Unsupervised Discovery of Corroborative Paths for Fact Validation

Zafar Habeeb Syed, Michael Röder, Axel-Cyrille Ngonga Ngomo

zsyed@mail.uni-paderborn.de, michael.roeder|axel.ngonga@upb.de

28 October, 2019

Motivation

- **Knowledge graphs** [1,2]

Motivation

- **Knowledge graphs** [1,2]
  - Facts used in real world applications

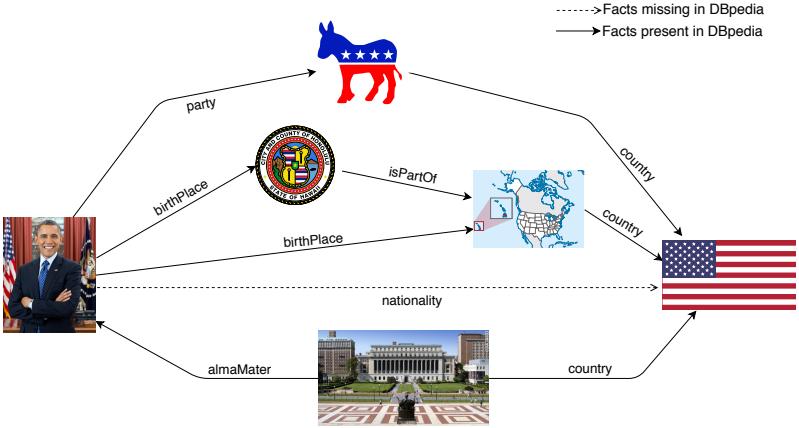
Motivation

- **Knowledge graphs** [1,2]
  - Facts used in real world applications
  - Drawbacks: incomplete, incorrect facts

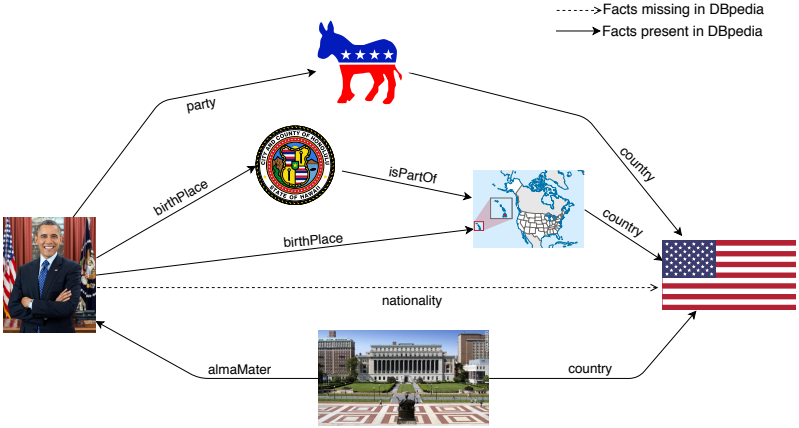
Motivation

- **Knowledge graphs** [1,2]
  - Facts used in real world applications
  - Drawbacks: incomplete, incorrect facts
  - Ensuring veracity

Motivation


Facts missing in DBpedia
Facts present in DBpedia
party
birthPlace
isPartOf
country
country
birthPlace
nationality
almaMater
country

Motivation



- **Idea:** Predicates carry mutual information with a set of other paths

- **Problem statement**: Given a triple $t = (s, p, o)$ and knowledge graph $\mathcal{G}$, compute the likelihood that t is true

Problem Statement

- **Problem statement**: Given a triple $t = (s, p, o)$ and knowledge graph $\mathcal{G}$, compute the likelihood that t is true
- **Knowledge graph:** An RDF knowledge graph $\mathcal{G}$ is a set of RDF triples, i.e.,

$$\mathcal{G} = \{(s, p, o) | s \in \mathbb{E} \cup \mathbb{B}, p \in \mathbb{P}, o \in \mathbb{E} \cup \mathbb{B} \cup \mathbb{L}\}$$

Directed path: $\pi^k(v_0, v_k) \rightarrow (v_0, p_1, v_1), (v_1, p_2, v_2), ..., (v_{k-1}, p_k, v_k)$

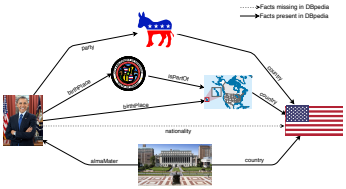
- $\forall i \in [0, k-1]\ (v_i, p_{i+1}, v_{i+1}) \in \mathcal{G}$
- $\forall i, j \in [0, k], i \neq j \rightarrow v_i \neq v_j$

Directed path: $\pi^k(v_0, v_k) \rightarrow (v_0, p_1, v_1), (v_1, p_2, v_2), ..., (v_{k-1}, p_k, v_k)$

- $\forall i \in [0, k-1]\ (v_i, p_{i+1}, v_{i+1}) \in \mathcal{G}$
- $\forall i, j \in [0, k], i \neq j \rightarrow v_i \neq v_j$

Undirected path: $\mu^k(v_0, v_k) \rightarrow (v_0, p_1, v_1), (v_1, p_2, v_2), ..., (v_{k-1}, p_k, v_k)$

- $\forall i \in [0, k-1]\ (v_i, p_{i+1}, v_{i+1}) \in \mathcal{G} \vee (v_{i+1}, p_{i+1}, v_i) \in \mathcal{G}$
- $\forall i, j \in [0, k], i \neq j \rightarrow v_i \neq v_j$



Example (Directed path)

Directed path: BarackObama $\xrightarrow{\texttt{birthPlace}}$ Hawaii $\xrightarrow{\texttt{country}}$ USA

Example (Undirected path)

Undirected path: BarackObama $\xleftarrow{\texttt{alumni}}$ GreenwichHighSchool $\xrightarrow{\texttt{country}}$ USA

- **Typed path:** Given set of types t_x and t_y, typed paths are given by

$$\Pi^k_{(t_x,t_y)} = \{\pi^k(v_0, v_k) \mid \gamma(v_0) \sqsubseteq t_x \wedge \gamma(v_k) \sqsubseteq t_y\}$$

- **Typed path:** Given set of types t_x and t_y, typed paths are given by

$$\Pi^k_{(t_x,t_y)} = \{\pi^k(v_0, v_k) \mid \gamma(v_0) \sqsubseteq t_x \wedge \gamma(v_k) \sqsubseteq t_y\}$$

Example (Typed path)

$t_x = \{\texttt{Person}\}$, $t_y = \{\texttt{Place}\}$

$$\Pi^2_{(t_x,t_y)} : \texttt{BarackObama} \xrightarrow{\texttt{party}} \texttt{DemocraticParty} \xrightarrow{\texttt{country}} \texttt{USA}$$

- **$\vec{q}$-restricted typed path:** $\vec{q} = q_1, \ldots, q_k$

$$\Pi^k_{(t_x,t_y),\vec{q}} = \{\pi^k(v_0, v_k) \mid \pi^k(v_0, v_k) \in \Pi^k_{(t_x,t_y)}, \\ \forall i \in [0, k-1] : (v_i, p_{i+1}, v_{i+1}) \in \pi^k(v_0, v_k) \rightarrow p_{i+1} = q_{i+1}\}$$

- **$\vec{q}$-restricted typed path:** $\vec{q} = q_1, \ldots, q_k$

$$\Pi^k_{(t_x,t_y),\vec{q}} = \{\pi^k(v_0, v_k) \mid \pi^k(v_0, v_k) \in \Pi^k_{(t_x,t_y)}, \\ \forall i \in [0, k-1] : (v_i, p_{i+1}, v_{i+1}) \in \pi^k(v_0, v_k) \rightarrow p_{i+1} = q_{i+1}\}$$

Example ($\vec{q}$-restricted typed path)

$t_x = \{\texttt{Person}\}$ and $t_y = \{\texttt{Place}\}$ and $\vec{q} = (\ \texttt{birthPlace}, \texttt{country})$,

$$\Pi^2_{(t_x,t_y)} : \texttt{BarackObama} \xrightarrow{\texttt{birthPlace}} \texttt{Hawaii} \xrightarrow{\texttt{country}} \texttt{USA}$$

- Given $t = (s, p, o)$: *Only typed paths in* $\Pi^k_{(D(p),R(p)),\vec{q}}$ *can corroborate facts with the predicate* p

- Given $t = (s, p, o)$: *Only typed paths in* $\Pi^k_{(D(p),R(p)),\vec{q}}$ *can corroborate facts with the predicate* p
- *Corroborative paths* for a predicate p

$$\Pi^k(p) = \bigcup_{j=1}^{k} \Pi^j_{(D(p),R(p)),\vec{q}}$$

- Association strength: NPMI [3]

- Association strength: NPMI [3]

$$\text{NPMI}(\Pi^{j}_{(t_x,t_y),\vec{q}}, p) = \frac{\log\left(\frac{\mathcal{P}\left(\Pi^{j}_{(t_x,t_y),\vec{q}},p\right)}{\mathcal{P}\left(\Pi^{j}_{(t_x,t_y)}\right)\cdot\mathcal{P}(p)}\right)}{-\log\left(\mathcal{P}\left(\Pi^{j}_{(t_x,t_y),\vec{q}},p\right)\right)}$$

- Association strength: NPMI [3]

$$\text{NPMI}(\Pi^{j}_{(t_x,t_y),\vec{q}}, p) = \frac{\log\left(\frac{\mathcal{P}\left(\Pi^{j}_{(t_x,t_y),\vec{q}},p\right)}{\mathcal{P}\left(\Pi^{j}_{(t_x,t_y)}\right)\cdot\mathcal{P}(p)}\right)}{-\log\left(\mathcal{P}\left(\Pi^{j}_{(t_x,t_y),\vec{q}}, p\right)\right)}$$

- Approximation to NPMI

$$\widehat{\text{NPMI}}(\Pi^j_{(t_x,t_y),\vec{q}}, p) = \frac{\log\left(\frac{\widehat{\mathcal{P}}\left(\Pi^j_{(t_x,t_y),\vec{q}},p\right)}{\widehat{\mathcal{P}}\left(\Pi^j_{(t_x,t_y)}\right)\cdot\mathcal{P}(p)}\right)}{-\log\left(\widehat{\mathcal{P}}\left(\Pi^j_{(t_x,t_y),\vec{q}}, p\right)\right)}$$

$$\widehat{\mathcal{P}}(\Pi^j_{(t_x,t_y),\vec{q}}) = \frac{|\Pi^j_{(t_x,t_y),\vec{q}}|}{|\lambda(t_x)|\cdot|\lambda(t_y)|}$$

$$\widehat{\mathcal{P}}(\Pi^j_{(t_x,t_y),\vec{q}}, p) = \frac{|\{\pi^j(a,b) \in \Pi^j_{(t_x,t_y),\vec{q}} : (a,p,b) \in \mathcal{G}\}|}{|\lambda(t_x)|\cdot|\lambda(t_y)|}$$

$$\mathcal{P}(p) = \frac{|\{(a,p,b) : \gamma(x) \sqsubseteq t_x \wedge \gamma(y) \sqsubseteq t_y \wedge (a,p,b) \in \mathcal{G}\}|}{|\lambda(t_x)|\cdot|\lambda(t_y)|}$$

- Calculate association strength for each path
  - $\forall \Pi^j_{(t_x,t_y),\vec{q}} \in \Pi^k(p)$, calculate $\widehat{\mathcal{P}}(\Pi^j_{(t_x,t_y),\vec{q}})$, $\widehat{\mathcal{P}}(\Pi^j_{(t_x,t_y),\vec{q}}, p)$ and $\mathcal{P}(p)$
  - $\forall \Pi^j_{(t_x,t_y),\vec{q}} \in \Pi^k(p)$ calculate $\widehat{\mathrm{NPMI}}(\Pi^j_{(t_x,t_y),\vec{q}}, p)$

- Calculate association strength for each path
  - $\forall \Pi^j_{(t_x,t_y),\vec{q}} \in \Pi^k(p)$, calculate $\widehat{\mathcal{P}}(\Pi^j_{(t_x,t_y),\vec{q}})$, $\widehat{\mathcal{P}}(\Pi^j_{(t_x,t_y),\vec{q}}, p)$ and $\mathcal{P}(p)$
  - $\forall \Pi^j_{(t_x,t_y),\vec{q}} \in \Pi^k(p)$ calculate $\widehat{\text{NPMI}}(\Pi^j_{(t_x,t_y),\vec{q}}, p)$
- Veracity calculation

$$Ver(p) = 1 - \prod_{\Pi^j_{(t_x,t_y),\vec{q}} \in \Pi^k(p)} \left(1 - \widehat{\text{NPMI}}(\Pi^j_{(t_x,t_y),\vec{q}}, p)\right).$$

Knowledge graph

- DBpedia version 2016-10
- Preprocessing: literals, links to external sources

Knowledge graph

- DBpedia version 2016-10
- Preprocessing: literals, links to external sources

Competing approaches

- **Knowledge Stream (KS):** Maximum information flow between paths [1]
- **Relational Knowledge Linker (KL-REL):** Information flow via single best path [1]
- **PredPath:** Supervised approach for learning discriminative paths [4]

- **Real-World:** Extracted from GREC and WSDM'17 challenge
- **Synthetic:** Facts from the domains of entertainment, sports etc.
- **FactBench-DBpedia:** Facts pertaining to DBpedia extracted from **FactBench** [5]

Dataset	*FactBench-DBpedia*						*Real-World*				*Synthetic*						
Subset	Domain	Domain-Range	Range	Mix	Property	Random	Birth-Place	Death-Place	Education	Nationality	US-CAP	NBA-Team	Oscars	CEO	US-WAR	US-VP	FLOTUS
Positive	1,124	1,124	1,124	1,124	1,124	1,124	273	126	466	50	50	41	78	201	126	47	16
Negative	1,119	1,006	1,123	1,014	1,153	511	819	378	1,395	150	250	123	4,602	1,007	584	227	240
Total	2,243	2,130	2,247	2,138	2,277	1,635	1,092	504	1,861	200	300	164	4,680	1,208	710	274	256

- Results of COPAAL using directed and undirected paths

	Domain	**Domain-Range**	**Range**	**Mix**	**Random**	**Property**
Undirected Paths	**0.9348**	**0.9389**	**0.8937**	**0.8561**	**0.9411**	**0.7307**
Directed Paths	0.7741	0.7824	0.7416	0.5914	0.6411	0.4713

- Results of COPAAL using directed and undirected paths

	Domain	**Domain-Range**	**Range**	**Mix**	**Random**	**Property**
Undirected Paths	**0.9348**	**0.9389**	**0.8937**	**0.8561**	**0.9411**	**0.7307**
Directed Paths	0.7741	0.7824	0.7416	0.5914	0.6411	0.4713

- Directed paths limit the evidences

- Results comparison of COPAAL against other approaches

	Birth Place	**Death Place**	**Education**	**Nationality**
COPAAL	**0.9441**	0.8997	**0.8731**	**0.9831**
PredPath	0.8997	0.8054	0.8644	0.9520
KL-REL	0.9254	**0.9095**	0.8547	0.9692
KS	0.7197	0.8002	0.8651	0.9789

- Results comparison of COPAAL against other approaches

	Birth Place	**Death Place**	**Education**	**Nationality**
COPAAL	**0.9441**	0.8997	**0.8731**	**0.9831**
PredPath	0.8997	0.8054	0.8644	0.9520
KL-REL	0.9254	**0.9095**	0.8547	0.9692
KS	0.7197	0.8002	0.8651	0.9789

- PredPath is limited to handle specificity of path w.r.t predicate in question
- KL-REL is limited to single best path

- Comparison of all approaches on the balanced dataset

	Domain	**DomainRange**	**Mix**	**Property**	**Random**	**Range**
COPAAL	**0.9348**	0.9389	**0.8561**	**0.7307**	**0.9411**	0.8937
PredPath	0.9301	**0.9447**	0.8408	0.7154	0.9354	**0.8992**
KL-REL	0.8453	0.8619	0.7721	0.6154	0.8547	0.8219
KS	0.8019	0.8124	0.7215	0.6047	0.7911	0.8047

- Comparison of all approaches on the balanced dataset

	Domain	**DomainRange**	**Mix**	**Property**	**Random**	**Range**
COPAAL	**0.9348**	0.9389	**0.8561**	**0.7307**	**0.9411**	0.8937
PredPath	0.9301	**0.9447**	0.8408	0.7154	0.9354	**0.8992**
KL-REL	0.8453	0.8619	0.7721	0.6154	0.8547	0.8219
KS	0.8019	0.8124	0.7215	0.6047	0.7911	0.8047

- Ontological information is important indeed

Summary:

- Evaluation on 17 datasets

Summary:

- Evaluation on 17 datasets
- Unsupervised - requires no labeled data

Summary:

- Evaluation on 17 datasets
- Unsupervised - requires no labeled data
- SPARQL - no alternate representation of knowledge graphs

Summary:

- Evaluation on 17 datasets
- Unsupervised - requires no labeled data
- SPARQL - no alternate representation of knowledge graphs
- Co-occurrence measure account for specificity and generality
  - $birthPlace(x, y) \leftarrow hometown(x, y) : 0.65$
  - $birthPlace(x, y) \leftarrow country(x, v_0) \wedge country(y, v_0) : 0.16$

Summary:

- Evaluation on 17 datasets
- Unsupervised - requires no labeled data
- SPARQL - no alternate representation of knowledge graphs
- Co-occurrence measure account for specificity and generality
  - $birthPlace(x, y) \leftarrow hometown(x, y) : 0.65$
  - $birthPlace(x, y) \leftarrow country(x, v_0) \wedge country(y, v_0) : 0.16$
- Co-occurrence measure applicable to paths of any length

Summary:

- Evaluation on 17 datasets
- Unsupervised - requires no labeled data
- SPARQL - no alternate representation of knowledge graphs
- Co-occurrence measure account for specificity and generality
  - $birthPlace(x, y) \leftarrow hometown(x, y) : 0.65$
  - $birthPlace(x, y) \leftarrow country(x, v_0) \wedge country(y, v_0) : 0.16$
- Co-occurrence measure applicable to paths of any length
- Verbalization of evidences
  - Demo: `http://copaal.dice-research.org/demo/`

Summary:

- Evaluation on 17 datasets
- Unsupervised - requires no labeled data
- SPARQL - no alternate representation of knowledge graphs
- Co-occurrence measure account for specificity and generality
  - $birthPlace(x, y) \leftarrow hometown(x, y) : 0.65$
  - $birthPlace(x, y) \leftarrow country(x, v_0) \wedge country(y, v_0) : 0.16$
- Co-occurrence measure applicable to paths of any length
- Verbalization of evidences
  - Demo: `http://copaal.dice-research.org/demo/`

Limitations:

- Existence of ontological information

Summary:

- Evaluation on 17 datasets
- Unsupervised - requires no labeled data
- SPARQL - no alternate representation of knowledge graphs
- Co-occurrence measure account for specificity and generality
  - $birthPlace(x, y) \leftarrow hometown(x, y) : 0.65$
  - $birthPlace(x, y) \leftarrow country(x, v_0) \wedge country(y, v_0) : 0.16$
- Co-occurrence measure applicable to paths of any length
- Verbalization of evidences
  - Demo: `http://copaal.dice-research.org/demo/`

Limitations:

- Existence of ontological information
- Limited to evidences found in one RDF graph

Acknowledgements

This project has been supported by German Federal Ministry of Transport and Digital Infrastructure (BMVI) in the projects **LIMBO** (no. 19F2029I) and **OPAL** (no. 19F2028A):

References

1. Shiralkar, P., Flammini, A., Menczer, F., Ciampaglia, G.L.: Finding streams in knowledgegraphs to support fact checking. In: 2017 IEEE International Conference on Data Mining(ICDM). pp. 859–864. IEEE (2017)
2. Lao, N., Cohen, W.W.: Relational retrieval using a combination of path-constrained randomwalks. Machine learning 81(1), 53–67 (2010)
3. Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. Proceed-ings of GSCL pp. 31–40 (2009)
4. Shi, B., Weninger, T.: Discriminative predicate path mining for fact checking in knowledgegraphs. Knowledge-based systems 104, 123–133 (2016)
5. https://github.com/DeFacto/FactBenc