点的表示向量
  - 现代 GNN 遵循邻居聚集策略：通过聚集邻居的表示来迭代更新节点的表示
  - 经过 k 次聚集后，节点的表示捕获其 k-跳网络邻居内的结构信息
  - GNN 的第 k 层

    $a_v^k = Aggregate^k\left(\{h_u^{k-1}: u \in N(v)\}\right), \qquad h_v^k = Combine^k\left(h_v^{k-1}, a_v^k\right)$

    其中，h_v^k 是节点 v 在第 k 次循环/层的特征向量，$N(v)$ 是邻接 v 的节点集合

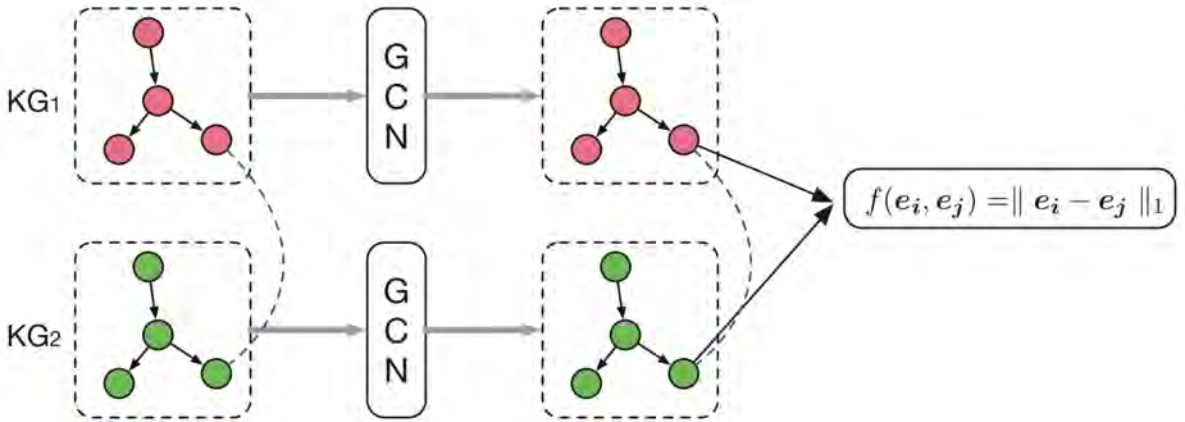
- 图卷积网络 (GCN)

$$h_v^k = ReLU\left(W \cdot Mean\left\{h_u^{k-1}, \forall u \in N(v) \cup \{v\}\right\}\right)$$

使用 mean pooling，并且合并 $Aggregate$ 和 $Combine$ 步骤

基于图神经网络的方法：GCN-Align

- GCN-Align 的基本思想是使用 GCN 将来自不同语言的实体嵌入到统一的向量空间中，其中对齐实体应该尽可能地接近
- GCN-Align 假设
  - 等价实体往往具有相似的属性
  - 等价实体的相邻实体通常也是相互等价的



来源：Cross-lingual knowledge graph alignment via graph convolutional networks. EMNLP, 2018

基于图神经网络的方法：GCN-Align

- GCN-Align 的卷积计算

$$[H_s^{l+1}; H_a^{l+1}] = ReLU(\widehat{D}^{-\frac{1}{2}}\hat{A}\widehat{D}^{-\frac{1}{2}}[H_s^l W_s^l; H_a^l W_a^l])$$

其中，h_s^l is the structure feature vector of l layer, h_a^l is the attribute feature vector of l layer; A is $n \times n$ connectivity matrix $\hat{A} = \mathrm{A} + \mathrm{I}, \mathrm{I}$ is the identity matrix, $\widehat{D}$ is the diagonal node degree matrix of $\hat{A}$; $W^l \in R^{d^l \times d^{l+1}}$ is the weight matrix of l-th layer in the GCN

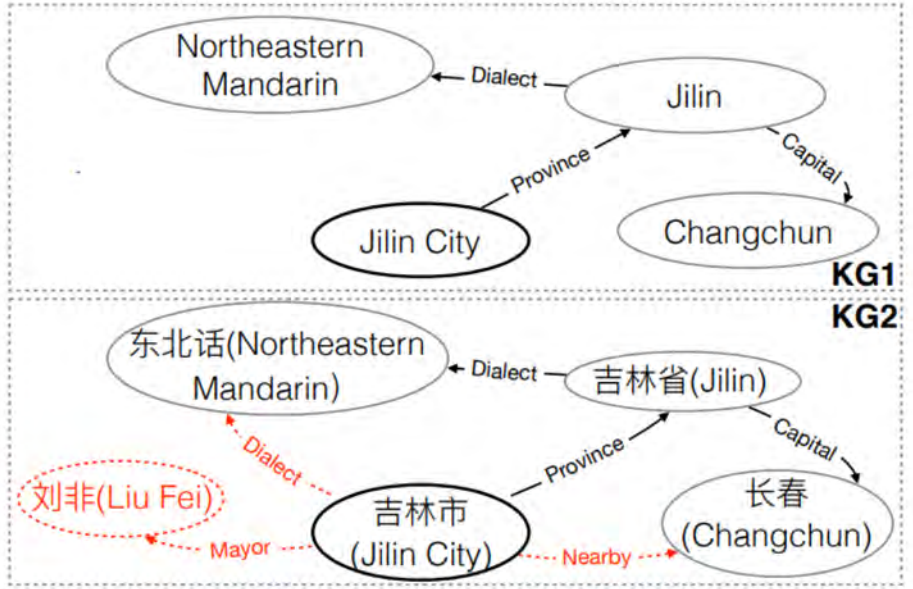
- 连接矩阵计算

$$fun(r) = \frac{\#Head\_Entities\_of\_r}{\#Triples\_of\_r} \quad ifun(r) = \frac{\#Tail\_Entities\_of\_r}{\#Triples\_of\_r} \quad a_{ij} = \sum_{<e_i,r,e_j>\in G} ifun(r) + \sum_{<e_j,r,e_i>\in G} fun(r)$$

其中，$\#Triples\_of\_r$ is the number of triples of relation r; $\#Head\_Entities\_of\_r$ and $\#Tail\_entities\_of\_r$ are the numbers of head entities and tail entities of r, respectively; a_{ij} is the element of position (i, j) in A

基于图神经网络的方法：MuGNN

- MuGNN 关注结构异构性和有限的种子对齐
  - 结构的异构性
    - 由于不完整而导致的关系缺失
    - 由应用程序或语言的不同构建需求引起的专有实体
- MuGNN 包括两个部分
  - 知识图谱补全
  - Multi-channel GNN

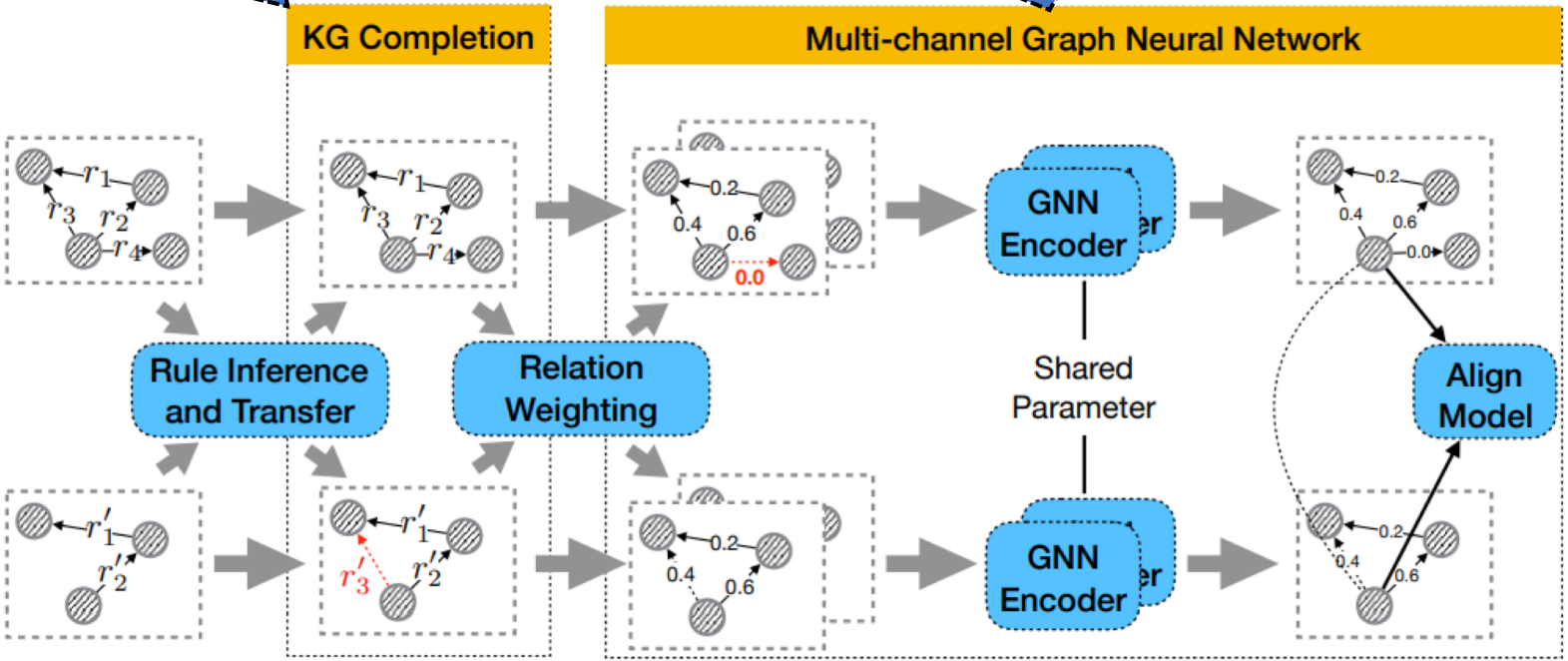


来源：Multi-channel graph neural network for entity alignment. ACL, 2019

MuGNN

通过补全缺失的关系来调和结构差异 (AMIE+)

- 关系加权：根据两种方案为每个 KG 生成权重矩阵
  - KG self-attention 和 cross-KG attention
- GNN encoder：通过 pooling 技术组合不同通道的输出
- 对齐模型：将两个 KG 嵌入到统一的向量空间



基于图神经网络的方法：MuGNN

- Multi-channel GNN
  - KG self-attention

$$a_{ij} = softmax(c_{ij}) = \frac{\exp(c_{ij})}{\sum_{e_k \in N_{e_i} \cup e_i} \exp(c_{ik})}$$

$$c_{ij} = attn(We_i, We_j) = LeakyReLU(p[We_i || We_j])$$

其中，||表示向量连接，W 和 p 是训练参数

  - Cross-KG attention

$$a_{ij} = \max_{r \in R, r' \in R'} 1\left((e_i, r, e_j) \in T\right) sim(r, r')$$

其中，相似度度量 sim(·) 定义为 inner-product $sim(r, r') = r^T r'$

  - Multi-channel GNN encoder

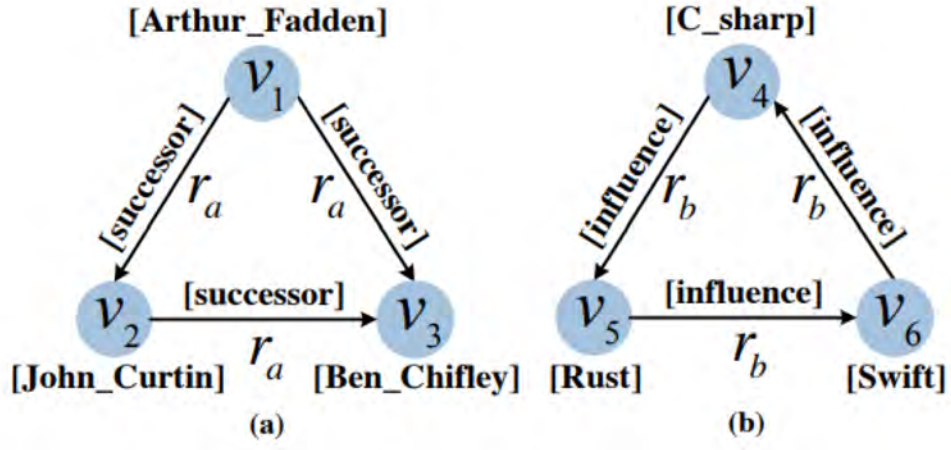
$$MultiGNN(H^l; A_1, \cdots, A_c) = Poling(H_1^{l+1}, \cdots, H_c^{l+1})$$

$$H_i^{l+1} = GNN(A_i, H^l, W_i)$$

其中，c 是通道数量，A_i 是第 i 个通道的连接矩阵，H_i^{l+1} 是第 $(l+1)$ 层第 i 个通道的隐状态

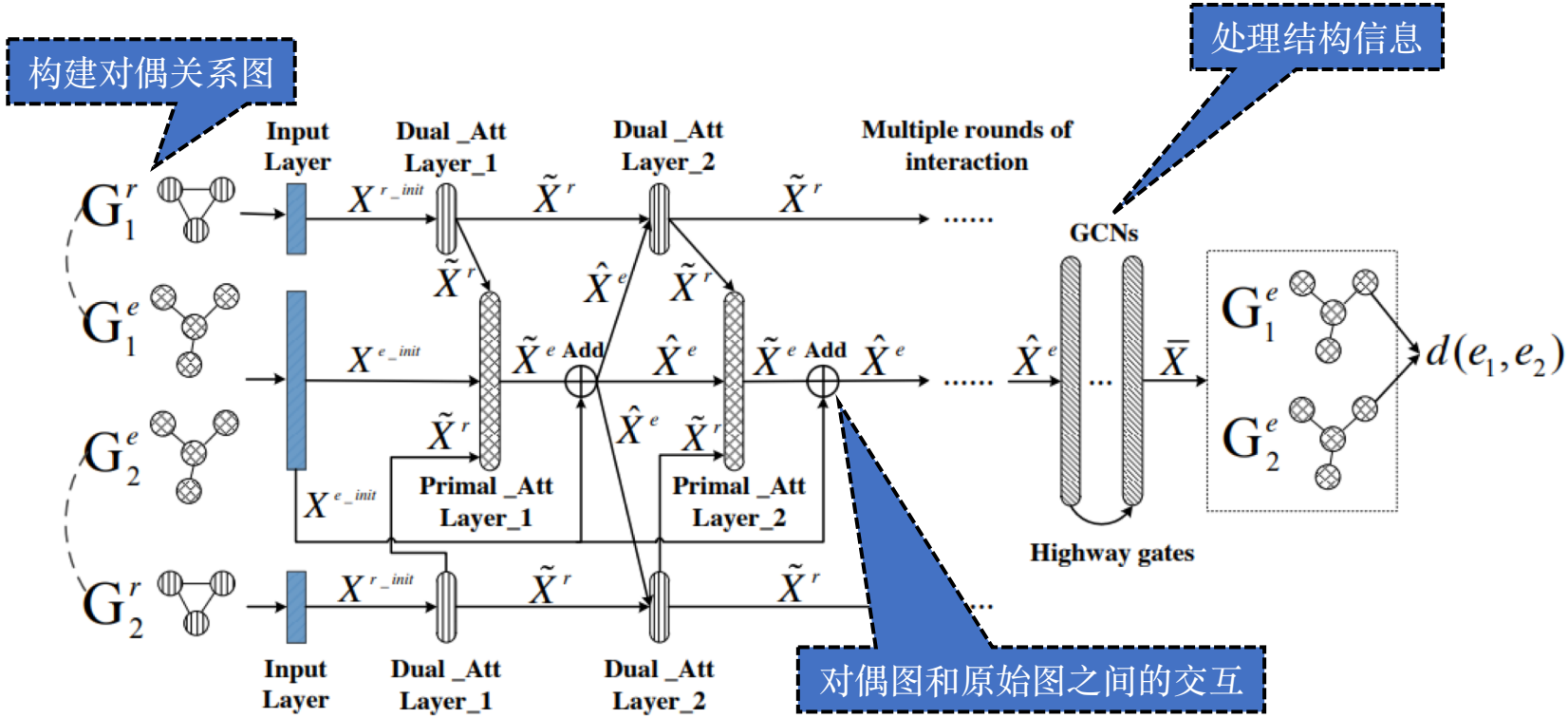
基于图神经网络的方法：RDGCN

- 现有实体对齐工作通常无法正确捕获在多关系 KG 中常见的复杂关系信息
- RDGCN 试图通过 KG 与其对偶关系对应之间的注意力交互来处理关系信息

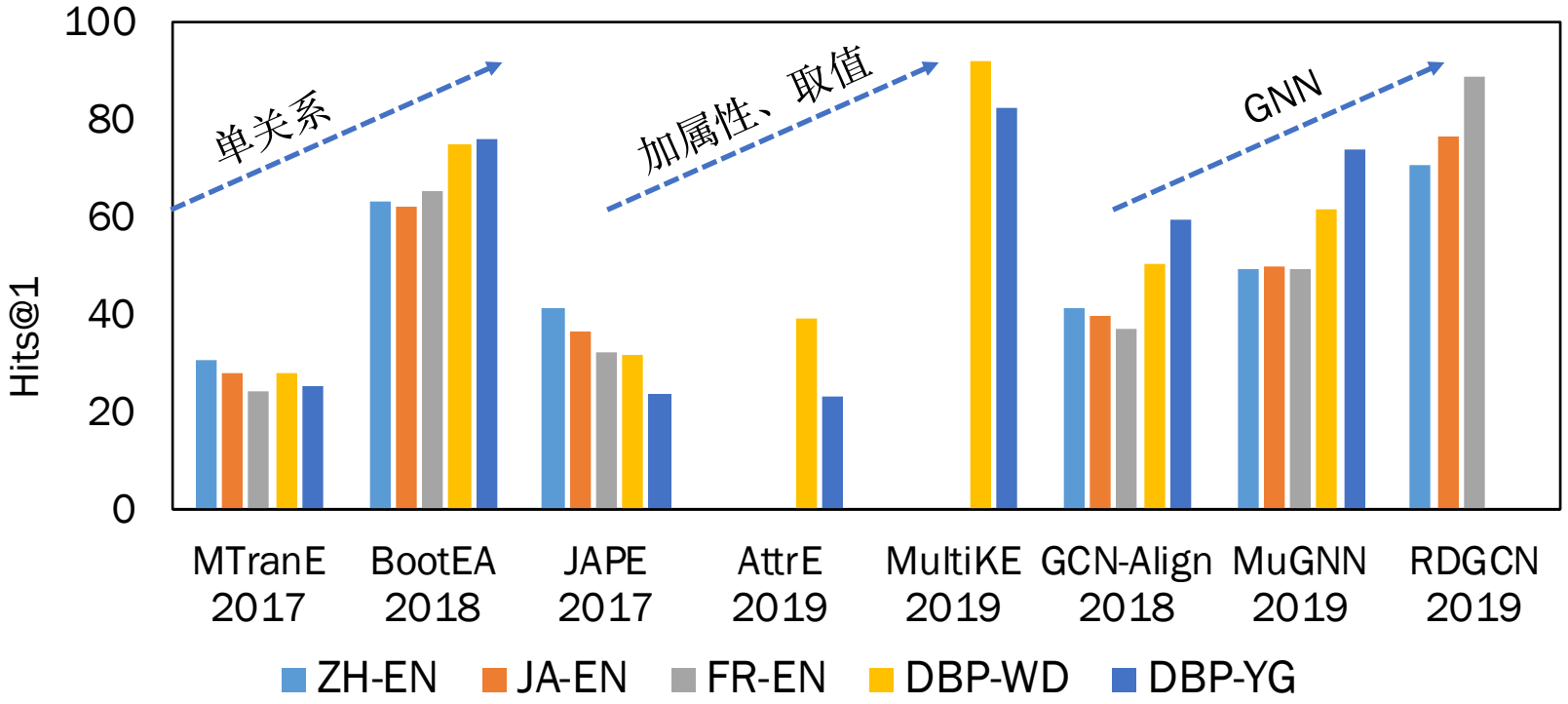


来源：Relation-aware entity alignment for heterogeneous knowledge graphs. IJCAI, 2019

基于图神经网络的方法：RDGCN



基于 embedding 的实体对齐方法实验对比



5. 知识融合

知识融合

对象

不同 KG 中共指实体在同一属性上的取值存在冲突

Entity	Attribute	Value	Source
Mississippi River	Length	2,320 mi (3,730 km)	Wikipedia
Mississippi River (Q1497)	length	2,340 mi	Wikidata
Mississippi River (m.04yf_)	length	3,766 km	Freebase
Mississippi River	length	3,733 km	DBpedia
Mississippi River	Length	2,348 mi	Google KG
Missouri River	Length	2,341 mi (3,767 km)	Wikipedia
Missouri River (Q5419)	length	3,767 km	Wikidata
Missouri River (m.04ykz)	length	3,767 km	Freebase
Missouri River	length	3,767 km	DBpedia
Missouri River	Length	2,341 mi	Google KG
...	...	...	...

事实

找到对齐实体在目标属性上的潜在真值

Entity	Attribute	Value
Mississippi River	Length	?
Missouri River	Length	?

Voting / averaging ?

Entity	Attribute	Value
Mississippi River	Length	?
Missouri River	Length	3,767 km

可行，但效果有限！

知识融合

- 知识抽取时，对于某一个实体，从不同数据源中抽取得到的实体在同一属性上的值存在冲突

Entity	Attribute	Value	Source
Barack Obama	birth of place	Kenya	US Census
Barack Obama	birth of place	Hawaii	Wikipedia
Barack Obama	birth of place	Kenya	BadSource.com
George Washington	birth of place	Virginia	US Census
George Washington	birth of place	Virginia	Wikipedia
George Washington	birth of place	Maryland	BadSource.com
Harry Potter	cast	Daniel Radcliffe	Wikipedia
Harry Potter	cast	Emma Waston	Wikipedia
Harry Potter	cast	Emma Waston	IMDB
...	...	....	...

当低质量数据源很多时，表现差

建模数据源的质量来获得精确的结果!

知识融合

- Voting / averaging
  - Take the value claimed by majority of the sources
  - Compute the mean of all the claims
- 缺点
  - 忽略了数据源的可靠性
  - 当低质量数据源很多时，表现差
- 解决方案
  - 评估数据源的可靠性
    - 通常事先未知

知识融合

Unsupervised

迭代模型

- TruthFinder
  KDD, 2007
- ACCU
  VLDB, 2009
- 3-Estimate
  WSDM, 2010
- Investment
  COLING, 2010
- MBM
  CIKM, 2015

优化模型

- CRH
  SIGMOD, 2014
- CATD
  VLDB, 2015

概率图模型

- LTM
  VLDB, 2012
- GTM
  QDB, 2012
- LCA
  WWW, 2013
- IATD
  CIKM, 2016
- BWA
  WWW, 2019

(Semi-)supervised

- SLiMFast
  SIGMOD, 2017
- SSTF
  WWW, 2011

迭代模型

事实的置信度  数据源的可信度

- 真值的计算
  - 假设数据源的质量是固定的
  - 事实的置信度通过加权聚合提供该事实的数据源来推断
- 数据源质量的估计
  - 数据源的可靠性基于当前识别出的事实来估计

迭代，直至达到稳定状态

迭代模型：TruthFinder

- 数据源 s 的置信度：$w(s)$
  - 它提供的事实的置信度的平均值

$$w(s) = \frac{\sum_{f \in F(s)} t(f)}{|F(s)|}$$

事实的置信度之和

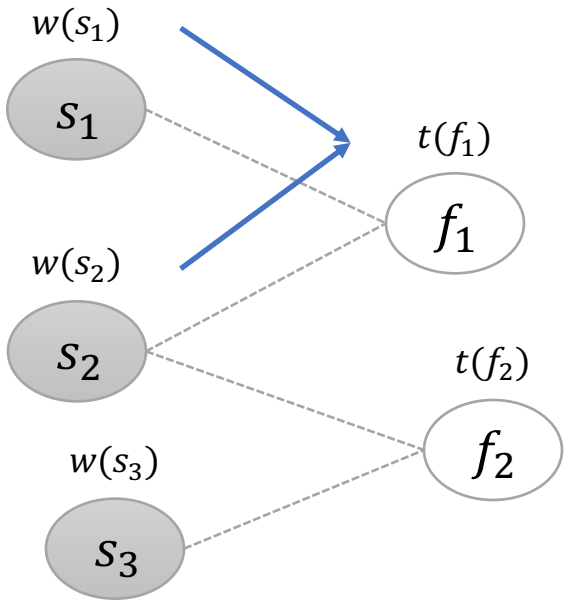
s 提供的事实集合

- 事实 f 的置信度：$t(f)$

$$t(f) = 1 - \prod_{s \in S(f)} (1 - w(s))$$

s 是错误的概率

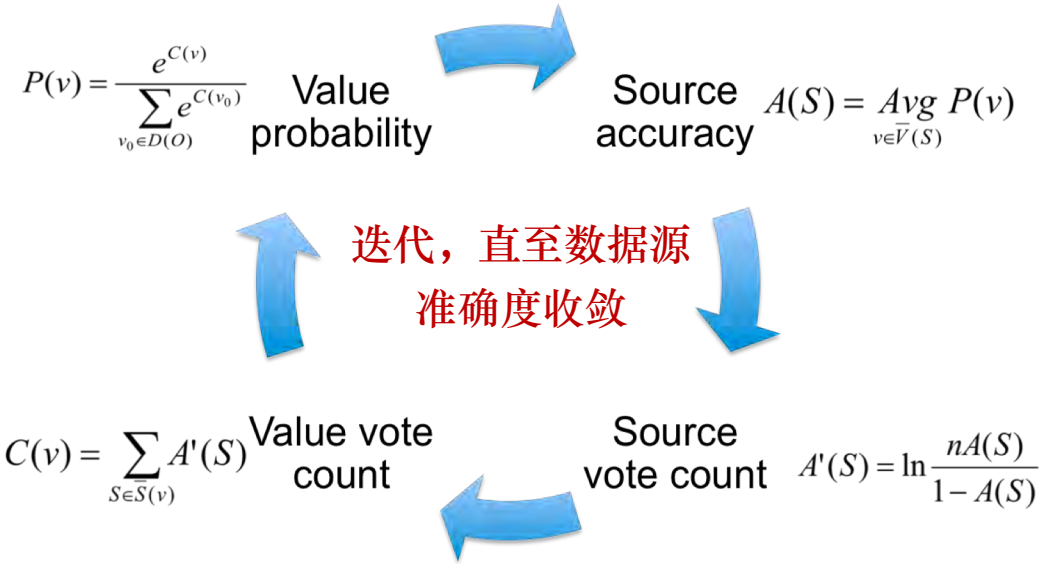
提供 f 的数据源集合



来源：Truth discovery with multiple conflicting information providers on the web. TKDE, 2008

迭代模型：ACCU

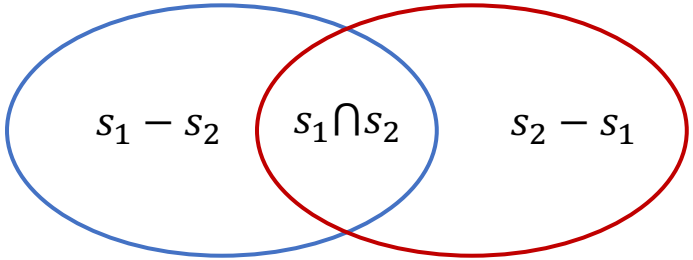
- 联合建模和预测

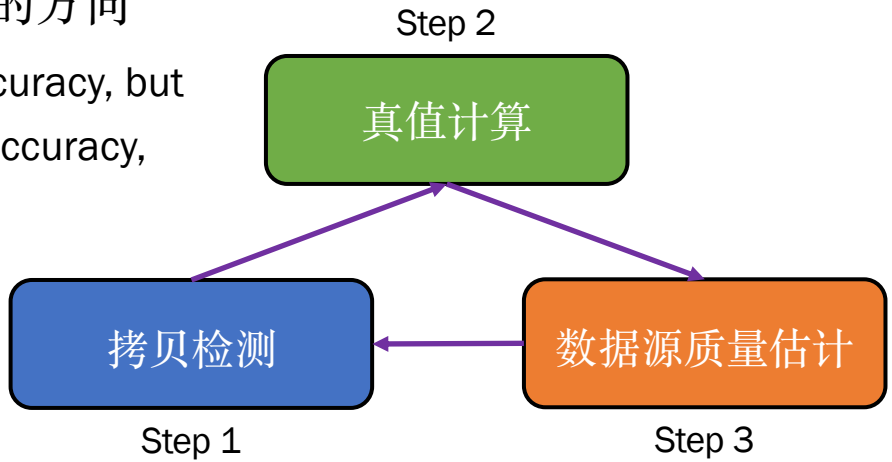


来源：Integrating conflicting data: The role of source dependence. VLDB, 2009

迭代模型：ACCU with source copy detection

- Source copy detection 的直观想法
  - 共同的错误意味着拷贝关系
    - Many same errors in $s_1 \cap s_2$ imply source s_1 and s_2 are related
  - 数据源可靠性的差异意味着拷贝的方向
    - $s_1 \cap s_2$ and $s_1 - s_2$ has similar accuracy, but $s_1 \cap s_2$ and $s_2 - s_1$ has different accuracy, so source s_2 may be a copier





优化模型

- 通用模型

$$\arg\min_{\{w_s\},\{v_o^*\}} \sum_{o\in O}\sum_{s\in S} g(w_s, v_o^*)$$
$$\text{s.t.} \quad \delta_1(w_s), \delta_2(v_o^*)$$

  - 联合估计真值 v_o^* 和数据源可靠性 w_s，满足约束 $\delta_1, \delta_2, \ldots$
  - $g(\cdot,\cdot)$ 可以是距离、熵等
  - 如果每个子问题都是突且光滑的，那么最优解可以通过最小化目标函数获得
  - 原始问题可以转换为（拉格朗日）对偶形式
  - 可以采用坐标下降来更新参数并求出解

优化模型：CRH

- 真值应该接近于可靠数据源的取值
  - Minimize the overall weighted distances to the truths in which reliable sources have high weights

$$\min_{\mathcal{X}(*),\mathcal{W}} f(\mathcal{X}(*),\mathcal{W}) = \sum_{k=1}^{K} w_k \sum_{i=1}^{N}\sum_{m=1}^{M} d_m(v_{im}^{(*)}, v_{im}^{(k)})$$

$$\text{s.t. } \delta(\mathcal{W}) = 1, \mathcal{W} \geq 0$$

$d_m(\cdot,\cdot)$ denotes the loss on the data type of the m^{th} property

◆ 类别数据

$$d_m\left(v_{im}^{(*)}, v_{im}^{(k)}\right) = \begin{cases} 1 & \text{if } v_{im}^{(k)} \neq v_{im}^{(*)} \\ 0 & \text{otherwise} \end{cases}$$

◆ 连续数据

$$d_m\left(v_{im}^{(*)}, v_{im}^{(k)}\right) = \frac{\left(v_{im}^{(*)}, v_{im}^{(k)}\right)^2}{std\left(v_{im}^{(1)}, \ldots, v_{im}^{(K)}\right)}$$

来源：Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. SIGMOD, 2014

优化模型：CRH

- 真值计算
  - 最小化真值和数据源提供的取值之间的加权距离

$$v_{im}^{(*)} \leftarrow \arg\min_{v} \sum_{k=1}^{K} w_k \cdot d_m(v, v_{im}^{(k)})$$

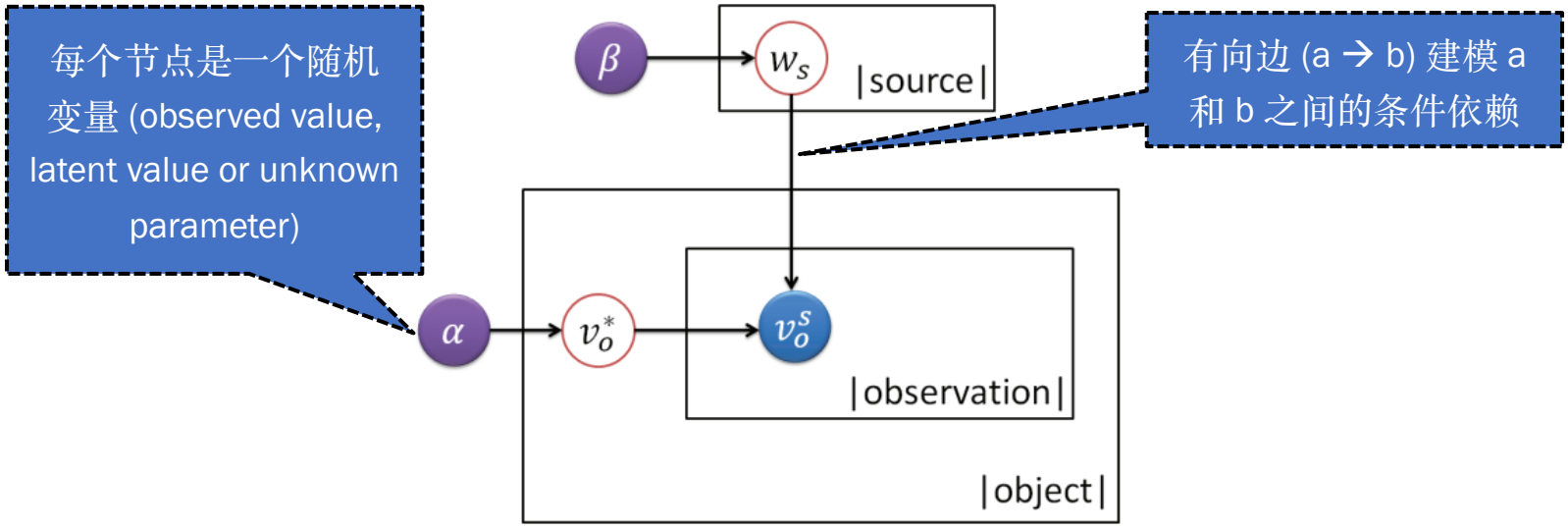
- 数据源质量估计
  - 基于真值和数据源观测值之间的差异，给每个数据源分配一个权重

$$\mathcal{W} \leftarrow \arg\min_{\mathcal{W}} f(\mathcal{X}(*), \mathcal{W})$$

概率图模型

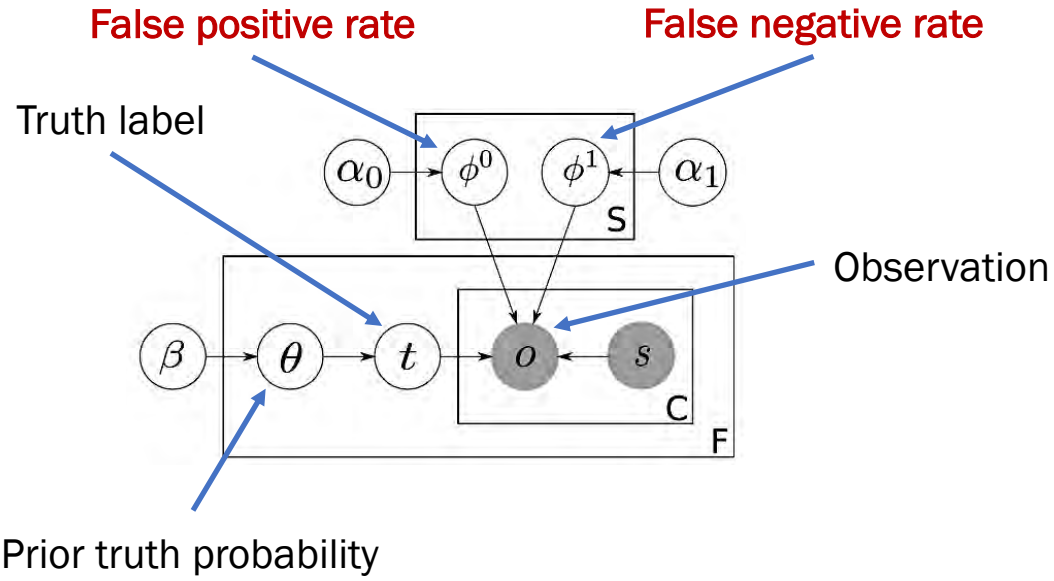
- 通用模型

$$\prod_{s\in S} p(w_s|\beta) \prod_{o\in O}\left(p(v_o^*|\alpha)\prod_{s\in S} p(v_o^s|v_o^*, w_s)\right)$$



概率图模型：LTM for categorical data

- Use Bayesian Network to model the source trustworthiness, fact truthfulness and claims from sources



$$\phi_s^0 \sim Beta(\alpha_{0,1}, \alpha_{0,0})$$

$$\phi_s^1 \sim Beta(\alpha_{1,1}, \alpha_{1,0})$$

$$\theta_f \sim Beta(\beta_1, \beta_0)$$

$$t_f \sim Bernoulli(\theta_f)$$

$$o_c \sim Bernoulli(\phi_{s_c}^{t_f})$$

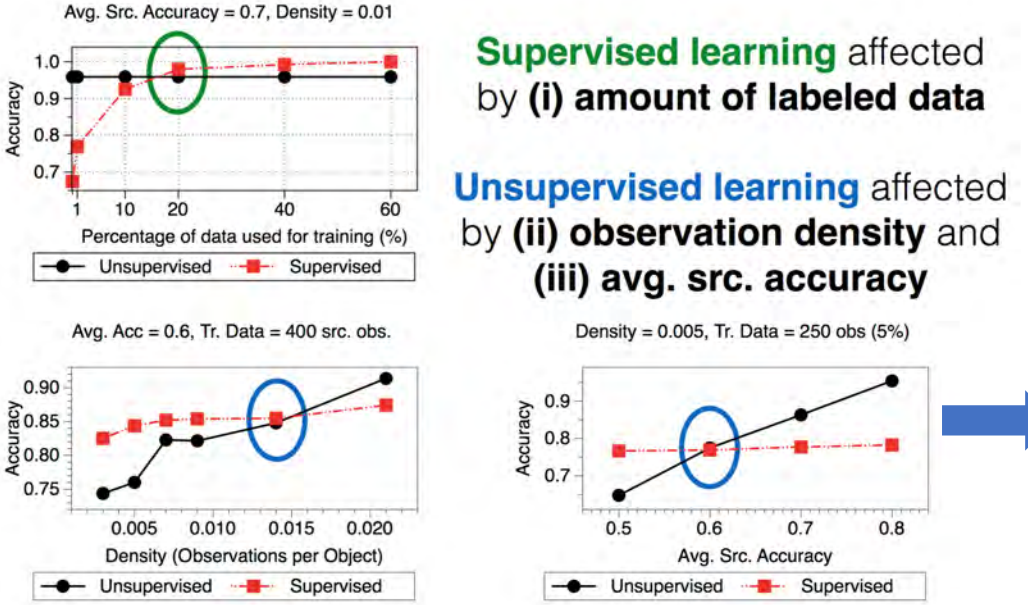
来源：A Bayesian approach to discovering truth from conflicting sources for data integration. VLDB, 2012

知识融合

- 无监督模型利用数据冲突和数据源质量估计
- 有监督模型利用特定领域特征和少量标记数据

Unsupervised *vs.* Supervised

知识融合

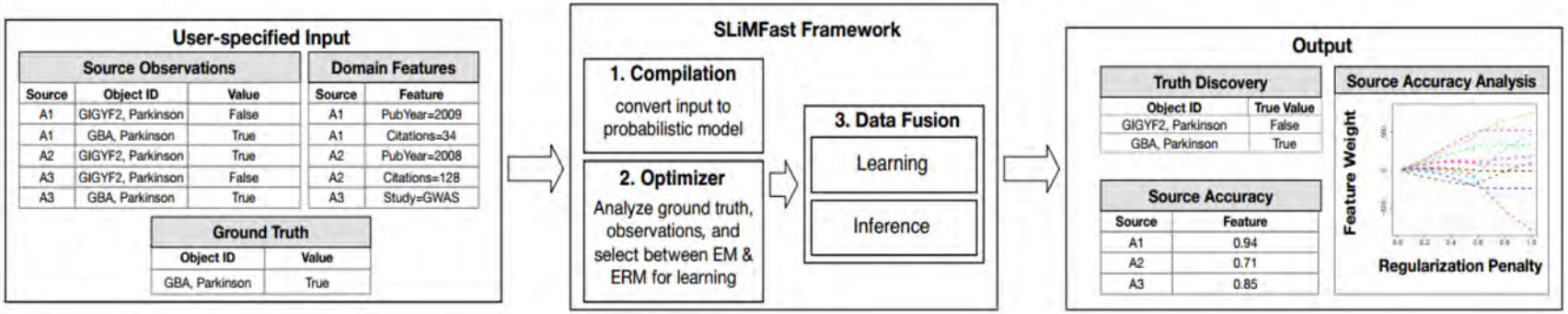


- 如果标记数据充足，则 使用有监督学习
- 如果观测密度较高、平均数据源准确度较高，则用无监督学习

来源：Fact checking: Theory and practice. KDD, 2018

半监督模型：SLiMFast

- 使用具有严格理论保障的判别概率模型
  - Combine cross-source conflicts with domain-specific features
  - Obviate the need for users to manually select an algorithm for learning parameters
  - Provide formal error bounds and present a series of theoretical guarantees



来源：SLiMFast: Guaranteed results for data fusion and source reliability. SIGMOD, 2017

半监督模型：SLiMFast

- SLiMFast 的判别模型

$$P(T_o = v|\Omega) = \frac{1}{Z} exp \sum_{(o,s)\in\Omega} \sigma_s \mathbf{1}_{v_{o,s}=v}$$

$$Z = \sum_{v\in V_o} exp \sum_{(o,s)\in\Omega} \sigma_s \mathbf{1}_{v_{o,s}=v}$$

$$\sigma_s = \log\left(\frac{A_s^*}{1-A_s^*}\right) = \log\left(\frac{P(v_{o,s} = v_o^*)}{1-P(v_{o,s} = v_o^*)}\right)$$

$$A_s = 1/(1+\exp(-w_s - \sum_{k\in K} w_k f_{s,k})$$

source reliability score (model parameters)

importance of domain-specific features (model parameters)

$$P(T_o = v|\Omega; w) = \frac{1}{Z}\exp\left(\sum_{(o,s)\in\Omega} (w_s + \sum_{k\in K} w_k f_{s,k})\, \mathbf{1}_{v_{o,s}=v}\right)$$

normalizing constant (valid probability)

indicator function

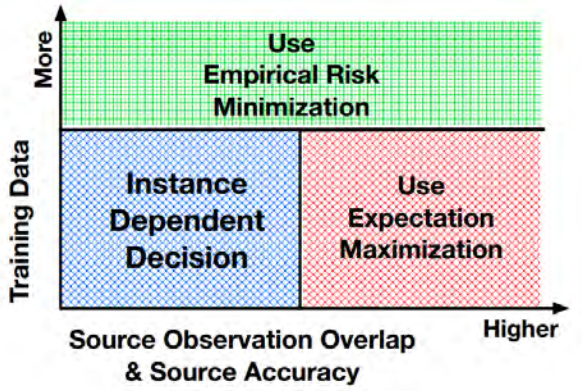
with $w = (\langle w_s \rangle_{s\in S}, \langle w_k \rangle_{k\in K})$

半监督模型：SLiMFast

- 模型求解
  - 通过优化似然 $\ell(w) = \log P(T|\Omega; w)$，学习逻辑回归的参数 w
  - 推断变量 T 的最大后验 (MAP) 分配

$$P(T_o = v|\Omega; w) = \frac{1}{Z}\exp\Big(\sum_{(o,s)\in\Omega}(w_s + \sum_{k\in K} w_k f_{s,k})\,\mathbf{1}_{v_{o,s}=v}\Big) \quad \text{with } w = (\langle w_s\rangle_{s\in S}, \langle w_k\rangle_{k\in K})$$

- 优化器
  - 当有足够的 ground truth 可用时，使用经验风险最小化 (ERM) 来计算其逻辑回归模型的参数
  - 当 ground truth 有限或不可用时，使用期望最大化 (EM) 来计算使数据源观测似然最大的参数 Ω



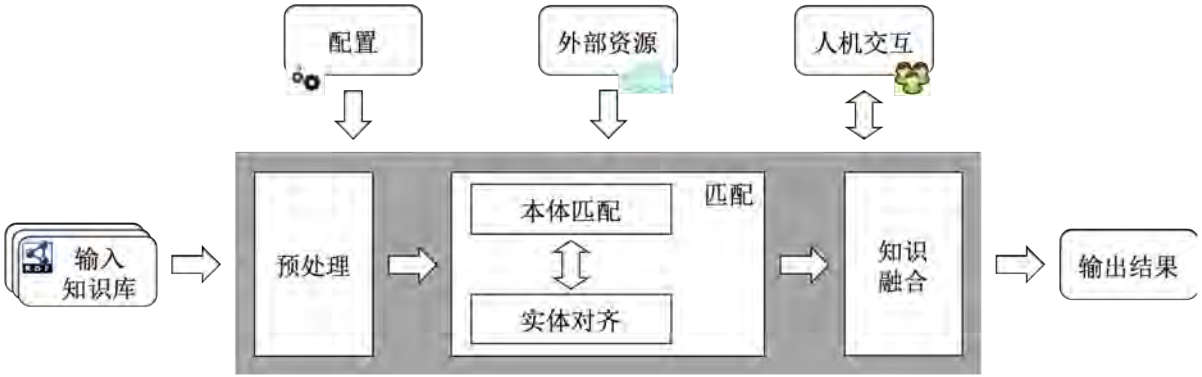
知识融合

- 对象关系
  - 绝大多数模型假设对象间相互独立，但是知识图谱中的对象有多种关联方式
    - 例如，“the birth year of a person” 和 “the age of a person”
- 模型选择
  - No free lunch theorem
  - 给定各种模型，如果针对特定任务选择合适的模型？或者是否可以同时应用多种模型，再组合它们的输出？
- 理论分析
  - 模型是否会收敛？收敛率如何？是否可以知道收敛结果的误差界线？

6. 总结与展望

总结

- 知识图谱已经得到了广泛的研究及应用
  - 知识图谱融合是“知识图谱构建”到“基于知识图谱的智能应用”中的重要一环
    - 目标：将不同知识图谱融合为一个统一、一致、简洁的形式
    - 主要挑战：异构性、大规模、动态性 …
    - 关键技术：知识表示与推理、自然语言处理、机器学习、深度学习 …



展望

- 深度学习与知识推理的结合
  - 描述逻辑、约束规则、众包知识 ...
- 动态 / 多模态知识图谱融合
  - Emerging 实体、长尾实体 / 数据源、图片 ...
- 大规模测试集构建
  - 人造数据集质量高，但规模小
  - 真实世界数据集规模大，但缺乏高质量黄金标准

南京大学 NANJING UNIVERSITY

南京大学万维网软件研究组

PPT 下载：

https://github.com/nju-websoft/KnowledgeGraphFusion

致谢：

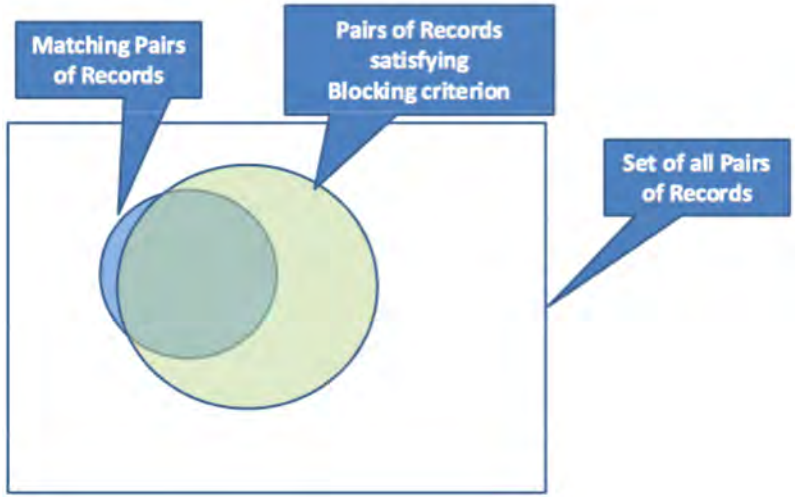
- CCF 学科前沿讲习班第 108 期：知识图谱
- 学生：黄佳程、曹二梅、王成名
- 国家重点研发计划课题 (2018YFB1004304)，国家自然科学基金项目 (61872172)

联系方式：胡伟，南京大学，邮箱：whu@nju.edu.cn

补充： Blocking

Blocking

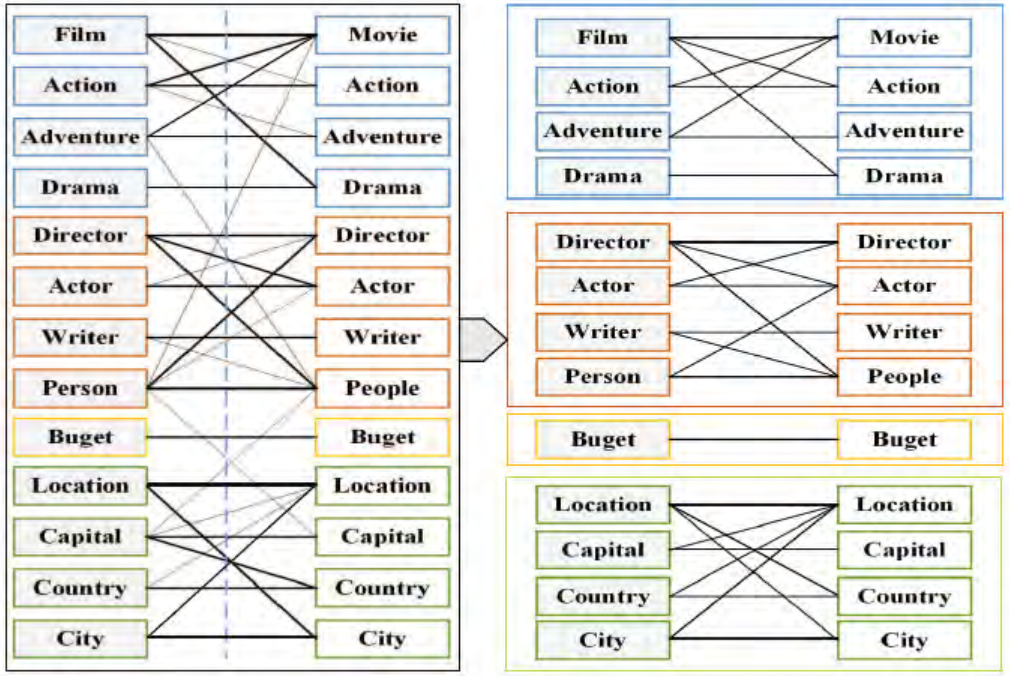
- Naïve pairwise: N^2 pairwise comparisons
  - 1,000 business listings each from 1,000 different cities across the world
  - 1 trillion comparisons, 11.6 天 (if each comparison is 1 μs)
- Mentions from different cities are unlikely to be matches
  - Blocking criterion: City
  - 1 billion comparisons, 16 分钟
- Class-based blocking
- Token-based blocking
- Attribute-clustering token-based blocking



Class-based Blocking

- Different classes of entities are not required to be compared
  - One entity is a person, the other entity is a film

- Heterogeneous classes
  → class 协同聚类



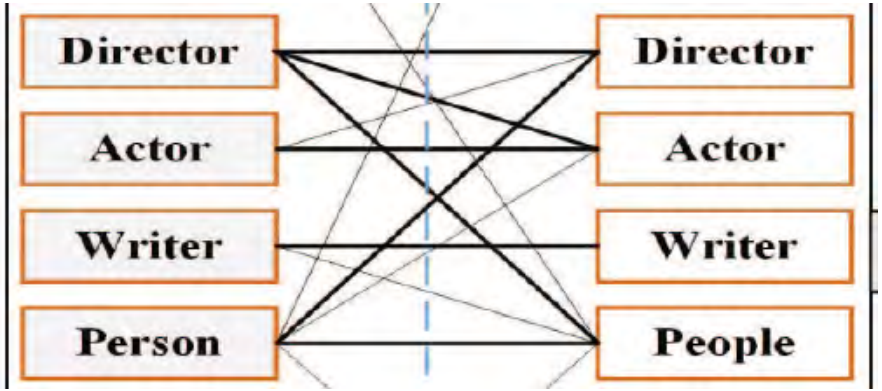
来源：PBA: Partition and blocking based alignment for large knowledge bases. DASFAA, 2016

Class-based Blocking

- Class 协同聚类
  - Jaccard similarity of two class clusters

$$\mathrm{sim}(C_1, C_2) = \frac{\mathbb{N}_I(C_1, C_2)}{|C_1| + |C_2| - \mathbb{N}_I(C_1, C_2)}$$

where $\mathbb{N}_I$ is the number of matched instances



1. Merge **Director** and **Person** in left
2. Merge **Director**, **Actor** and **People** in right

...

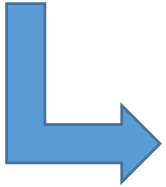
- 层次化协同聚类算法
  - Iteratively merge the partitions from the priority queue according to the similarity measure θ
  - Terminate until either the number of partitions in the queue reaches 1 or there is no possibility to merge more partitions

Token-based blocking

- 当属性相同或对齐时
  - Generate BKVs for entities, build an inverted index for BKVs, and compare entity pairs that have at least one same BKVs

Record fields				
Identifiers	Givennames	Surnames	Postcodes	Suburb names
R1	Peter	Christen	2010	North Sydney
R2	Pedro	Kristen	2000	Sydney
R3	Paul	Smith	2600	Canberra
R4	Pablo	Smyth	2700	Canberra S

Generate blocking key values (BKVs)



Number of comparisons: $\frac{4\times3}{2} = 6 \Rightarrow 2$

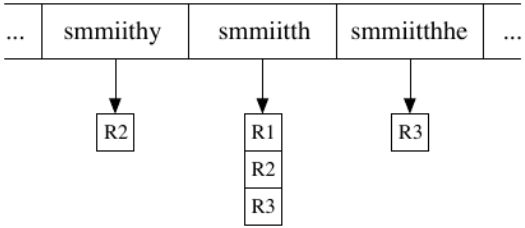
Blocking keys and BKVs		
Sndx(GiN) + PC	Fi2D(PC) + DMe(SurN)	Sndx(SubN) + La2D(PC)
P360-2010	20-KRST	N632-10
P360-2000	20-KRST	S530-00
P400-2600	26-SM0	C516-00
P140-2700	27-SM0	C516-00

来源：A survey of indexing techniques for scalable record linkage and deduplication. TKDE, 2012

Token-based blocking

- Q-gram 索引

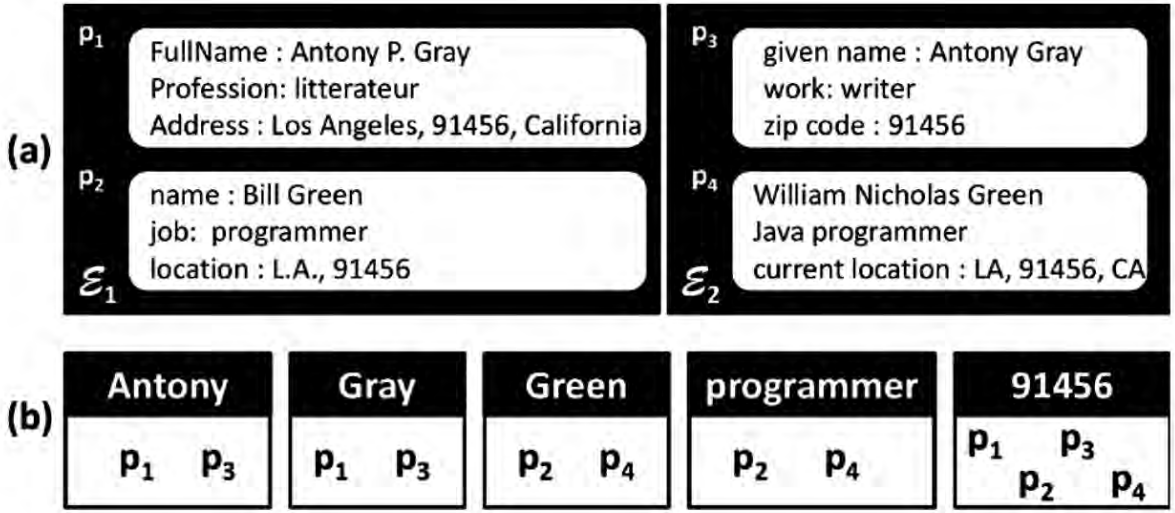
Identifiers	BKVs (Surname)	Bigram sub-lists	Index key values
R1	Smith	[sm,mi,it,th], [mi,it,th], [sm,it,th], [sm,mi,th], [sm,mi,it]	**smmiitth**, miitth, smitth, smmith, smmiit
R2	Smithy	[sm,mi,it,th,hy], [mi,it,th,hy], [sm,it,th,hy], [sm,mi,th,hy], [sm,mi,it,hy], [sm,mi,it,th]	smmiitthhy, miitthhy, smitthhy, smmithhy, smmiithy, **smmiitth**
R3	Smithe	[sm,mi,it,th,he], [mi,it,th,he], [sm,it,th,he], [sm,mi,th,he], [sm,mi,it,he], [sm,mi,it,th]	smmiitthhe, miitthhe, smitthhe, smmithhe, smmiithe, **smmiitth**



- 其他索引
  - Sorted neighborhood, canopy clustering, string map, suffix array …

Attribute-clustering Token-based Blocking

- 当属性异构时
  - Group attribute names into clusters such that we can apply token-based blocking independently inside each cluster



来源：A blocking framework for entity resolution in highly heterogeneous information spaces. TKDE, 2012

Attribute-clustering Token-based Blocking

- 算法
  - Create a graph, every node represents an attribute name and its attribute values
  - For each attribute name/node n_i
    - Find the most similar node n_j
    - If $\text{sim}(n_i, n_j) > 0$, add an edge (n_i, n_j)
  - Extract connected components
  - Put all nodes in a cluster

- Once attribute clusters are obtained, we can use token-based blocking