HEINZ NIXDORF INSTITUT
UNIVERSITÄT PADERBORN

Knowledge Graph Summarization Project Group (KGSUMM PG)

Brainstorming Section

Prof. Dr. Axel-Cyrille Ngonga Ngomo
Tutor: Asep Fajar Firmansyah

DICE
Data Science Group
Paderborn University

Project Group - WiSe 2025
October 16, 2025

Part I

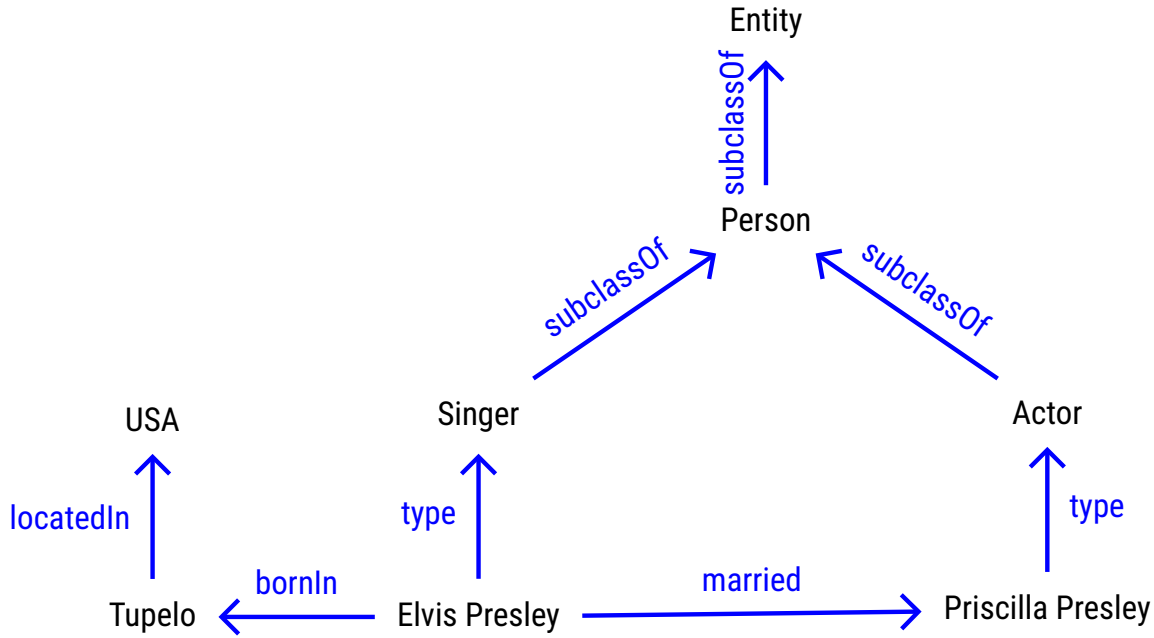
Introduction to Knowledge Graphs and Entity Summarization

Knowledge Graph

Introduction

- A graph of data that models real-world knowledge.
- A knowledge graph $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times (\mathcal{C} \cup \mathcal{L} \cup \mathcal{E})$

- **Nodes**
  - Entities ($\mathcal{E}$)
  - Classes ($\mathcal{C}$)
  - Literal ($\mathcal{L}$)
- **Directed edges**
  - properties of entities
  - Typing of entities
  - Class hierarchies



source:https://sites.google.com/view/entity-summarization-tutorials/www2020

Knowledge Graph

Aibel Knowledge Graph

1. Aibel provides services in oil, gas, offshore wind industries
   - 1.1 produces offshore oil platforms
2. Aibel MMG KG
   - 2.1 Captures
     - 2.1.1 national and international standards (industrial ISO, environmental)
     - 2.1.2 catalogs of components for oil platforms
     - 2.1.3 compatibilities between components
   - 2.2 Facilitates 3D design of oil platforms
   - 2.3 Optimises weight control, logistics, storage

How can entity summaries be useful?

- An entity in the Aibel KG (e.g., a pump) can be compatible with 100s of other entities
- A summary can help in getting a better overview of KG elements

M. G. Skjæveland, A. Gjærver, C. M. Hansen, J. W. Klüver, M. R. Strand, A. Waaler, P. Ø. Øverlid: Semantic Material Master Data Management at Aibel. ISWC 2018

Knowledge Graph

Festo Knowledge Graph

1. Festo offers drive products
   - 1.1 To move things, for factory automation
   - 1.2 E.g. a drive train
     - 1.2.1 Composed from basic components
     - 1.2.2 Modular structure
     - 1.2.3 Billions of combinations / variations
2. Festo Knowledge Graph
   - 2.1 Captures basic components
     - label=i. Their characteristics
     - lbbel=ii. Compatibilities
   - 2.2 Help to enhance design of drive trains

S. Elmer, F. Jrad, T. Liebig, A. ul Mehdi, M. Opitz, T. Stauß, D. Weidig: Ontologies and Reasoning to Capture Product Complexity in Automation Industry. ISWC 2017.

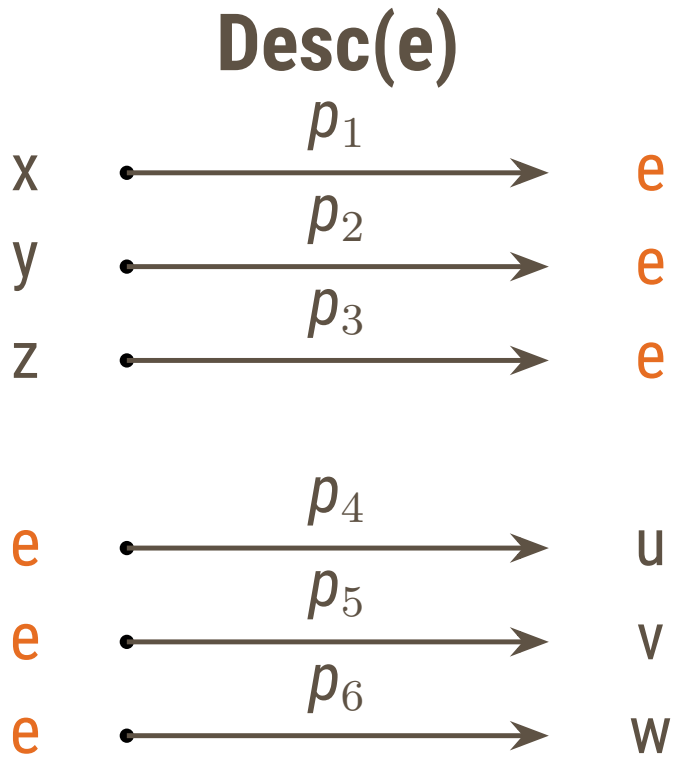
How can entity summaries be useful?

- In the Festo KG a drive train can be compatible with 100s of basic elements, drive trains
- A summary can help in finding the most relevant compatibilities

Entity Summarization

Entity Description

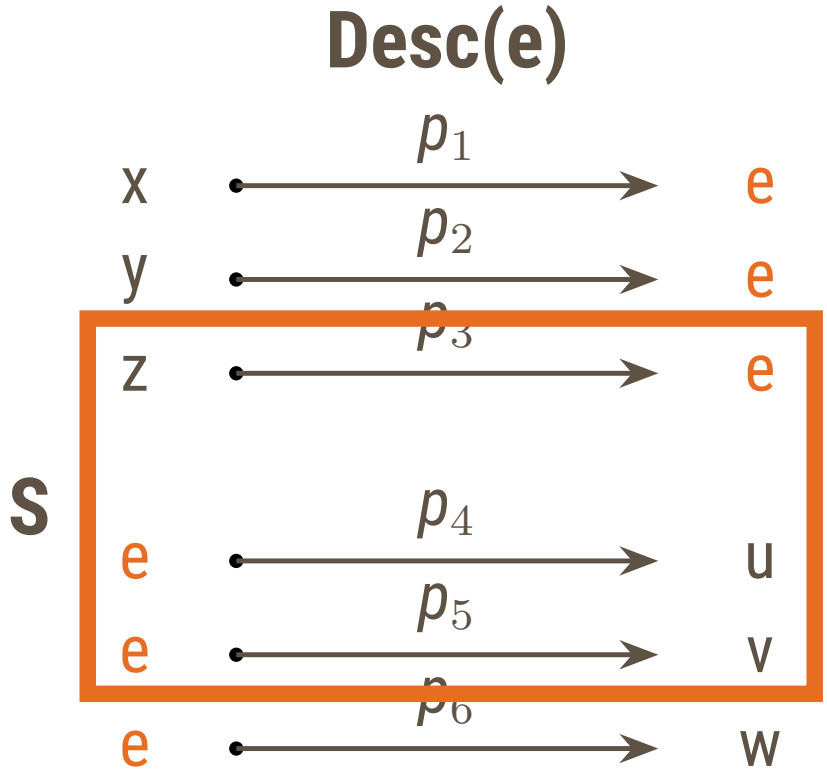
- Given:
  1. A knowledge graph $T \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, i.e., a set of triples (s, p, o) with $s, e \in \mathcal{E}$, $p \in \mathcal{R}$, and $o \in (\mathcal{C} \cup \mathcal{L} \cup \mathcal{E})$.
  2. An entity description $Desc(e, T) \subseteq T$ (e.g., the subset of triples in T that mention an entity e).
     - $Desc(e, T) = \{(e, p, o) \in T \vee (s, p, e) \in T\}$



Entity Summarization

Entity Summary

- Given:
  1. A knowledge graph $T \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, i.e., a set of triples (s, p, o) with $s, e \in \mathcal{E}$, $p \in \mathcal{R}$, and $o \in (\mathcal{C} \cup \mathcal{L} \cup \mathcal{E})$.
  2. An entity description $Desc(e, T) \subseteq T$ (e.g., the subset of triples in T that mention an entity e).
     - $Desc(e, T) = \{(e, p, o) \in T \vee (s, p, e) \in T\}$
- Entity summary (S)
  - A summary of an entity e in $\mathcal{T}$ of size k
  - Problem definition [1]



$$\underset{S \subseteq \text{Desc}(e)}{\text{find}} \quad \underset{S}{\arg\max}\ \text{score}(S \mid \mathcal{T}), \quad \text{subject to } |S| \leq k$$

Part II

Datasets & Existing Methods

Datasets

Existing Benchmark Datasets

Dataset	Entities	URL
FACES Evaluation Dataset (Gunarathna et al., AAAI 2015)	50 DBpedia entities	http://wiki.knoesis.org/index.php/FACES / alternative link [1]
ESBM v1.2 (Liu et al., ESWC 2020)	125 DBpedia entities, 50 LinkedMDB entities	https://w3id.org/esbm/
Wiki Entity Summarization Benchmark (WikES) (Javadi et al, arXiv 2024)	1948 Wikidata entities with different size (small, medium, large)	https://github.com/msorkhpar/wiki-entity-summarization

[1] https://files.dice-research.org/users/asep/datasets/entity-summarization/original_faces_dataset.zip

Existing Methods

Unsupervised Learning

- RELIN: RELatedness and Informativeness-based centrality for entity summarizatioN (2011)
- Linksum: Using link analysis to summarize entity data (2016)
- PageRank and Generic EntitySummarization for RDF Knowledge Bases (2018)
- KAFCA: Entity Summarization Based on Formal Concept Analysis (2018)
- BAFREC: Balancing Frequency and Rarity for Entity Characterization in Open Linked Data (2018)
- MPSUM: Predicate-Based Matching for RDF Triples with Application to LDA (2018)
- IRES: Untapping the Power of Indirect Relationships in Entity Summarization (2025)

Existing Methods

Results Comparison (F-Measure)

Method	DBpedia		LinkedMDB	
	$k = 5$	$k = 10$	$k = 5$	$k = 10$
RELIN	0.242	0.455	0.203	0.258
DIVERSUM	0.249	0.507	0.207	0.358
FACES	0.270	0.428	0.169	0.263
FACES-E	0.280	0.488	0.313	0.393
CD	0.283	0.513	0.217	0.331
LinkSUM	0.287	0.486	0.140	0.279
BAFREC	0.335	0.503	0.360	0.402
KAFCA	0.314	0.509	0.244	0.397
MPSUM	0.314	0.512	0.272	0.423
ORACLE	**0.595**	**0.713**	**0.619**	**0.678**

Existing Methods

Supervised Learning

- ESA:Entity Summarization Attention (2019)
- DeepLENS: Deep Learning for Entity Summarization (2020)
- AutoSUM: Automating Feature Extraction and Multi-user Preference for Entity Summarization (2020)
- GATES: Graph Attention Networks for Entity Summarization (2021)
- ESCS: Entity Summarization via Exploiting Description Complementary and Salience (2023)
- ESLM: Improving Entity Summarization by Leveraging Language Models (2024)

Existing Methods

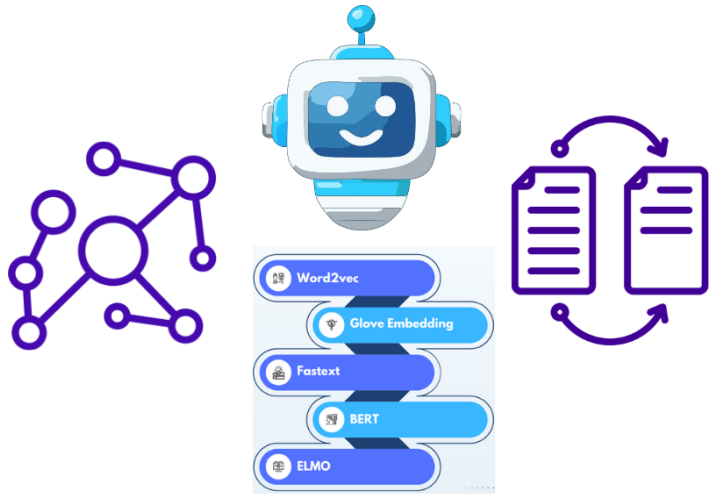
Results Comparison (F-Measure)

	DBpedia		LinkedMDB		FACES	
	k=5	k=10	k=5	k=10	k=5	k=10
ESA	0.332	0.532	0.353	0.435	0.153	0.261
NEST	0.354	0.540	0.332	0.465	0.272	0.346
AutoSUM	0.372	0.555	0.430	0.520	0.241	0.316
DeepLENS	0.404	0.575	0.469	0.489	0.130	0.248
GATES	0.423	0.574	0.437	**0.535**	0.254	0.324
ESCS	0.415	0.582	**0.494**	0.512	-	-
ESLM	**0.427**	**0.591**	0.467	0.498	**0.301**	**0.369**

Next Steps

- Identify gaps not covered by existing methods.
- Define and address the research problem(s).
- Propose a new approach informed by these insights.

That's all Folks!



Thank you!

dice-research.org/teaching/KGSUMM-2025

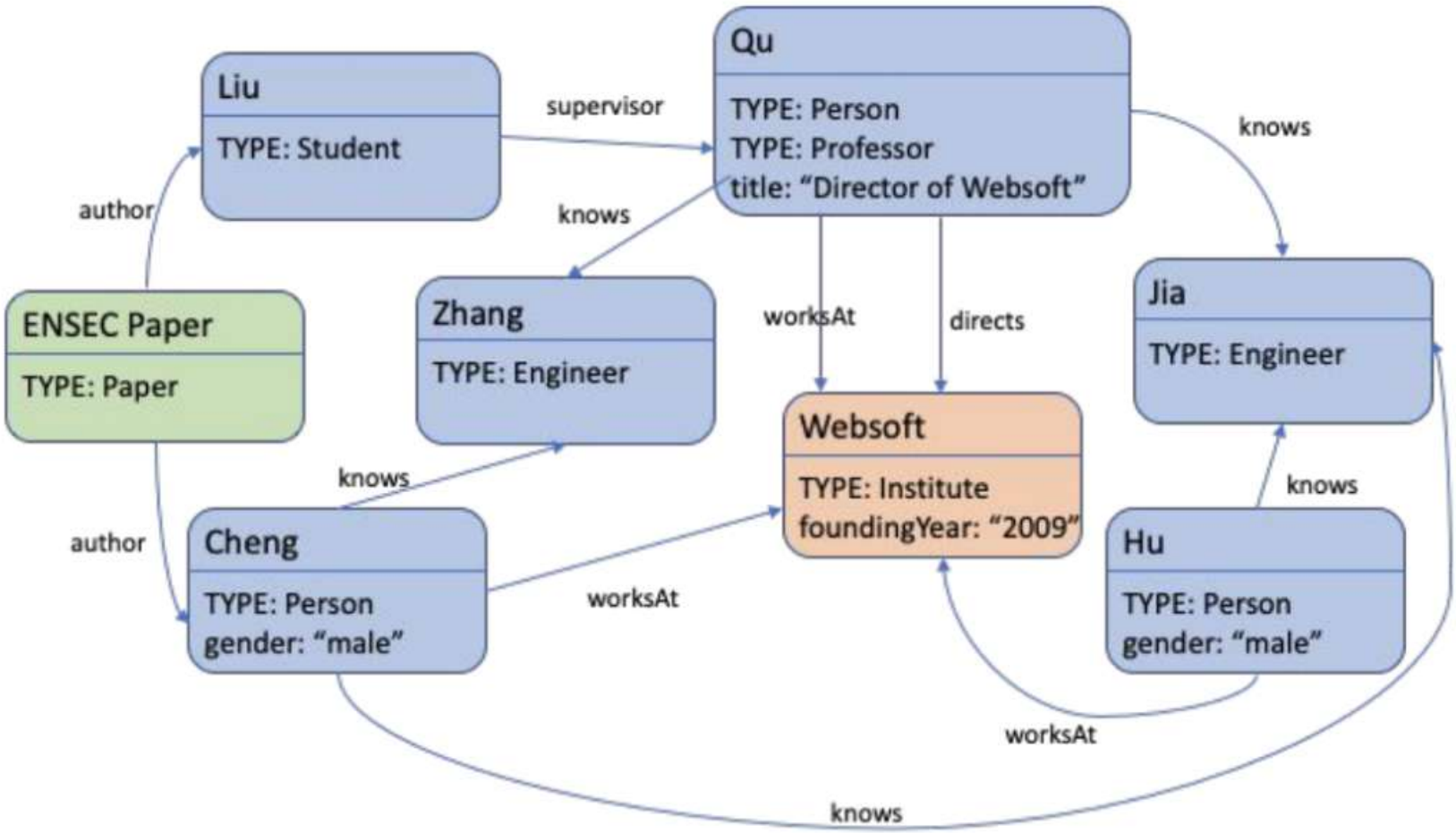
Have questions?

Email: asep.fajar.firmansyah@uni-paderborn.de
Matrix: @asepff:chat.dice-research.org

All source images are from Google Search and Canva.

Knowledge Graph

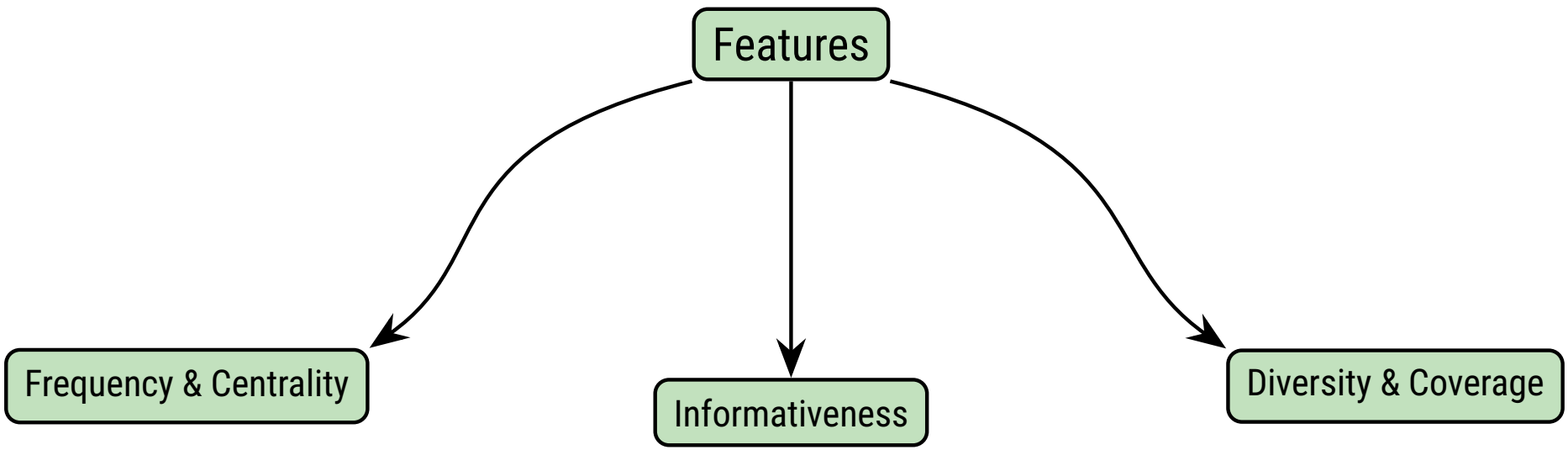
An example



Technical Features

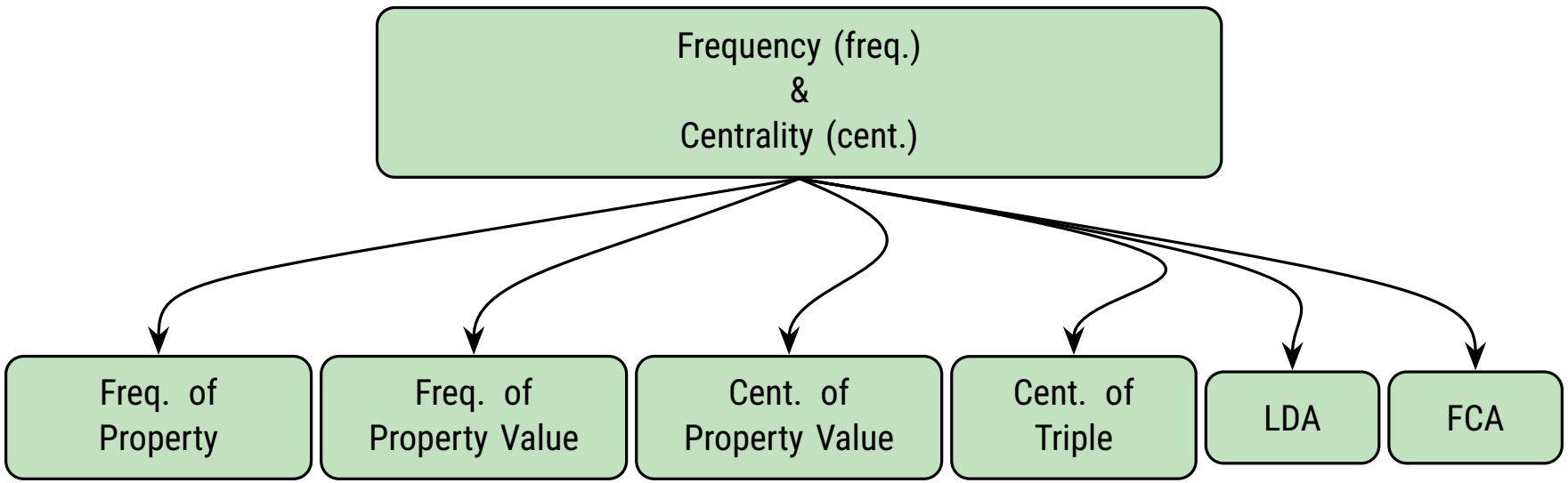
Main Categories

- These are the features that have been used mainly in the unsupervised (heuristic or feature engineered machine learning based) entity summarization methods.

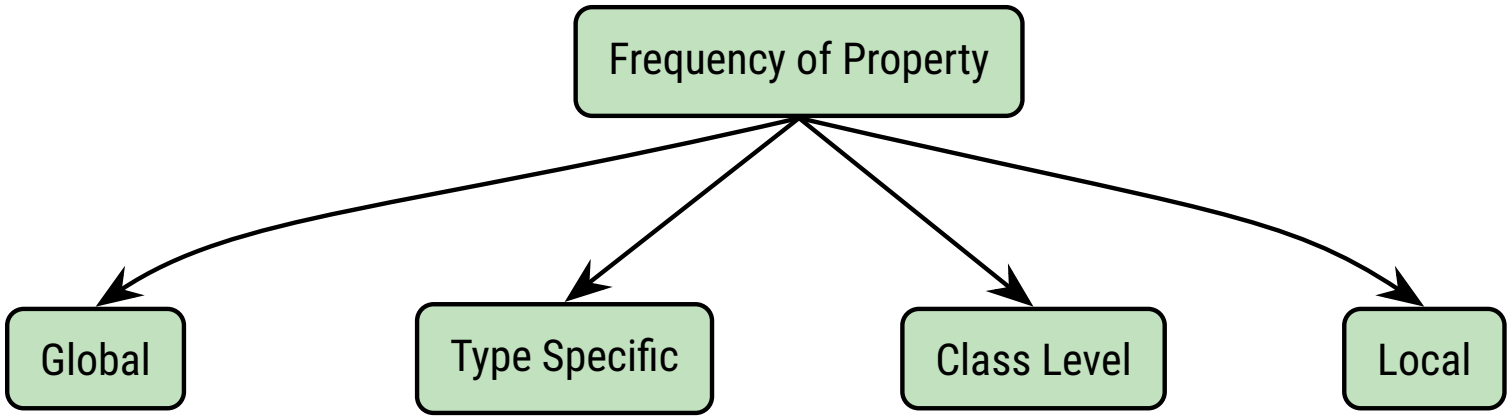


Technical Features

Frequency and Centrality



Frequency of Property



(E.g., Hasibi et al., SIGIR 2017)

$gf_g(p) = |\{e' \in \mathcal{E} : \exists t \in \mathrm{Desc}(e'), \mathrm{prp}(t) = p\}|$

Calculated over all entity descriptions in the dataset. Example: worksAt ranked higher than directs.

$gf_{\mathcal{T}}(p) = |\{t \in \mathcal{T} : \mathrm{prp}(t) = p\}|$

Calculated over all triples in the dataset. For instance, knows and worksAt have the same value over entity descriptions, but counting triples may differ.

(E.g., Kim et al., IEICE 2017)

$tf(p) = |\{e' \in \mathrm{Inst}(c) : \exists t \in \mathrm{Desc}(e'), \mathrm{prp}(t) = p\}|$

Number of entity descriptions of a specific type that have the property.

(E.g., Hasibi et al., SIGIR 2017; Kim et al., IEICE 2017)

$ntf(p) = |\{c' \in C : \exists e' \in \mathrm{Inst}(c'), t \in \mathrm{Desc}(e'), \mathrm{prp}(t) = p\}|$

Number of types (classes) in which the property appears in the descriptions.
Example: worksAt appears in two classes of entities.

(E.g., Cheng et al., WWW 2015)

$lf(p) = |\{t \in \mathrm{Desc}(e) : \mathrm{prp}(t) = p\}|$

Number of appearances of a property within one entity description.
Example: for entity Qu, knows may be ranked higher than worksAt.

Frequency of Property Value
(e.g., Hasibi et al., SIGIR 2017)

- The importance of a triple to the summary also depends on the frequency (i.e., popularity) of the property value.
- Measured at the entity level:

$$\mathrm{vf}_E(v) = \left|\{\, e' \in \mathcal{E} \;:\; \exists\, t \in \mathrm{Desc}(e'),\ \mathrm{val}(t) = v\}\right|$$

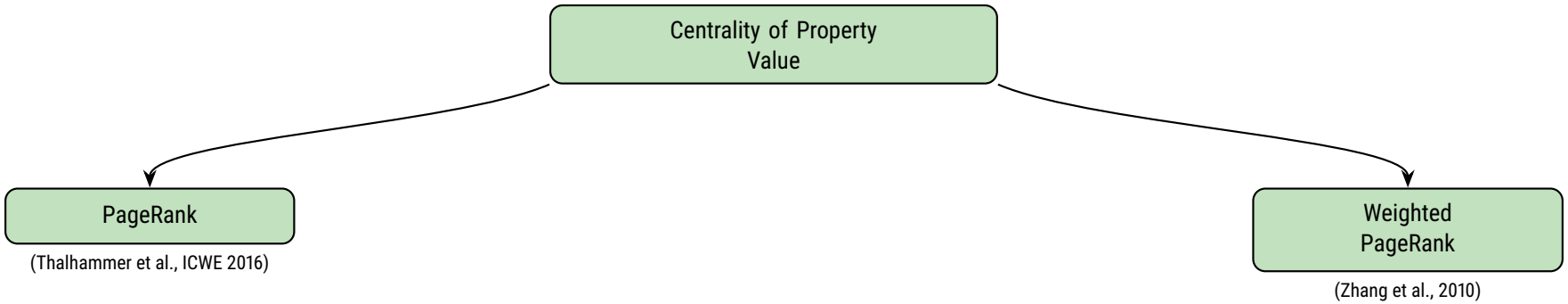
Calculates # of subject entities that the value appears in their descriptions

- Measured at the triple level:

$$\mathrm{vf}_T(v) = |\{\, t \in \mathcal{T} \;:\; \mathrm{val}(t) = v\}|$$

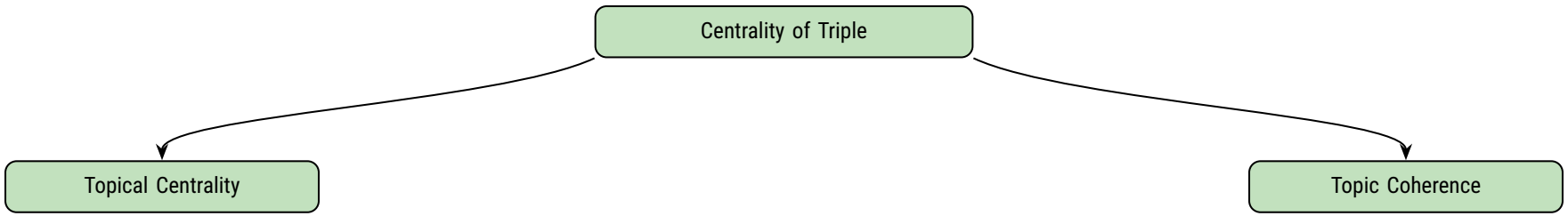
Calculates # of triples that the value appears in their descriptions

Centrality of Property Value



- Since a KG is a graph, centrality measures can quantify the importance of property values.
- Centrality (e.g., PageRank) captures connectivity beyond simple frequency.
- In semantic graphs, edges are labeled; a weighted PageRank is often more suitable.

Centrality of Triple



(e.g., Cheng et al., ISWC 2011)

Represent triples as nodes and connect them via similarity edges; then use PageRank to compute triple-level centrality.

(e.g., Yan et al., WAIM 2016)

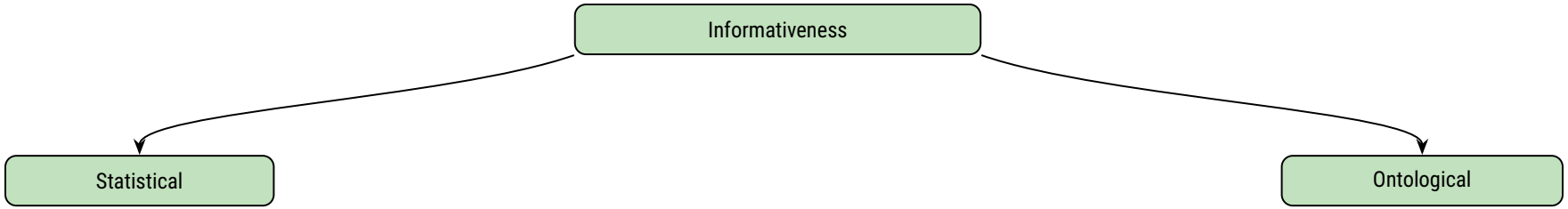
Extends topical centrality using query-based topic coherence.

LDA and FCA

(e.g., Pouriyeh et al., JCNLP 2017.) **(e.g., Kim et al., EYRE 2018.)**

- Latent Dirichlet Allocation (LDA) technique is adapted to process triples by:
  - Properties are considered as topics
  - Each property is a distribution over all the property values
- Formal Concept Analysis (FCA) technique is adapted by:
  - Aggregate properties and values into a hierarchy.
  - Then a triple is scored by its elements' depth in the hierarchy.
- FCA implicitly gives preference to infrequent properties and frequent property values.

Informativeness



(e.g., Gunaratna et al., AAAI 2015)

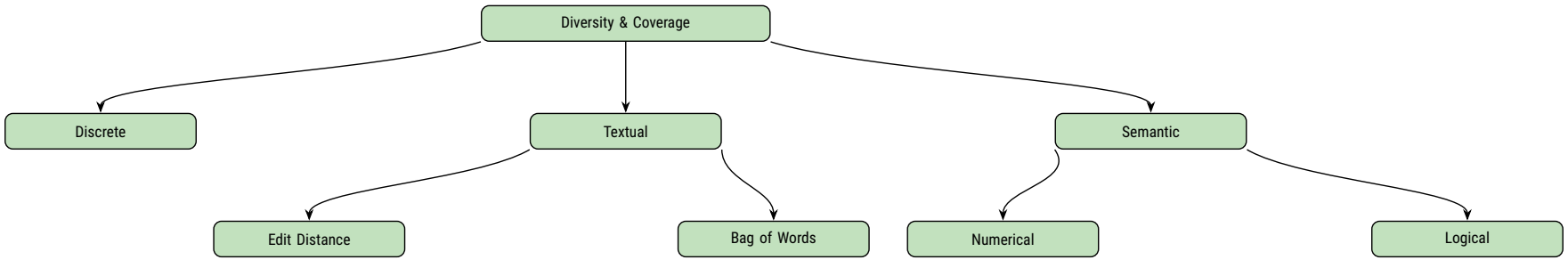
$$\mathrm{si}(< p, v >) \;=\; -\log\left(\frac{\left|\{\, e' \in \mathcal{E} : (e', p, v) \in \mathrm{Desc}(e')\,\}\right|}{|\mathcal{E}|}\right)$$

A property–value pair is considered more important to a summary when both are rare. For example, the pair ⟨worksAt, Websoft⟩ has a higher score than ⟨directs, Websoft⟩ if fewer entities work at Websoft than direct it.

(e.g., Kroll et al., EYRE 2018)

Ontological informativeness is used when the property is of type “TYPE” and the values are classes. Class hierarchy depth helps measure informativeness: deeper (more specific) classes convey more information and thus are considered more informative, typically because they have fewer instances.

Diversity & Coverage



(e.g., Sydow et al., JIIS 2013)

Discrete: Two properties (or values) i and j measured by binary similarity $ds(i, j)$:

$$ds(i,j) = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

(e.g., Cheng et al., WWW 2015; Gunaratna et al., AAAI 2015)

Edit Distance: similarity via edit distance of properties/values (e.g., Websoft vs. Websofr).

Bag of Words: similarity via overlap/cosine of word vectors of properties/values.

(e.g., Cheng et al., WWW 2015)

Numerical: similarity of numeric property values n_i, n_j

$$\mathrm{n}(n_i, n_j) = \begin{cases} 1, & n_i = n_j, \\ -1, & n_i \neq n_j \;\&\; n_i n_j \leq 0, \\ \dfrac{\min(|n_i|, |n_j|)}{\max(|n_i|, |n_j|)}, & n_i \neq n_j \;\&\; n_i n_j > 0. \end{cases}$$

Logical: class-level overlap/reasoning for property values (e.g., subclasses, shared types).

References

- [1] Liu, Q., Cheng, G., Gunaratna, K. and Qu, Y., 2021. Entity summarization: State of the art and future challenges. Journal of Web Semantics, 69, p.100647.

- Some content is adopted from `https://sites.google.com/view/entity-summarization-tutorials/www2020`