What is normal, what is strange, and what is missing in a Knowledge Graph: Unified characterization via Inductive Summarization

Tiphaine Viard, DIG seminar

A paper by Belth, Zeng, Vreeken and Koutra, WWW'20
All figures taken from the paper

Goals and overview

Provide a **unified solution** to **KG characterization**

Find what is normal, infer what is abnormal

Rules are **labeled, rooted** graphs

"Books are written by authors, who are born in countries" ✓
"Authors writes Books" ✗

Find the rules that **best compress** the KG

Main problem: Inductive KG Summarization

Given a KG G, and ideal KG $\hat{G}$, find a concise model $M*$ of inductive rules that summarize what is normal in both G and $\hat{G}$.

Rules should be (1) interpretable ($=$ readable in natural language), (2) their exceptions should reveal abnormal information in the KG:

- erroneous ($t \in E : t \notin \hat{E}$),
- missing ($t \in \hat{E} : t \notin E$),
- an exception ($t \in E : t \in \hat{E}$).

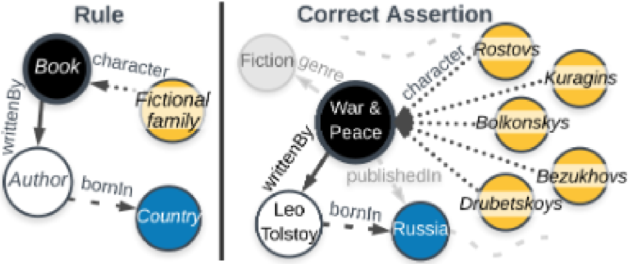
A KG $G = (V, E, \mathcal{L}_V, \mathcal{L}_E, \phi)$; also A matrix and L matrix
A model M is a set of rules
Authors use two-part MDL: for $M \in \mathcal{M}$, minimize $L(M) + L(\mathcal{D}|M)$

Rules and assertions

Rules are **recursive** and **compositional**: $g = (\mathcal{L}_g, \chi_g)$

- $\mathcal{L}_g$ root (e.g. Book)
- χ_g set of children $\{(p, \delta, \hat{g})\}$

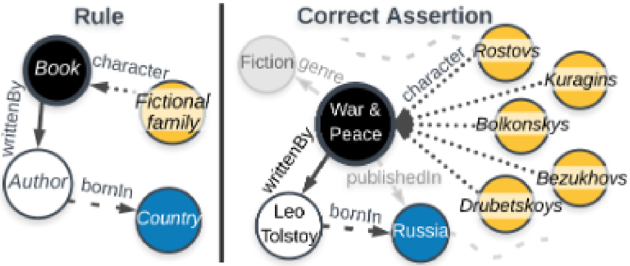
e.g. `Book`, (writtenBy, $\rightarrow$, (Author, $\emptyset$))



Rules and assertions

Assertions a_g are subgraphs asserted by rule g

Obtained by traversal with a start node that has $\mathcal{L}_g$ in its label



$\mathcal{A}^{(g)} = \mathcal{A}_c^{(g)} \cup \mathcal{A}_\xi^{(g)}$

- $\mathcal{A}_c^{(g)}$: all traversals a_g matching g's syntax (correct)
- $\mathcal{A}_\xi^{(g)}$: all traversals a_g not matching g's syntax (exceptions)

MDL model

Problem 2 (Inductive KG Summarization with MDL).
Given KG G, find the model $M*$ that minimizes the description length of the graph,

$$M* = argmin_{M \in M} L(G, M) = argmin_{M \in M} \{L(M) + L(G|M)\}$$

- $L(M)$ = cost of model M,
- $L(G|M)$ = cost of encoding G with M.

Cost = cost of **transmission (in bits) to reconstruct** G

Computing the cost of $L(M)$

$$L(M) = log(2 \cdot |\mathcal{L}_V|^2 \cdot \mathcal{L}_E + 1) + \sum_{g \in M} L(g) + L(\mathcal{A}^{(g)})$$

- $log(\dots) =$ number of rules, ✓
- $L(g) =$ cost of one rule $(p, \delta, \hat{g})$, ✗
- $L(\mathcal{A}^{(g)})$: cost of assertions. ✗

$$L(g) = L(\mathcal{L}_g) + L_{\mathbb{N}}(|\chi_g| + 1) + \sum_{\hat{g} \in \chi_g} -log \frac{n_p}{|E|} + 1 + L(\hat{g})$$

$$L(\mathcal{L}_g) = log|\mathcal{L}_V| + \sum_{\ell \in \mathcal{L}_g} -log \frac{n_\ell}{|V|},$$

Computing the cost of $L(M)$

$$L(M) = log(2 \cdot |\mathcal{L}_V|^2 \cdot \mathcal{L}_E + 1) + \sum_{g \in M} L(g) + L(\mathcal{A}^{(g)})$$

- $log(\dots) =$ number of rules, ✓
- $L(g) =$ cost of one rule $(p, \delta, \hat{g})$, ✓
- $L(\mathcal{A}^{(g)})$: cost of assertions. ✗

$$L(\mathcal{A}^{(g)}) = L(\mathcal{A}_c^{(g)}) + L(\mathcal{A}_\xi^{(g)})$$

$$L(\mathcal{A}_\xi^{(g)}) = log|\mathcal{A}^{(g)}| + log\binom{|\mathcal{A}^{(g)}|}{|\mathcal{A}_\xi^{(g)}|}$$

$$L(\mathcal{A}_c^{(g)}) = \sum_{a_g} L(a_g) = \sum_{\hat{g} \in \chi_g} log|V| + log\binom{|V|-1}{|\mathcal{A}_c^{(\hat{g})}|} + \sum_{a_{\hat{g}}} L(a_{\hat{g}})$$

Computing the cost of $L(M)$

$$L(M) = log(2 \cdot |\mathcal{L}_V|^2 \cdot \mathcal{L}_E + 1) + \sum_{g \in M} L(g) + L(\mathcal{A}^{(g)})$$

- $log(\dots) =$ number of rules, ✓
- $L(g) =$ cost of one rule $(p, \delta, \hat{g})$, ✓
- $L(\mathcal{A}^{(g)})$: cost of assertions. ✓

Computing the cost of $L(G|M)$

L, A : labels matrix, adjacency matrix
L_M, A_M: modelled labels, modelled edges
$L^- = L - L_M$, $A^- = A - A_M$
Sending what is not modelled:

- Unrevealed node labels
- Unmodelled edges

$$L(G|M) = L(L^-) + L(A^-)$$

With:

- $L(L^-) = log\binom{|\mathcal{L}_V|\cdot|V|-|L_M|}{|L-|}$
- $L(A^-) = log\binom{|\mathcal{L}_E|\cdot|V|^2-|A_M|}{|A-|}$

How to find all rules?

Naive approach: enumerate all rules from a set of candidates C

This is terrible! There are $2^{|C|}$ models to choose from

Contrary to support/confidence-based methods, there are no nice properties of the search space

No anti-monotonicity or (known) exploitable structure

Instead, use **compositionality** of rules

Start with **atomic rules** (assert one thing) and build up

Greedy approach is still costly (quadratic in $|C|$)

$$\Delta L(G|M_0 \cup \{g\}) = L(G|M_0) - L(G|M_0 \cup \{g\})$$

Rank using ΔL, descending

Constant number of passes on C

The KGist algorithm

Algorithm 1 KGist

Input: Knowledge graph G
Output: A model M, consisting of a set of rules

Read G and generate candidate rules C ▷ § 4.1.1
Qualify candidate rules with labels
Rank all rules $g \in C$ first by $\downarrow \Delta L(G|M_0)$ then by $\downarrow |\mathcal{A}_c(g)|$ and $\downarrow$ lexicographic $\mathcal{L}_g$ ▷ § 4.1.3, Eq. (12)
$M \leftarrow \emptyset$
while not converged **do** ▷ i.e., more rules can be added to M
    for $g \in C$ **do**
        if $L(G, M \cup \{g\}) < L(G, M)$ **then** ▷ § 4.2.1
            $M \leftarrow M \cup \{g\}$
            $C \leftarrow C \setminus \{g\}$
Optionally perform refinements **Rm** and **Rn** ▷ § 4.2.2

Complexity: $\mathcal{O}(m\phi_{max}^2 \cdot log(m\phi_{max}^2))$

Evaluation

Goal: answer the questions

1. Does KGIST characterize what is normal? How well can KGIST compress KGs?
2. Does KGIST identify what is strange? Can it identify and characterize multiple types of errors?
3. Does KGIST identify what is missing?
4. Is KGIST scalable?

Table 2: Description of KG datasets: number of nodes, edges, node labels, relations, and average / median labels per node, resp.

	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{L}_{\mathcal{V}}\rvert$	$\lvert\mathcal{L}_{\mathcal{E}}\rvert$	avg $\phi(v)$	med $\phi(v)$
NELL	46,682	231,634	266	821	1.53	1
DBpedia	976,404	2,862,489	239	504	2.72	3
Yago	6,349,336	12,027,848	629,681	33	3.81	3

[Q1] What is normal?

Setup. Compare compression as compared to an empty model M_0 (i.e. the whole graph is an error)

- Freq: select most often top k rules that apply (instead of MDL),
- Coverage: number of edges explained by the rule
- AMIE+ [Fabian's work] does not compress, so report only # rules

		Horn rules	Rules of the form $g = (\mathcal{L}_g, \chi_g)$				
Dataset	*Metric*	AMIE+	Freq	Coverage	KGist	KGist+m	KGist+n
NELL (6,268,200 bits)	% Bits needed	N/A	191.46%	192.72%	73.88%	73.00%	63.57%
	Edges Explained	N/A	57.33%	50.12%	78.52%	78.52%	74.67%
	# Rules	32,676	top-k	top-k	1,115	647	573
DBpedia (119,117,468 bits)	% Bits needed	N/A	674.51%	718.22%	69.88%	69.84%	69.77%
	Edges Explained	N/A	80.64%	71.70%	89.17%	89.17%	88.51%
	# Rules	~6,963 [17]	top-k	top-k	516	505	498
Yago (793,027,801 bits)	% Bits needed	N/A	896.33%	947.64%	76.13%	75.98%	75.04%
	Edges Explained	N/A	86.54%	83.44%	88.40%	88.40%	85.20%
	# Rules	failed	top-k	top-k	60,298	34,331	32,670

[Q2] What is strange?

Setup. Missing (A1), superfluous (A2), swapped (A4) labels, erroneous links (A3)
Baselines. KGist-freq, AMIE+, others

		Supervised		Unsupervised			
Task	*Metric*	ComplEx	TransE	SDValidate	AMIE+	KGist_Freq	KGist+m
All anomalies	AUC	0.5508 ± 0.02	0.5779 ± 0.04	0.4996 ± 0.00	0.4871 ± 0.04	0.5739 ± 0.01	**0.6052 ± 0.03***
	P@100	0.4820 ± 0.05	0.7040 ± 0.06	0.5100 ± 0.04	0.3980 ± 0.07	0.6816 ± 0.10	**0.7419 ± 0.07***
	R@100	0.0087 ± 0.00	0.0126 ± 0.00	0.0092 ± 0.00	0.0072 ± 0.00	0.0126 ± 0.00	**0.0139 ± 0.00***
	F1@100	0.0172 ± 0.00	0.0247 ± 0.00	0.0181 ± 0.00	0.0141 ± 0.00	0.0247 ± 0.01	**0.0273 ± 0.01***
***A1** missing labels*	AUC	0.5842 ± 0.04	0.6021 ± 0.06	0.4997 ± 0.00	0.4409 ± 0.06	0.5149 ± 0.02	**0.6076 ± 0.03***
	P@100	0.2640 ± 0.05	0.4280 ± 0.15	0.3040 ± 0.06	0.1200 ± 0.05	0.4067 ± 0.11	**0.4759 ± 0.05***
	R@100	0.0119 ± 0.00	0.0181 ± 0.01	0.0134 ± 0.00	0.0057 ± 0.00	0.0199 ± 0.01	**0.0244 ± 0.01***
	F1@100	0.0227 ± 0.01	0.0346 ± 0.01	0.0257 ± 0.01	0.0109 ± 0.01	0.0377 ± 0.01	**0.0463 ± 0.02***
***A2** superfluous labels*	AUC	0.5502 ± 0.02	0.5659 ± 0.03	0.4989 ± 0.01	0.4946 ± 0.03	0.4997 ± 0.04	**0.5115 ± 0.03**
	P@100	0.1780 ± 0.05	0.3160 ± 0.16	0.2160 ± 0.07	0.1040 ± 0.09	0.2081 ± 0.06	**0.2485 ± 0.09**
	R@100	0.0122 ± 0.00	0.0219 ± 0.01	0.0152 ± 0.00	0.0070 ± 0.01	0.0169 ± 0.01	**0.0175 ± 0.01**
	F1@100	0.0229 ± 0.00	0.0408 ± 0.02	0.0283 ± 0.01	0.0131 ± 0.01	0.0311 ± 0.01	**0.0326 ± 0.01**
***A3** erroneous links*	AUC	0.2495 ± 0.03	0.4126 ± 0.08	0.4966 ± 0.01	**0.8902 ± 0.08**	0.7383 ± 0.00	0.8423 ± 0.00
	P@100	0.1020 ± 0.04	0.0020 ± 0.00	0.0480 ± 0.02	**0.1860 ± 0.08***	0.0131 ± 0.01	0.0137 ± 0.01
	R@100	0.0374 ± 0.02	0.0007 ± 0.00	0.0176 ± 0.01	**0.0679 ± 0.03***	0.0051 ± 0.01	0.0052 ± 0.01
	F1@100	0.0548 ± 0.02	0.0011 ± 0.00	0.0257 ± 0.01	**0.0995 ± 0.05***	0.0074 ± 0.01	0.0075 ± 0.01
***A4** swapped labels*	AUC	0.5369 ± 0.03	0.5527 ± 0.02	0.4991 ± 0.00	0.4891 ± 0.03	**0.6904 ± 0.01***	0.6633 ± 0.07
	P@100	0.2160 ± 0.08	0.4200 ± 0.09	0.2080 ± 0.08	0.1240 ± 0.06	**0.5360 ± 0.15***	0.4768 ± 0.10
	R@100	0.0136 ± 0.00	0.0269 ± 0.01	0.0128 ± 0.00	0.0079 ± 0.00	**0.0379 ± 0.01***	0.0320 ± 0.01
	F1@100	0.0256 ± 0.01	0.0505 ± 0.01	0.0241 ± 0.01	0.0148 ± 0.01	**0.0705 ± 0.01***	0.0599 ± 0.01
Avg rank		4.10	2.90	4.15	5.00	2.90	**1.95**

[Q3] What is missing?

Assume PCA, removes $q\%$ nodes from G, identify $\mathcal{A}(^{(g)}_{\xi}$

Baselines.

Metrics.

		Supervised		Unsupervised	
Dataset	*Metric*	LP	AMIE+C [16]	Freq	KGIST
NELL	**R**	N/A	0.6587 ± 0.03	0.4589 ± 0.02	0.7598 ± 0.02
	$\mathbf{R_L}$	N/A	N/A	0.3924 ± 0.02	0.6636 ± 0.01
DBpedia	**R**	N/A	0.8187 ± 0.01	0.8049 ± 0.01	0.9288 ± 0.00
	$\mathbf{R_L}$	N/A	N/A	0.7839 ± 0.01	0.9179 ± 0.00

Conclusion and thoughts

In brief:

- MDL-based method for extracting rule sets out of knowledge graphs
- Tasks: description, error detection and KG completion tasks
- Data: NELL, DBPedia, YAGO
- Code is online : `github.com/GemsLab/KGist`

Thoughts:

- Really well written and thorough, easy to follow despite lots of contributions
- Hard for me to evaluate if it is performing well or if well chosen task