知识图谱的表示与推理

漆桂林
东南大学认知智能研究所所长

华为 杭州 25-10-2018

Schedule of My Talk

- What is knowledge graph?
- Representation of knowledge graph
- Reasoning with knowledge graph
- Knowledge Update
- Conclusion and Future Work

What is Knowledge Graph?

- The Knowledge Graph is a knowledge base used by Google and its services to enhance its search engine's results with information gathered from a variety of sources.
- Formal definition: a knowledge graph is a knowledge base with graph structure, where the nodes are instances or concepts, and edges are relations between them

Why Do We Care about Knowledge Representation?

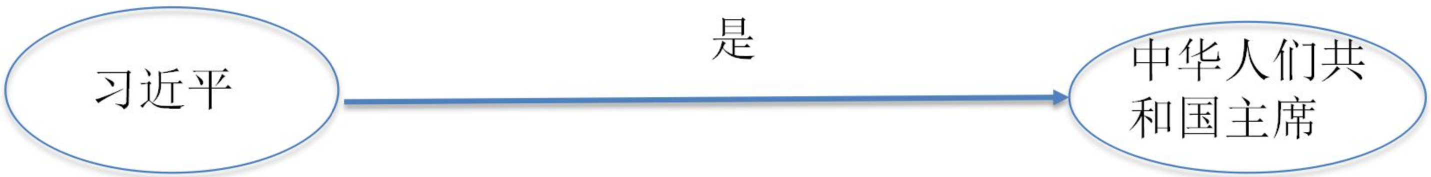
- How do we represent event in knowledge graph?
- How can we represent temporal and spatial relations in knowledge graph?
- How can we represent n-ary relations in knowledge graph?
- What can be the nodes of a knowledge graph?
  - Only instances or classes are allowed?
  - How about literals? How about text ?
- Is the graph representation the best way to represent knowledge?
- Is the knowledge static or dynamic?
- …

Schedule of My Talk

- What is knowledge graph?
- Representation of knowledge graph
- Reasoning with knowledge graph
- Knowledge Update
- Conclusion and Future Work

RDF idea

- Use (directed) graphs as data model



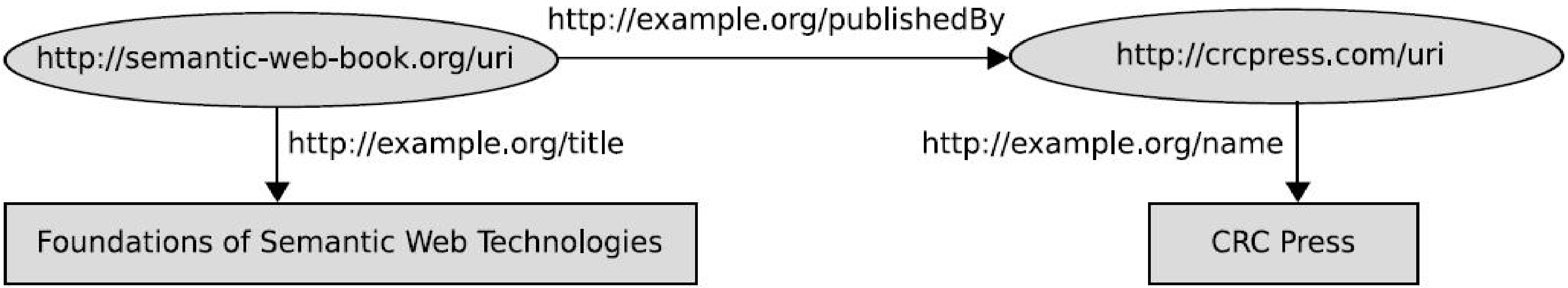
Subject Predicate Object

- Each piece of information is represented as a triple

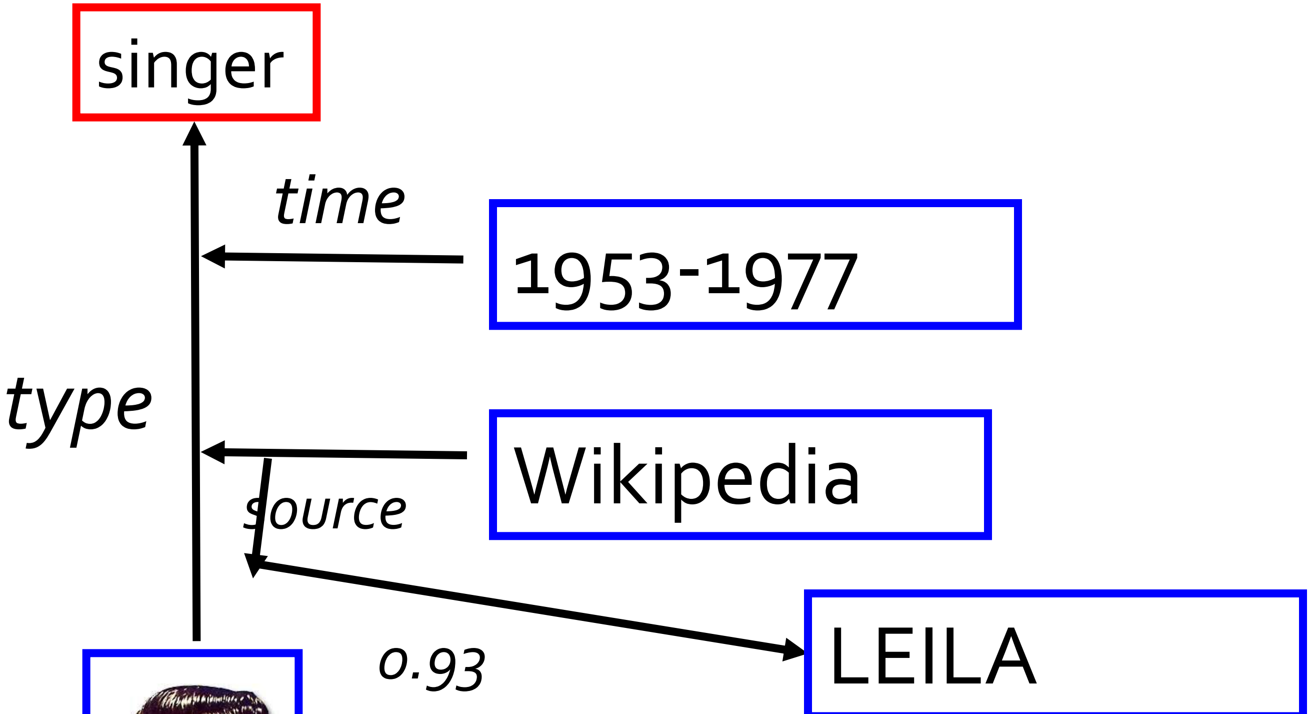
- A set of triple is a knowledge graph

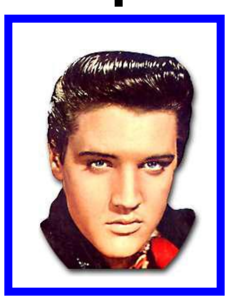
Literals

- for representing data values
- encoded as strings
- interpreted by means of datatypes
- literals without datatype are treated the same as strings



Why Binary is not Enough?

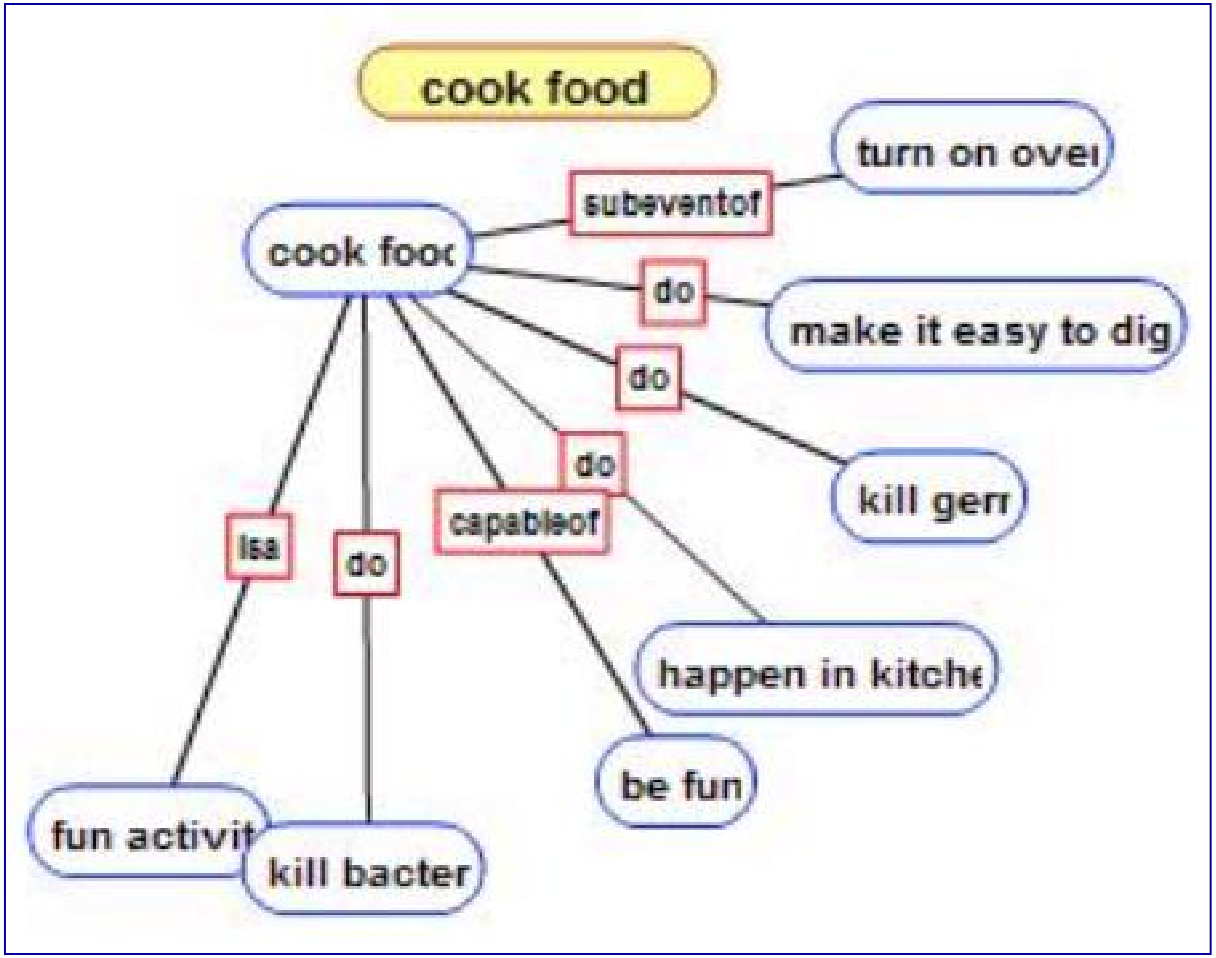




- **`#1 (Elvis, is_a, singer)`**
- **`#2 (#1, time, 1953-1977)`**
- **`#3 (#1, source,Wikipedia)`**

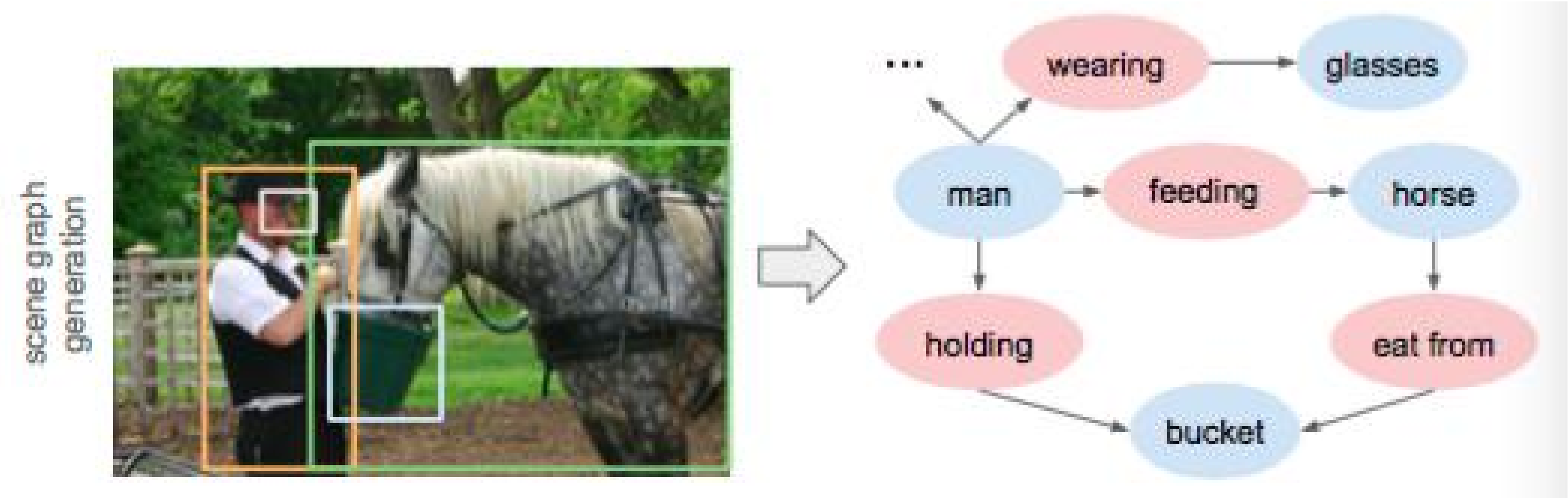
From Yago

Adding text to KG



From ConceptNet

Adding image to KG



Scene Graph

From FeiFei Li

RDFS:Class and Instance

- Classes: sets of instance
- Example：人工智能公司
- Classes can have hierarchy
  - Example: 人工智能公司是高科技公司

人工智能公司 subclass 高科技公司

RDFS:Domain and Range

- Property: a relation over instances
- Example：投资
- Property can domain and range:
  - Domain: 投资人
  - Range:公司

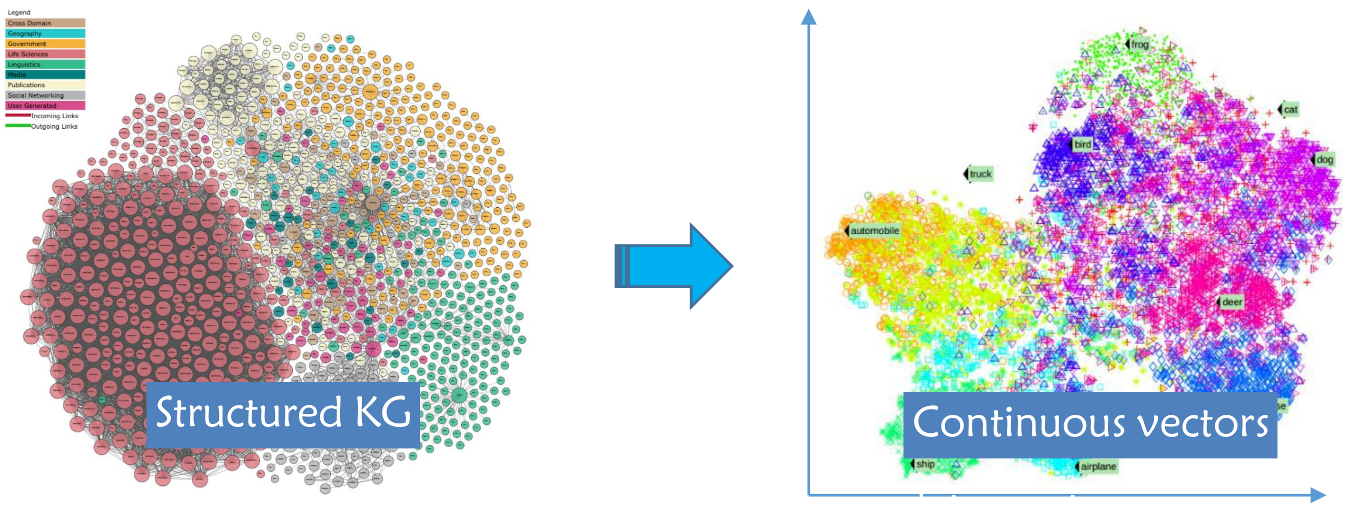
OWL

- Ontology Web Language
- Has description logics as its logical underpinning
- W3C standard ontology language
- Expressive logical language
  - Negation: Car $\sqsubseteq$ $\neg$Train （Disjontness(Car,Train)）
  - Existential restriction:

$$\exists hasChild.Boy \sqsubseteq Parent$$

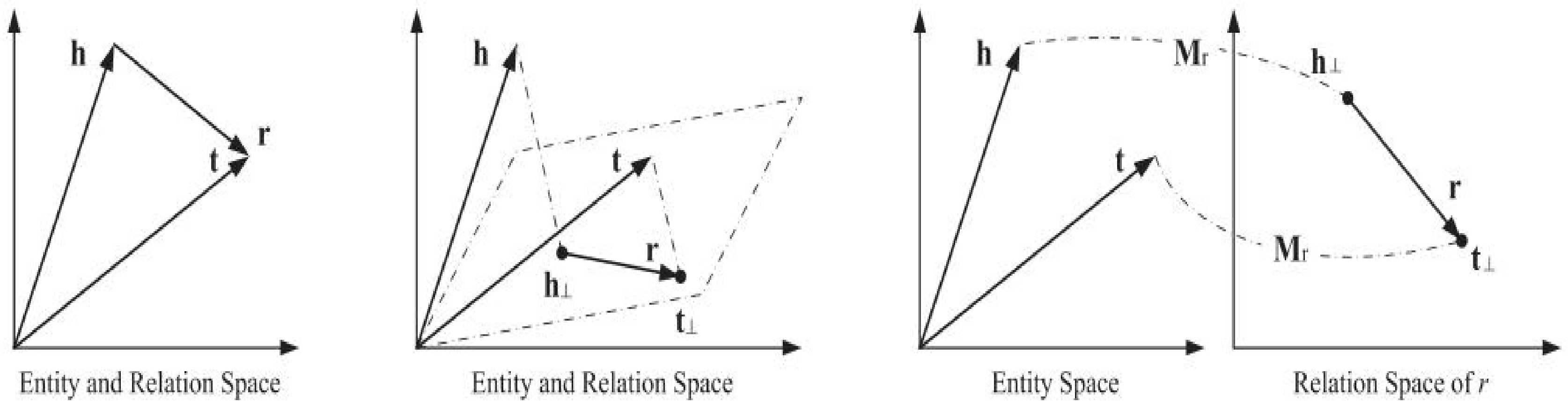
Knowledge Graph Embedding

Knowledge graph (KG) embedding is to embed components of a KG including entities and relations into continuous vector spaces, so as to simplify the manipulation while preserving the inherent structure of the KG.



Knowledge Graph Embedding

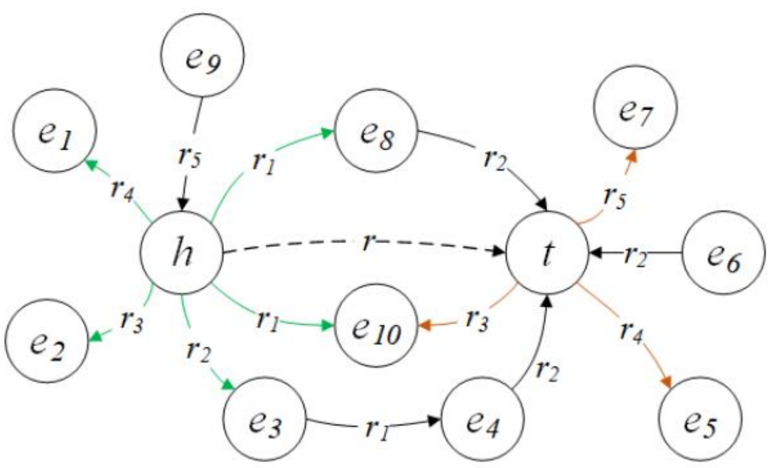
- Translational Distance Models
  - Definition: Translational distance models exploit distance-based scoring functions. They measure the plausibility of a fact as the distance between the two entities, usually after a translation carried out by the relation.



(a) TransE. (b) TransH. (c) TransR.

Context-based Representation Learning

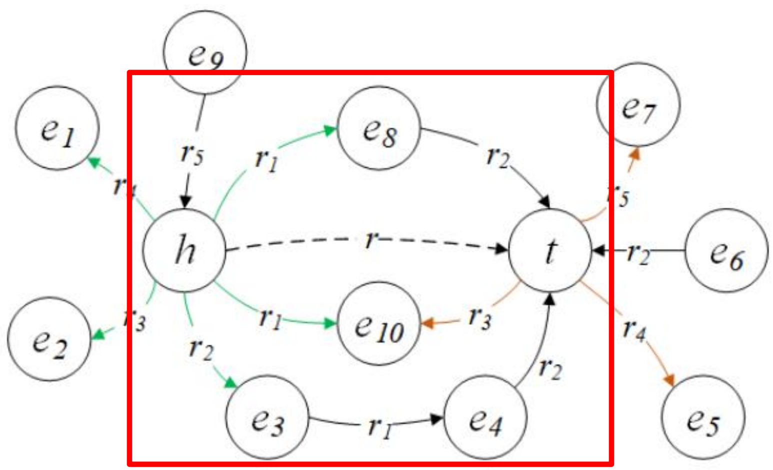
- Knowledge representation learning with structured information
  - Triple Context=Triple + Path Context + Neighbor Context



Jun Shi, Huan Gao, Guilin Qi, Zhangquan Zhou: Knowledge Graph Embedding with Triple Context. **CIKM 2017**: 2299-2302

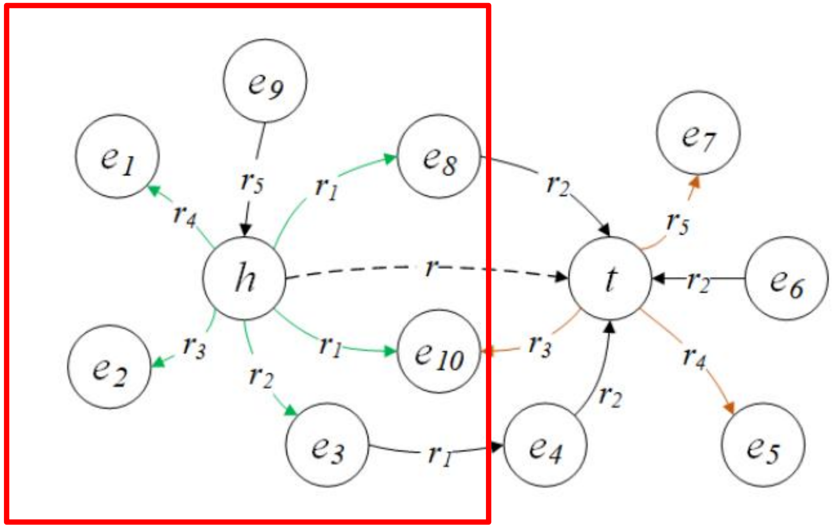
Context-based Representation Learning

- Knowledge representation learning with structured information
  - Triple Context=Triple + Path Context + Neighbor Context



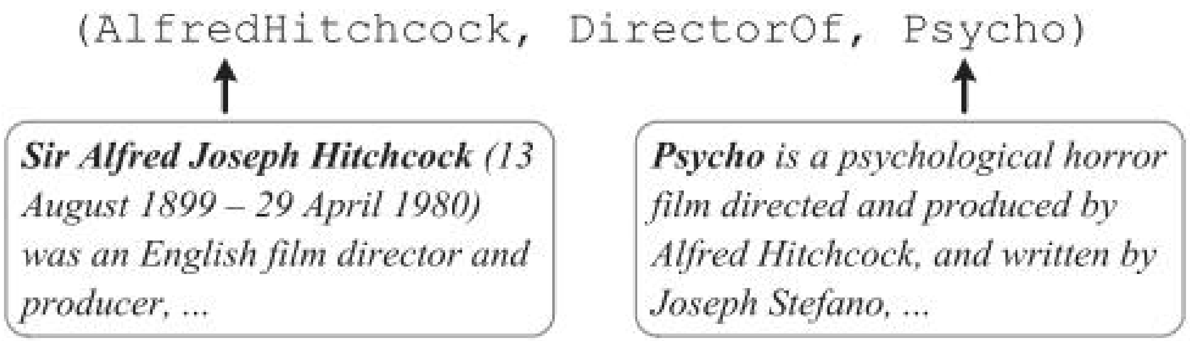
Context-based Representation Learning

- Knowledge representation learning with structured information
  - Triple Context=Triple + Path Context + Neighbor Context



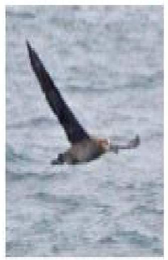
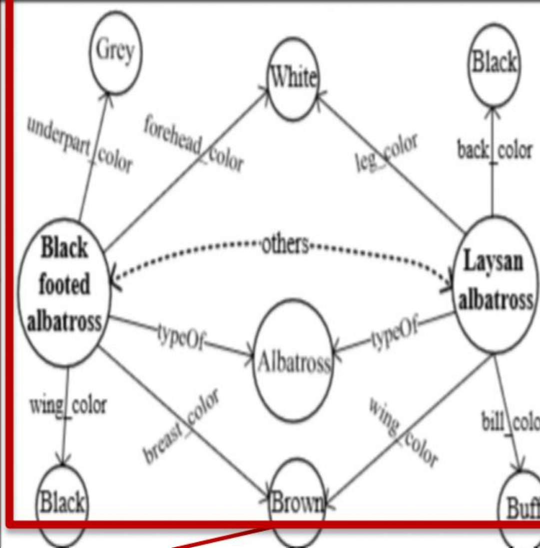
Knowledge Graph Embedding with Side Information

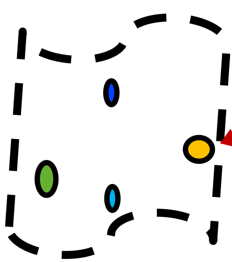
- Textual Descriptions
  - Incorporate descriptions for entities to improve the quality of the Embedding.



- Simple method
  - Textual information is simply used to initialize entity representations.
- Joint method
  - Align the given KG with an auxiliary text corpus, and then jointly conduct KG embedding and word embedding.

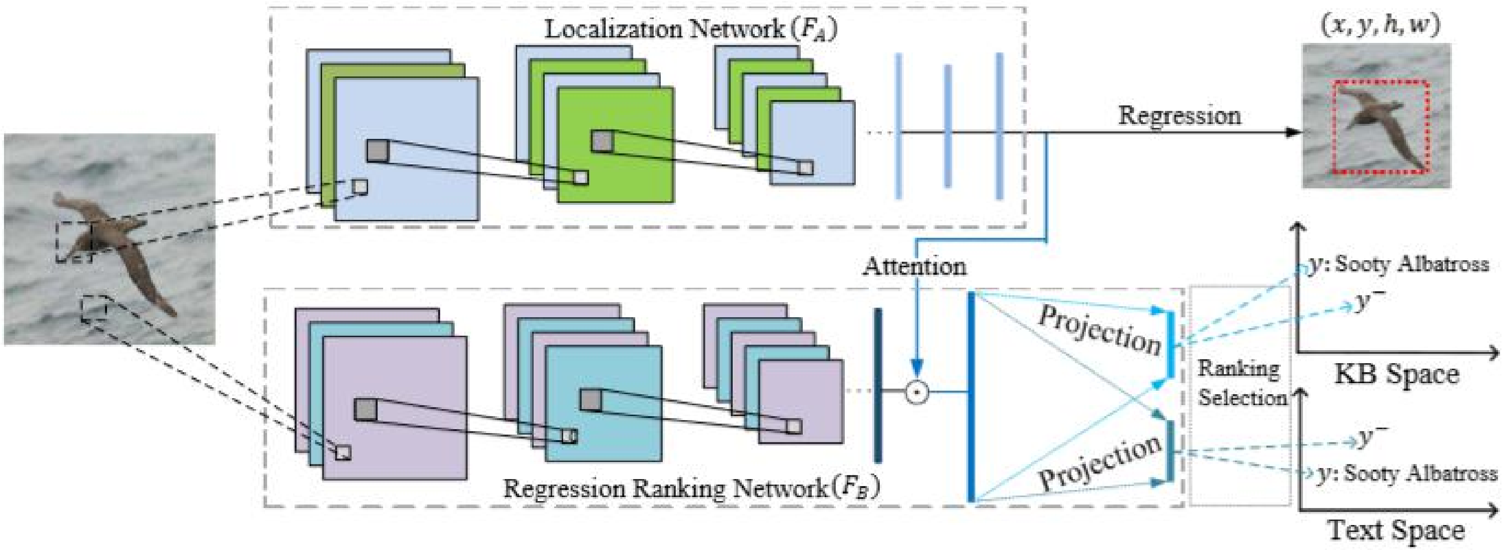
Application of KG in Computer Vision: Multi-Modal

Category	Image	Text	Knowledge base
Black-footed albatross		Black-footed albatross is between two and three feet long with a large wingspan. It has brown to black feathers with white around its eyes and bill and has a large brown bill with a curved tip and black feet.	
Laysan albatross		Laysan albatross has blackish-gray upperwing, back, and tail, and its head, underparts are white. It has a black smudge around the eye, and its underwing pattern with some having narrower black margins	



Embedding space: Encoded prior knowledge into the class vectors

Visual Semantic Embedding



$$f(x,y) = arg_y \max \prod_{i=1}^{2} p(\delta_i(y)|\boldsymbol{x};\boldsymbol{W})$$

A visual semantic embedding model which explores semantic embedding from knowledge bases and text, and further trains a novel end-to-end CNN framework to linearly map image features to a rich semantic embedding space

Huapeng Xu, Guilin Qi et.al Fine-grained Image Classification by Visual-Semantic Embedding. **IJCAI 2018**

Schedule of My Talk

- What is knowledge graph?
- Representation of knowledge graph
- Reasoning with knowledge graph
- Conclusion and Future Work

What is Reasoning in KG?

- Inferring new knowledge from a knowledge graph
- Tasks of reasoning
  - Query relationship between two entities
  - Predict relationship between two entities
  - Query instances of a class
  - Predict instances of a class
  - Query instances that has some relationship with a given instance
  - Checking inconsistency of a KG
  - …
- What methods can be used for reasoning
  - Logical reasoning
  - Statistical reasoning
  - Spreading activation

RDFS:Reasoning

- From

  Google **RDF:type** 人工智能公司

  and

  人工智能公司 **subclass** **高科技公司**

  we can infer

  Google **RDF:type** 高科技公司

RDFS:Reasoning (cont.)

- From

投资 **domain** 投资人
投资 **range** 公司

and

大卫·切瑞 **投资** **Google**

we can infer

大卫·切瑞 **RDF:type** 投资人

Ontology Reasoning-Materialization

- An ontology consists of **axioms** and **assertions**.
- Ontology reasoning is to compute **all implicit statements**.

Example.

Before reasoning

axioms:

$ElementarySchoolChild \sqcap Male \sqsubseteq Boy$

$\exists hasChild.Boy \sqsubseteq Parent$

assertions:

ElementarySchoolChild (Tom)

Male(Tom)

hasChild(Jack, Tom)

reasoning →

After reasoning

$ElementarySchoolChild \sqcap Male \sqsubseteq Boy$

$\exists hasChild.Boy \sqsubseteq Parent$

ElementarySchoolChild (Tom)

Male(Tom)

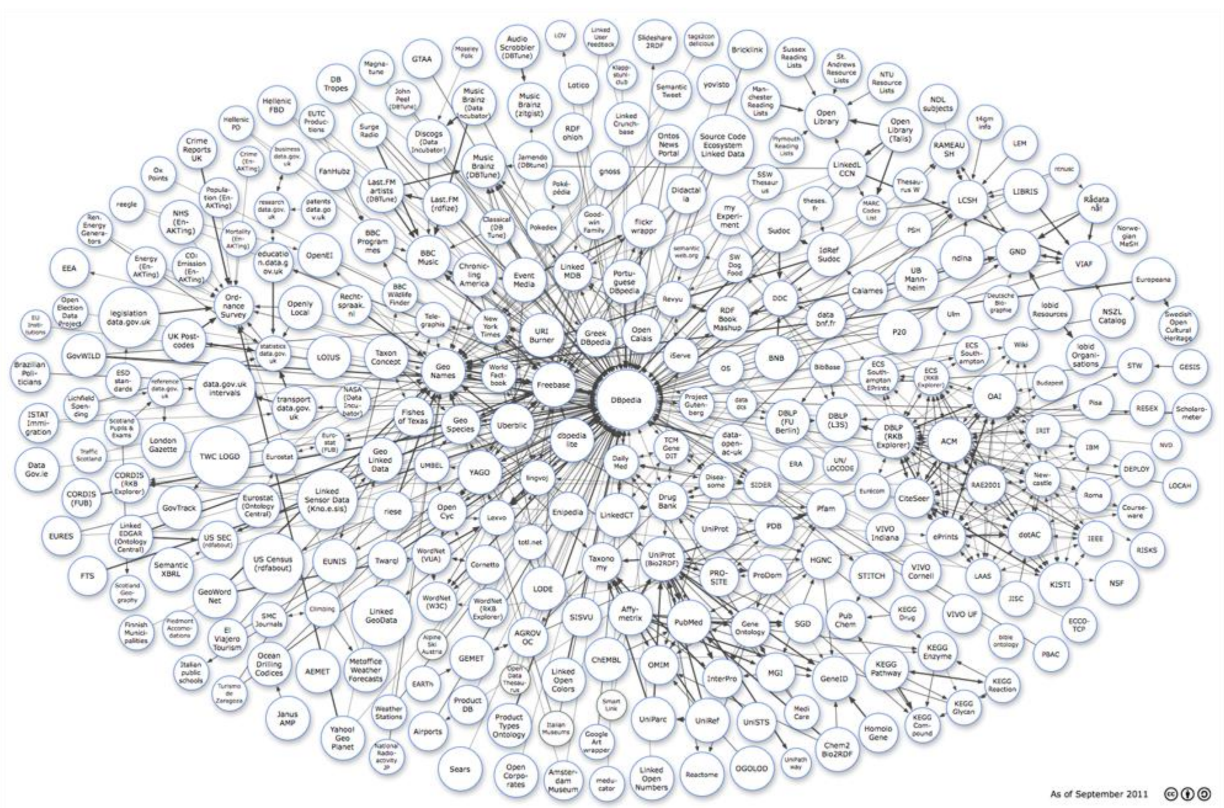
hasChild(Jack, Tom)

Boy(Tom)

Parent(Jack)

Ontology Reasoning

- In practice, we have to face the problem of **large-scale ontology reasoning**.



Ontology Reasoning

What we do for large-scale reasoning**.**

- Current works focuses on devising efficient algorithms. However, we find that the performance of large-scale reasoning depends on **the structure of datasets** to a large extent.
- We aim to **identify classes** of ontologies, for which ontology reasoning is **in NC complexity**.

NC problems are **efficient for parallel** implementation and **large-scale processing**.

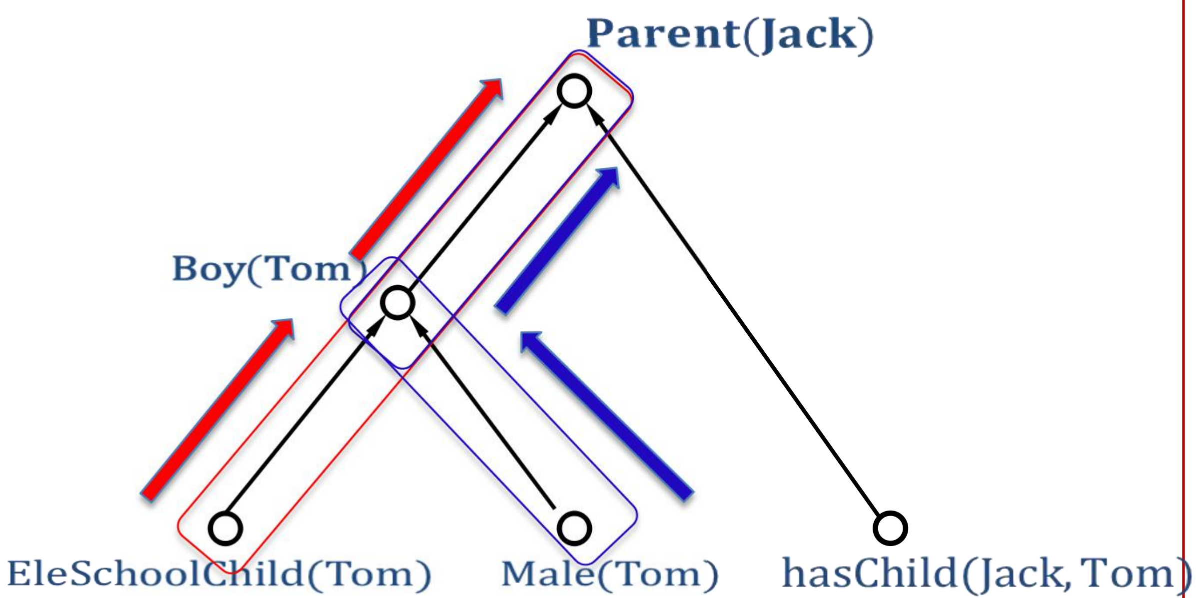
Zhangquan Zhou, Guilin Qi, Birte Glimm: Exploring Parallel Tractability of Ontology Materialization. **ECAI 2016**: 73-81

Ontology Reasoning

Critical Paths in Reasoning Graph

ElementarySchoolChild(x), Male(x) → Boy(x).
hasChild(x, y), Boy(y) →Parent(x).

ElementarySchoolChild (Tom)
Male(Tom)
hasChild(Jack, Tom)
Boy(Tom)
Parent(Jack)



- We proved that the performance of reasoning depends on **the critical paths** (the longest paths) in reasoning graphs when using parallel techniques.
- If the length of critical paths is bounded with poly-log, then ontology reasoning is in NC complexity.
- **Optimization**: identify the critical paths, and process them beforehand.

Ontology Reasoning

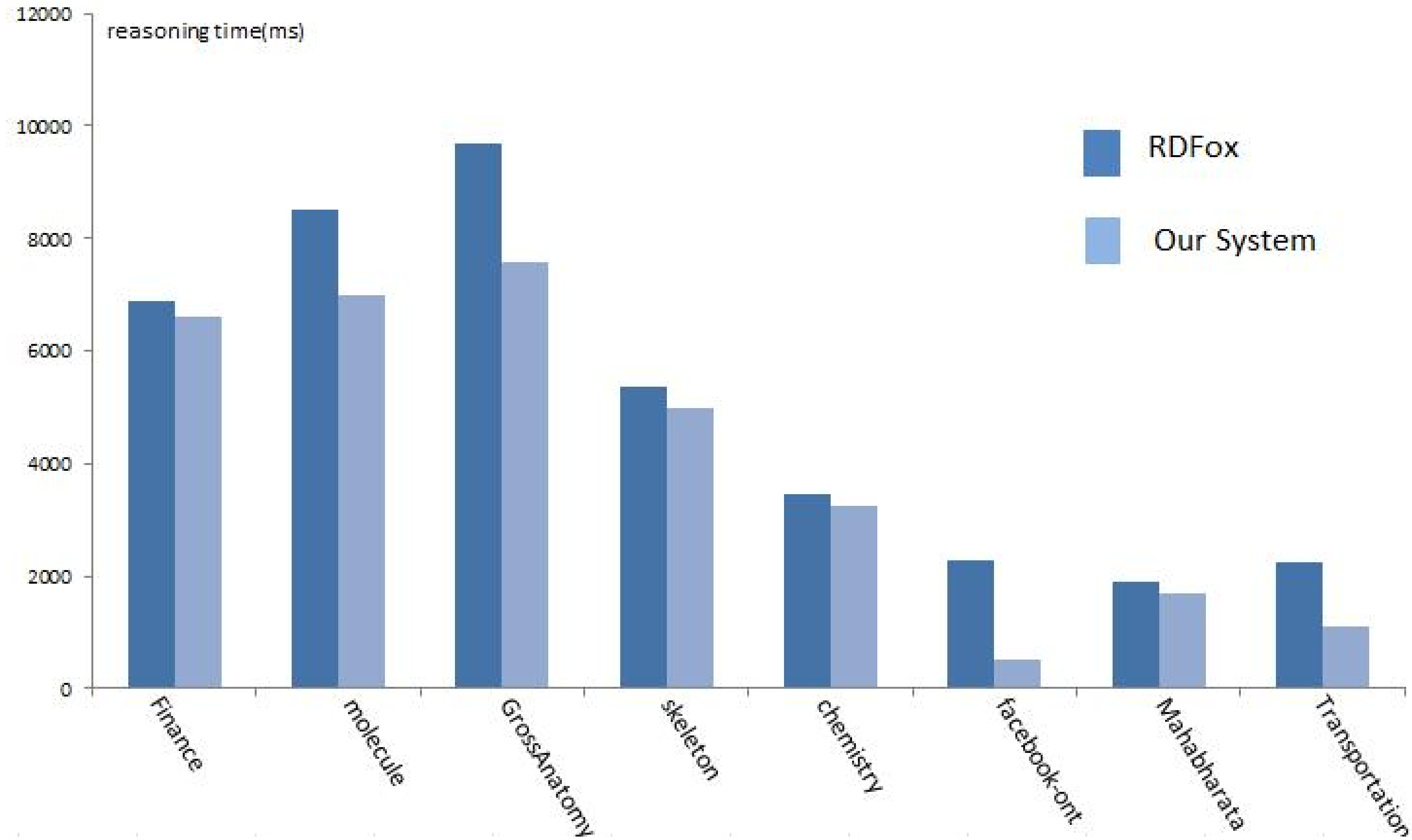
- We have **evaluated** many **real datasets** and **benchmarks**. We find that many of them have critical paths in their reasoning graphs with poly-log bounded lengths.

- One can use our results to optimize large-scale reasoning for the following real ontologies.

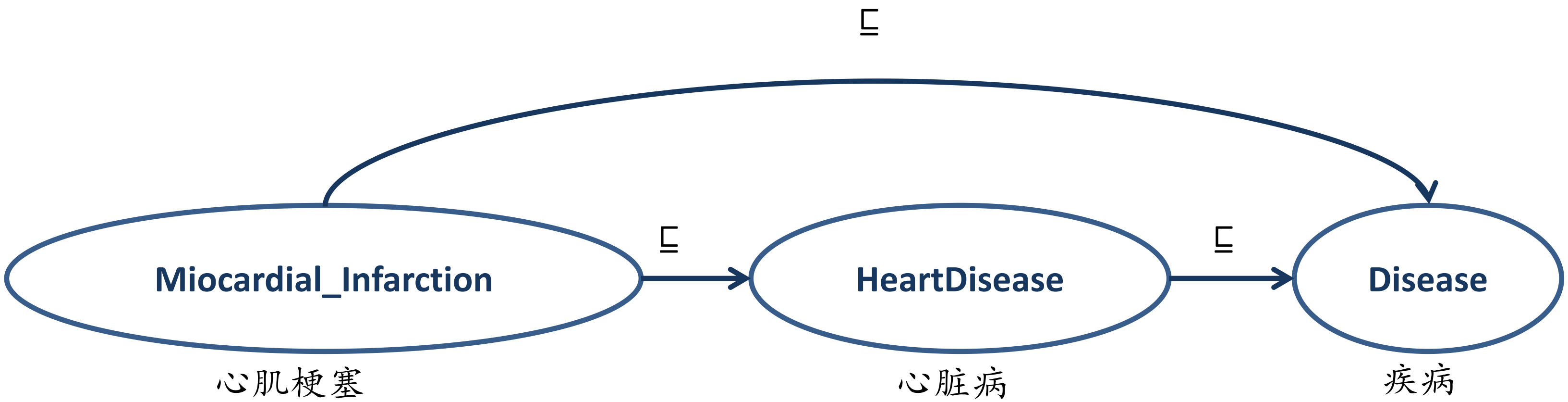
Benchmarks	**SIB** (*Social Network Intelligence BenchMark*), **BSBM** (*Berlin SPARQL Benchmark*) **LODIB** (*Linked Open Data Integration Benchmark*) **IIMB** (*The ISLab InstanceMatching Benchmark*)
YAGO	**follows our results.**
Real ontologies	We **investigated 151 ontologies** that are collected from the Protege ontology library, Swoogle and Oxford ontology lib. Among these ontologies, **111** of them are **parallelly tractable**, i.e., in NC complexity.

Ontology Reasoning

- We implement an optimized system based on our method, and compare it to the currently fastest reasoner RDFox, our system is **averagely 20% faster** then RDFox over the following real large-scale ontologies.



Ontology Reasoning-Classification



Graph Representation of OWL EL Ontology

Why using graph representation?

1) The syntax of OWL EL is specific, and not fit for parallelism;
2) Graph is a popular platform-independent model in database and parallel data processing (Pregel, Neo4j).

What to do?

Translate an OWL EL ontology to a graph structure, and perform classification on the translated graph.

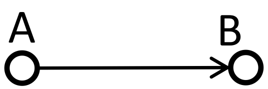
Challenges:

1) Represent ternary relationships:

   binary relation: $A \sqsubseteq B$

   ternary relation: $A \sqsubseteq \exists r.B$

   Conceptual Graph (complicated!!!)



2) The graph operations should be easily scaled.

Graph Representation

Our method:

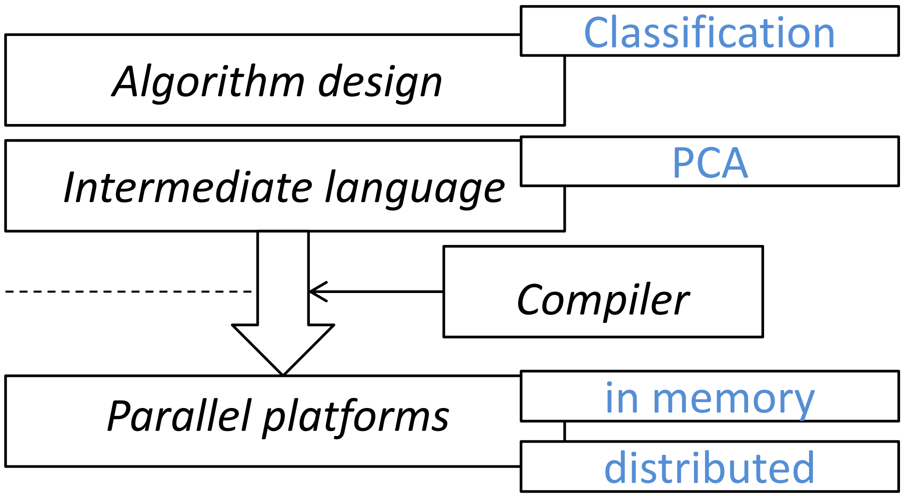
Encode some information into the edge.

Specifically:

$A \sqsubseteq B$	$A \xrightarrow{\sqsubseteq} B$	$\exists r.A \sqsubseteq B$	$A \xrightarrow{-r} B$
$A_1 \sqcap A_2 \sqsubseteq C$	$A_1 \xrightarrow{\sqcap A_2} C$	$r \sqsubseteq s$	$r \xrightarrow{\sqsubseteq} s$
$A \sqsubseteq \exists r.B$	$A \xrightarrow{+r} B$	$r \circ s \sqsubseteq t$	$r \xrightarrow{\circ s} t$

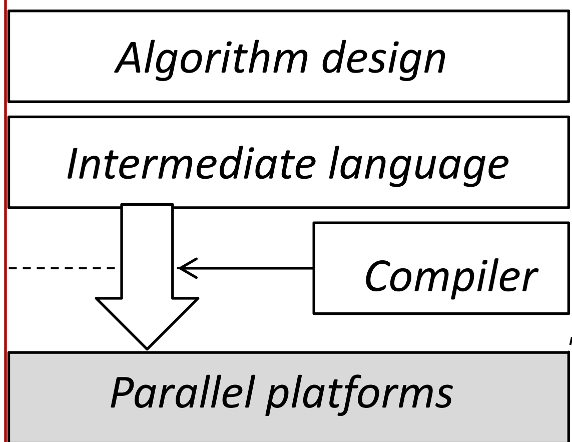
Theorem: *the graph representation is model-equivalent to OWL EL syntax.*

A Platform-independent Model



Z. Zhou, G. Qi, et.al. A Platform-independent Approach for Parallel Reasoning with OWL EL Ontologies using GraphRepresentation， 2015 IEEE 27th International Conference on Tools with Artificial Intelligence,pp 80-87, 2015. (Best student paper award)

Evaluation



Platforms:

1) centralized platform:

A SuperCloud server with a 128 Gigabyte memory and 12 physical cores.

2) A distributed cluster consisting of 5 nodes and each one of them is an economic Lenovo ThinkCentre machine with a 2 Gigabyte RAM and two physical cores

Datasets:

	#node	#edge
GO	49,375	125,336
Galen	35,989	83,948
NCI	27,933	47,360
SCT	464,601	887,846
SGG	511,511	982,237

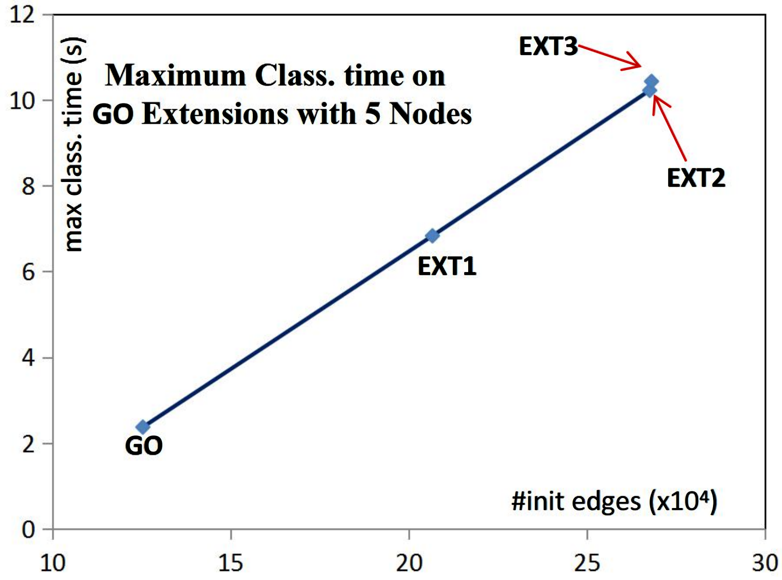
Evaluation

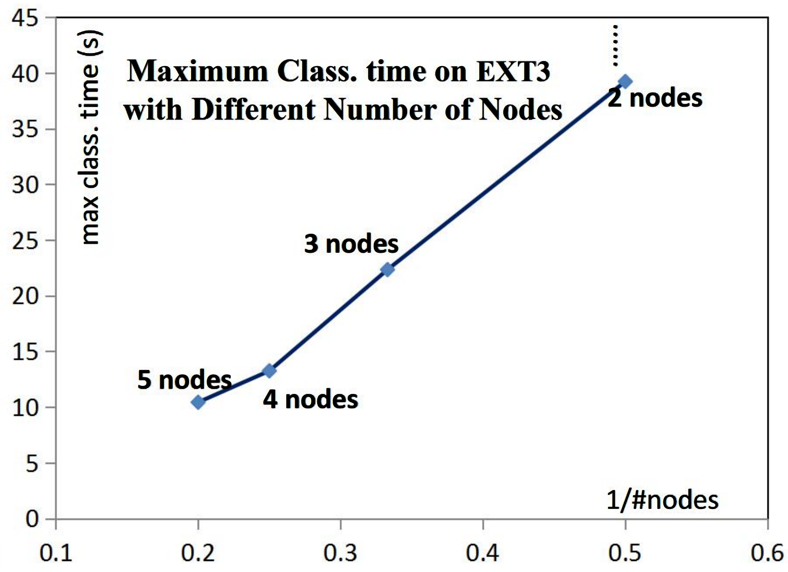
Evaluation on a multi-core system: Compare our system with ELK and jCEL.

Time/s	#thread	GO	Galen	NCI	SCT	SGG
jCEL	1	20.22	33.25	1.43	676.75	776.93
ELK	1	3.96	3.27	2.07	25.42	33.81
	5	4.16	2.42	1.92	15.93	18.83
	10	3.55	3.01	2.19	14.99	18.47
	15	3.94	2.98	2.08	13.86	18.57
	20	3.27	2.87	2.14	15.27	18.79
GEL (our system)	1	8.69	5.87	2.58	34.57	43.22
	5	2.84	2.13	1.69	16.55	17.93
	10	2.44	1.88	1.31	15.46	16.42
	15	1.87	1.74	1.30	14.33	14.66
	20	1.53	1.74	1.27	14.58	13.75

Evaluation

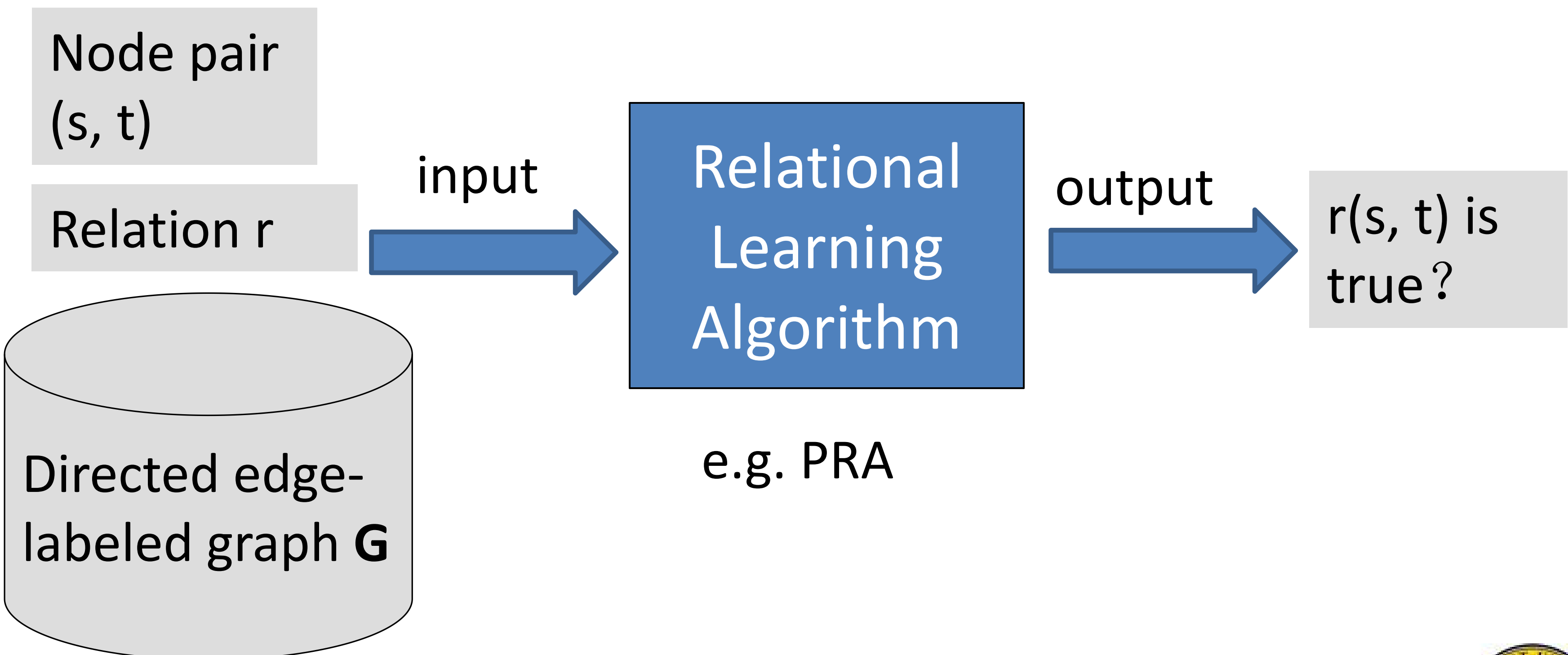
Evaluation on a distributed cluster: A distributed cluster consisting of 5 nodes and each one of them is an economic Lenovo ThinkCentre machine with a 2 Gigabyte RAM and two physical cores.



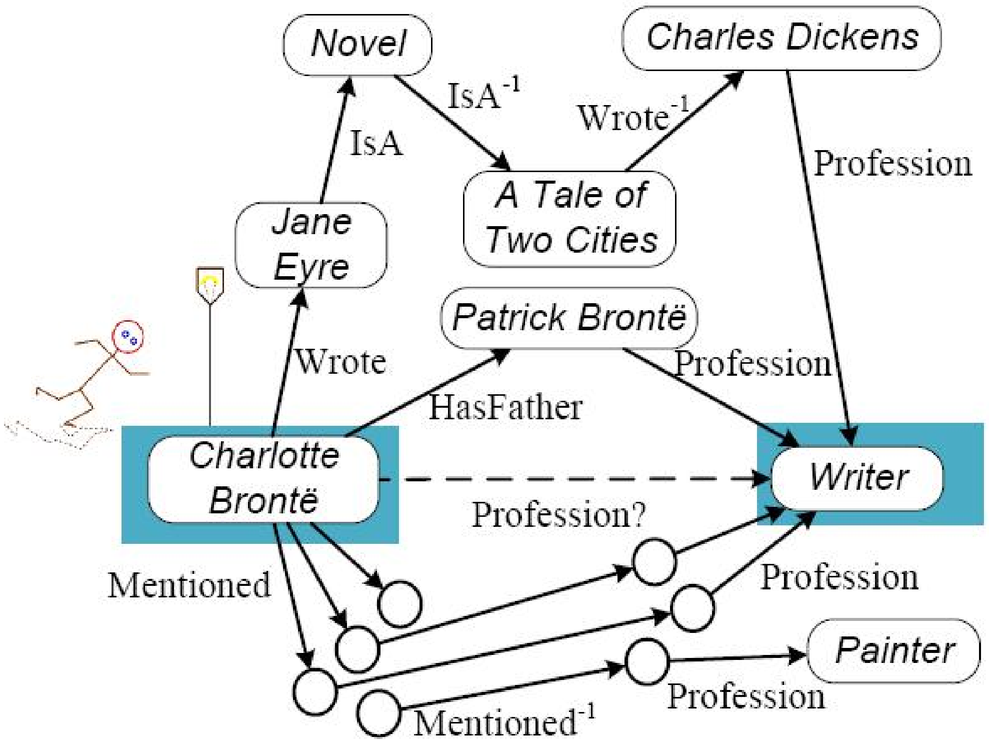


Relational Learning

- The generic relational learning task



Path Ranking Algorithm (PRA)



Lao et.al.

G=(N,E, R)

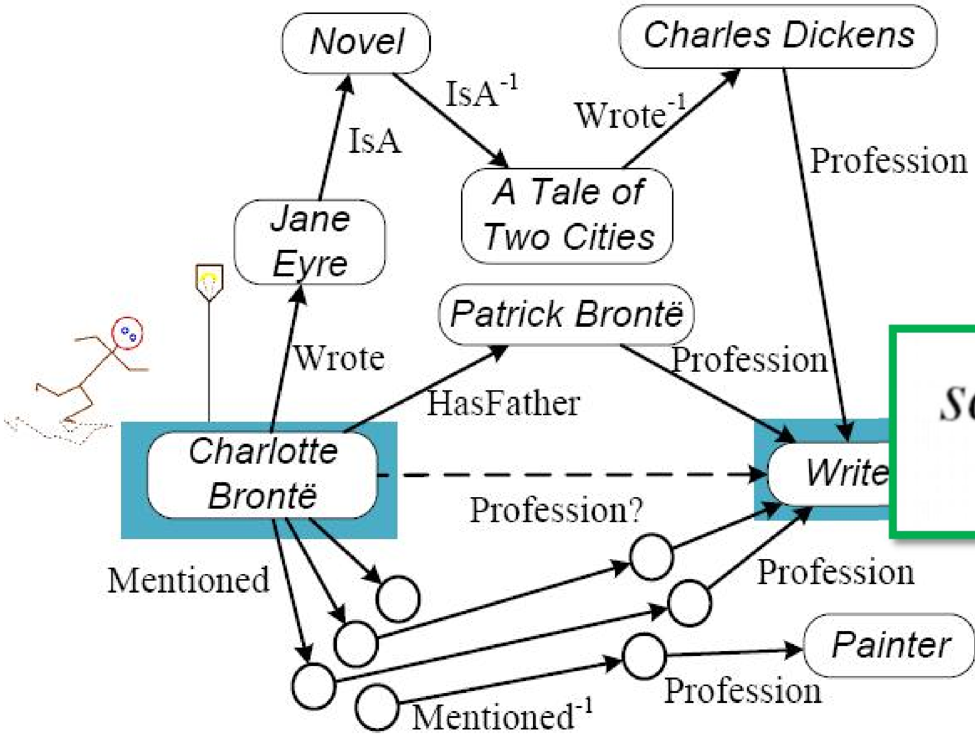
- N: nodes (instances or concepts)
- E: edges
- R: edge types

Note: r^{-1}: reverse of edge type r

Path type π: $<r_1,r_2,\ldots,r_n>$

e.g. <HasFather, Profession>

Path Ranking Algorithm (PRA)



Profession(Charlotte Bonte, Writer)?

$$score(s,t) = \sum_{\pi \in Q} P(s \to t; \pi)\theta_\pi$$

Q: all path types starting from s and ending with t (with length of n)

θ_π: weights obtained by training

Rules

类型	支持的规则形式	举例
0型.	*If…then…* 或者 *If…normally then…*	百度的研发岗位工资高： *if* 就职(?X, 百度), 岗位(?X, 研发) *then* 收入(?X, 高) 一般来说，头痛发烧是感冒： *if* 症状(?X, 头痛), 症状(?X, 发烧) *normally then* 患病(?X, 感冒)
1型.	[0.6] *If…then…* 或者 [0.8] *If…normally then…*	咳痰带血是支气管扩张的重要症状： [0.6] *If* 患病(?X,支气管扩张)，患者(?X) *then* 症状(X,咳痰带血) 一般来说，咯血是肺癌的关键症状： [0.8] *If* 患病(?X,支气管扩张)，患者(?X) *normally then* 症状(X,咯血)
2型.	规则+本体	规则：一般来说，头痛发烧是感冒： *if* 症状(?X, 头痛), 症状(?X, 发烧) *normally then* 患病(?X, 感冒) 本体：感冒是一种上呼吸道疾病： 感冒⊂上呼吸道疾病
3型.	认知规则 K	右下腹突发剧痛，可能是急性盲肠炎或者肾结石： if 症状(?X, 右下腹剧痛) then 患病(?X, 盲肠炎) 或者 患病(?X, 肾结石) 如果盲肠炎确诊无误，则实施切除手术： if K 患病(?X, 盲肠炎 then 手术切除(?X, 盲肠)

Performance

类型	支持的规则形式	目前性能
0型.	*If…then…* 或者 *If…normally then…*	万级规则，毫秒级速度
1型.	[0.5] *If…then…* 或者 [0.8] *If…normally then…*	万级规则，毫秒级速度
2型.	规则+本体	万级规则，十万级本体，秒级速度
3型.	认知规则	万级规则，毫秒级速度

Applications

类型	应用场景
0型.	• 疾病诊断知识表示和推理
1型.	• 国家863项目，高考机器人，地理知识表示和推理 • **态势综合分析 • 多模态人机交互中的交互知识表示和推理
2型.	• 国家863项目，高考机器人，地理知识表示和推理 • 疾病诊断知识表示和推理
3型.	• 多模态人机交互中的交互知识表示和推理

Application Case

问题及知识的自然语言描述	规则	本体（RDF）
咳嗽、咳痰或原有呼吸道疾病症状加重一般是社区获得性肺炎诊断的重要依据。 John咳嗽较之前更重。 问：John是否有可能得了呼吸系统疾病	If NULL then 病症(John, 咳嗽). If NULL then 症状加重(John, 咳嗽). [0.8] If 患者(?X), 症状加重(?X, ?呼吸道疾病症状) normally then 患病(?X,社区获得性肺炎) ……	(咳嗽, is_a,呼吸道疾病症状) …… (呼吸道疾病症状, is_a, 症状) …… (咳嗽, 相关症状,社区获得性肺炎) …… (社区获得性肺炎, subclass_of, 呼吸系统疾病)

第一步：利用本体得到“John有呼吸道疾病症状加重”

第二步：利用规则推理得到“John有0.8的可能患社区获得性肺炎”

第三步：利用本体得到“John有0.8可能患呼吸系统疾病”

Schedule of My Talk

- What is knowledge graph?
- Representation of knowledge graph
- Reasoning with knowledge graph
- Knowledge Update
- Conclusion and Future Work

Background of KG update

1. New entities appear rapidly
2. Knowledge relevant to entities changes quickly
   – a large number of outdated knowledge will be produced without update

<Donald Trump, profession, businessman>

New facts

<Donald Trump, profession, businessman>

<Donald Trump, profession, politicians>

<USA, president, Barack Obama>

Changed facts

<USA, president, Donald Trump>

Motivation

Periodical Update

- replacing the knowledge base with a new version once after a period of time

1. If the update cycle is too long, there will be a large amount of outdated knowledge existing in KB.

2. If the update cycle is too short, there will be a huge consumption of bandwidth and computing resources

Goals of Dynamic Update

Update KB with a higher frequency and lower consumption

a. Most entities in KB do not need to be updated
   - No need to check every entity for updating

b. In each update cycle, select entities which are most likely to change to check for update.

Framework

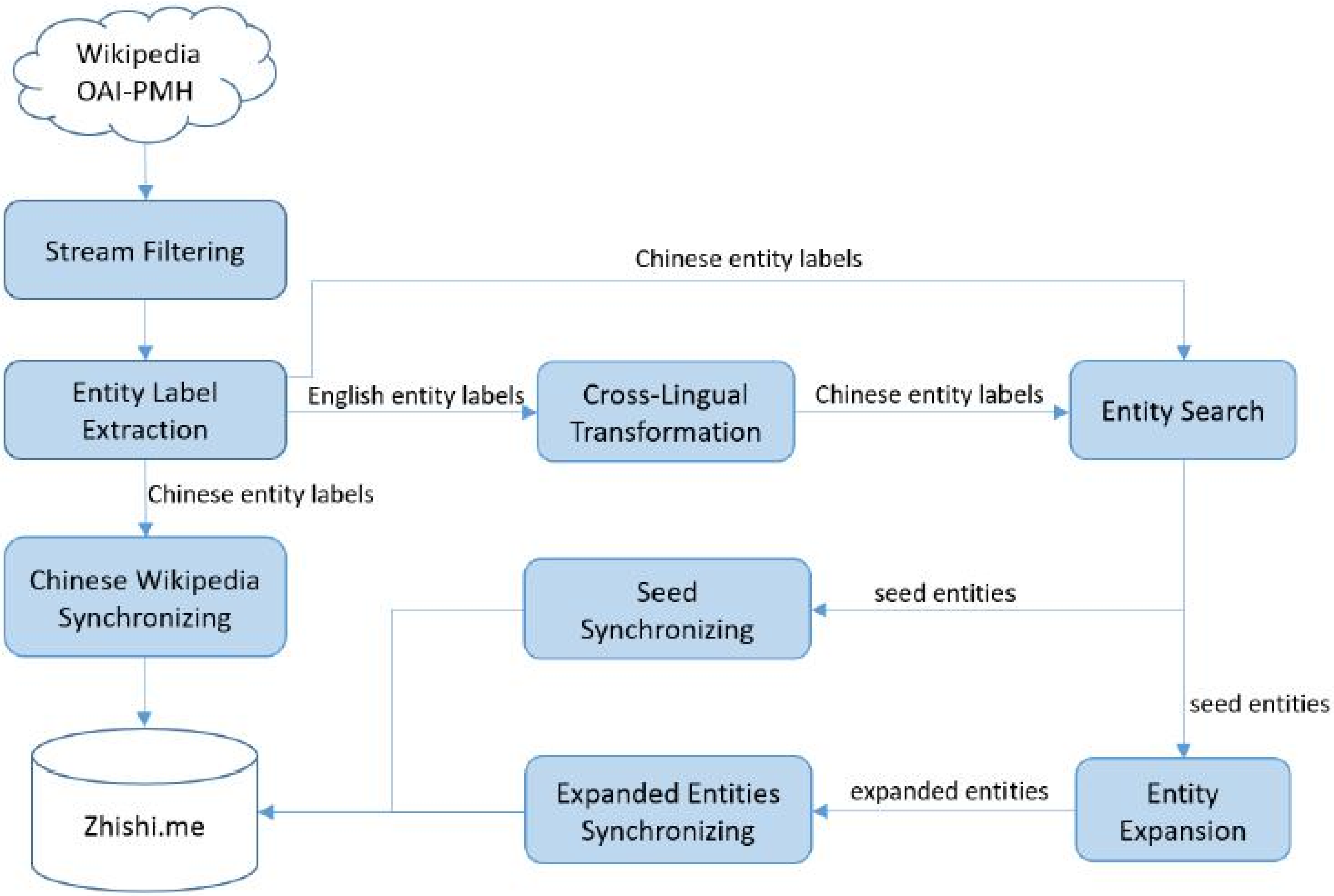


Fig. 2 Overview of Zhishi.me 2.0 Live Extraction Framework

Schedule of My Talk

- What is knowledge graph?
- Representation of knowledge graph
- Reasoning with knowledge graph
- Conclusion and Future Work

Conclusion

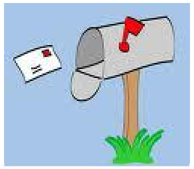
- Representation of KG
  - Has foundation on ontology languages
  - But is not confined to them
- Logical reasoning with KGs
  - Parallel reasoning can achieve good performance
  - Non-standard reasoning can be done efficiently
- Statistical reasoning with KGs
  - Path-based reasoning is practical
  - Knowledge representation learning has potential to be useful
- Applications
  - Question answering
  - Decision making
  - …

Future Work

- We need a new representation language that has the following features
  - Has logical foundation
  - Has different levels of abstraction
  - Can do uncertainty reasoning, like Bayesian network
  - Can be extended to support temporal and spatial reasoning
  - Has network-like structure
  - Reasoning can be easily computed in parallel

Thank You

Question?



gqi@seu.edu.cn