UNIVERSITÄT PADERBORN

KNOWLEDGE GRAPH SUMMARIZATION

UNIVERSITÄT PADERBORN

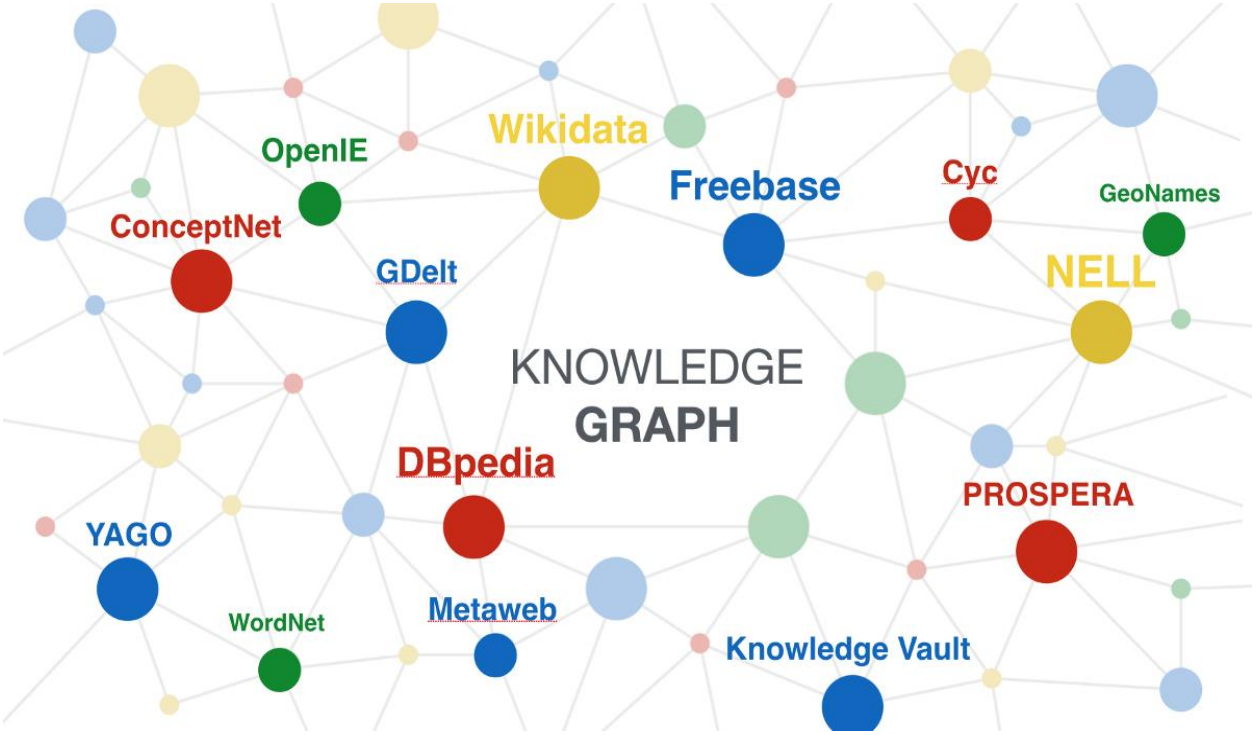
Group Members

- Shreyas Kottur Shivananda
- Pavan Kumar Sheshanarayana
- Muhammad Haseeb Javaid
- Usman Ashraf

UNIVERSITÄT PADERBORN

Knowledge Graphs

Knowledge Graph : A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge. **(Ehrlinger et al., 2016)**



UNIVERSITÄT PADERBORN

Problem Statement

- **Problem:** Enormous amounts of data is available.
  - Unstructured data.
  - Structured data is presented in the form of Knowledge Graph (KG).
- **Requirement:** Reduction of KG size.
- **Proposed Solution:** KG summarization.
- **End Goal**
  - Creation of summarized KG.
  - Evaluation of summarized KG.
  - UI functionality implementation.

UNIVERSITÄT PADERBORN

Summarization Algorithms

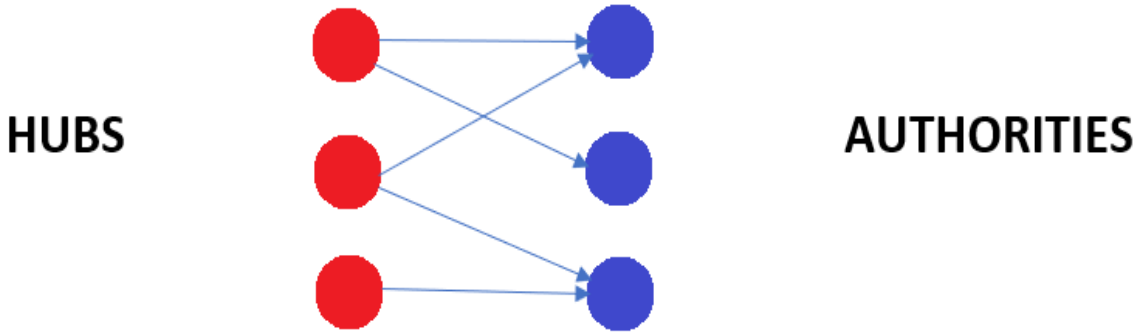
- **KG summarization**
  - Hyperlink-Induced Topic Search (HITS). **(Kleinberg, 1999)**
  - Stochastic Approach for Link-Structure Analysis (SALSA).**(Lampel et al., 2001)**

- **Entity summarization**
  - LinkSUM **(Thalhammer et al., 2016)**

UNIVERSITÄT PADERBORN

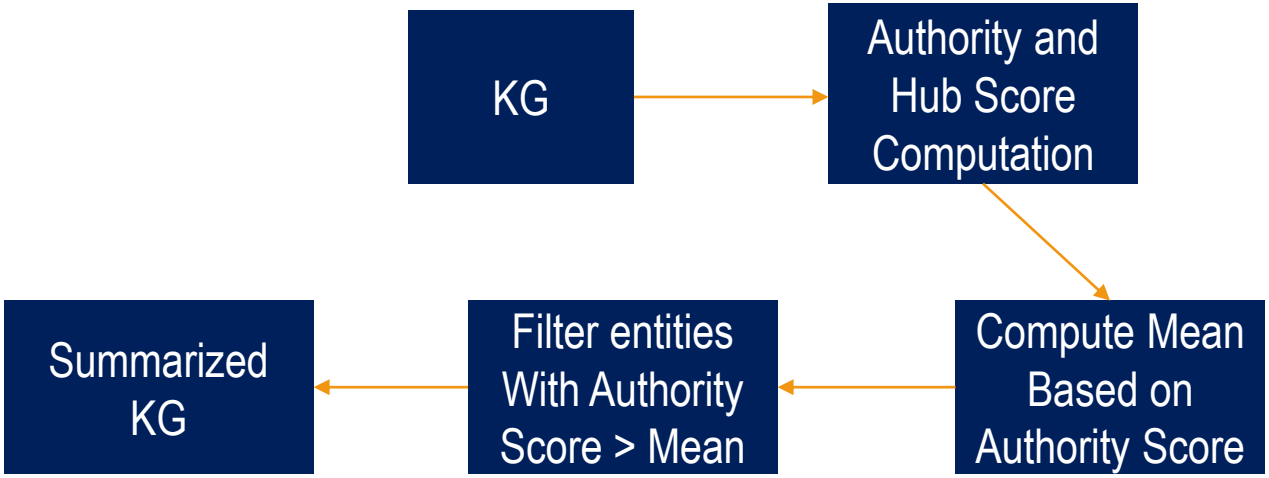
HITS - Introduction

- Link Analysis Algorithm.
- Uses weights for assigning scores.
- Each node gets two scores: Authority score and Hub score.
- Authority scores represent number of links towards the respective node.
- Hub scores represent numbers of links from the respective node.



UNIVERSITÄT PADERBORN

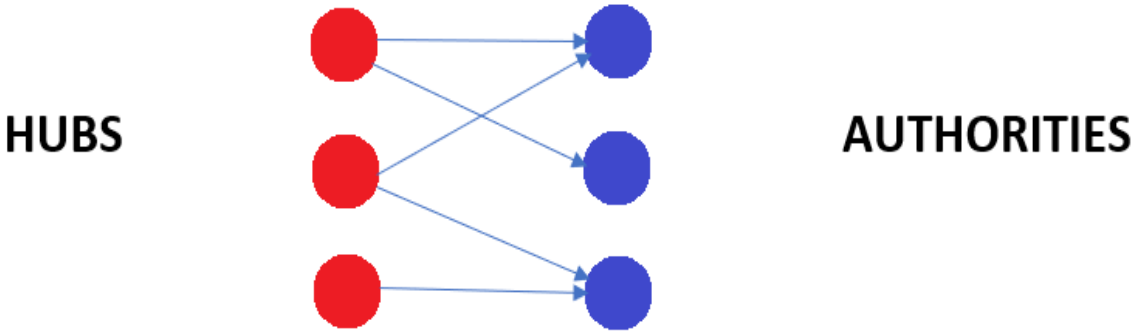
HITS (continued…)



UNIVERSITÄT PADERBORN

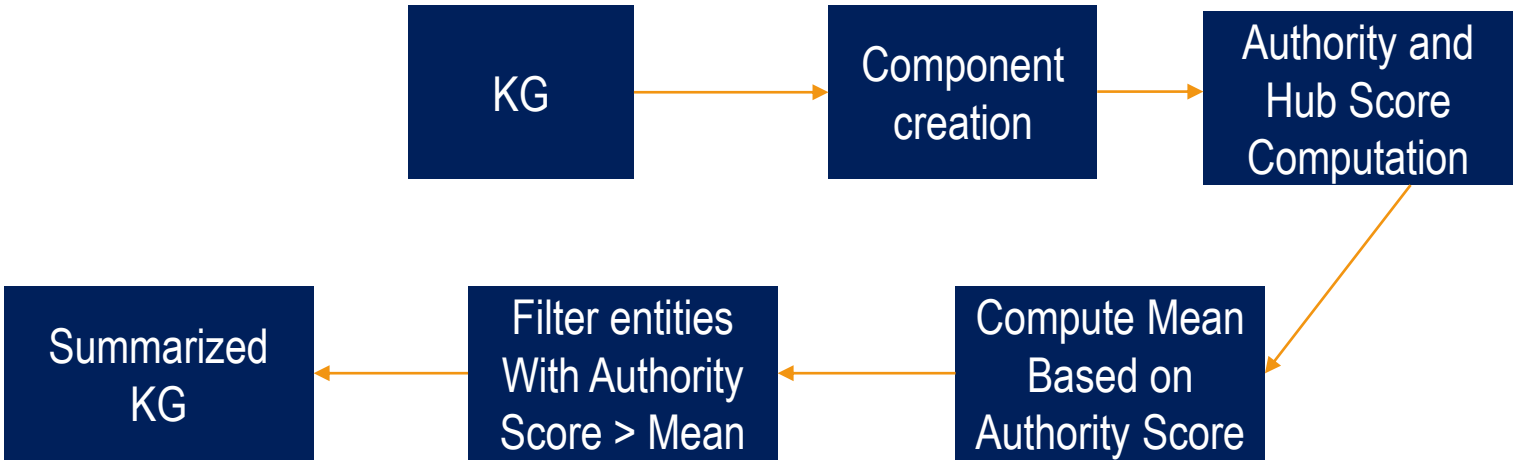
SALSA - Introduction

- The Stochastic Approach for Link-Structure Analysis
- Root set.
- Base set.
- Authoritative nodes and Hub nodes are represented as Bipartite graph.
- Nodes with good authoritative scores and hub scores are visited often.
- Computes on hubs and authority nodes separately.



UNIVERSITÄT PADERBORN

SALSA (Continued….)



UNIVERSITÄT PADERBORN

SALSA vs HITS

- Mutual reinforcement relationship.
- Iterative computation of authority and hub scores.
- SALSA not affected by TKC effect.

UNIVERSITÄT PADERBORN

HITS and SALSA - Statistics

- **HITS**
  - Original KG Resources – 3,116,745
  - Summarized KG Resources – 1,235,850
- **SALSA**
  - Original KG Resources – 3,116,745
  - Summarized KG Resources – 1,027,786

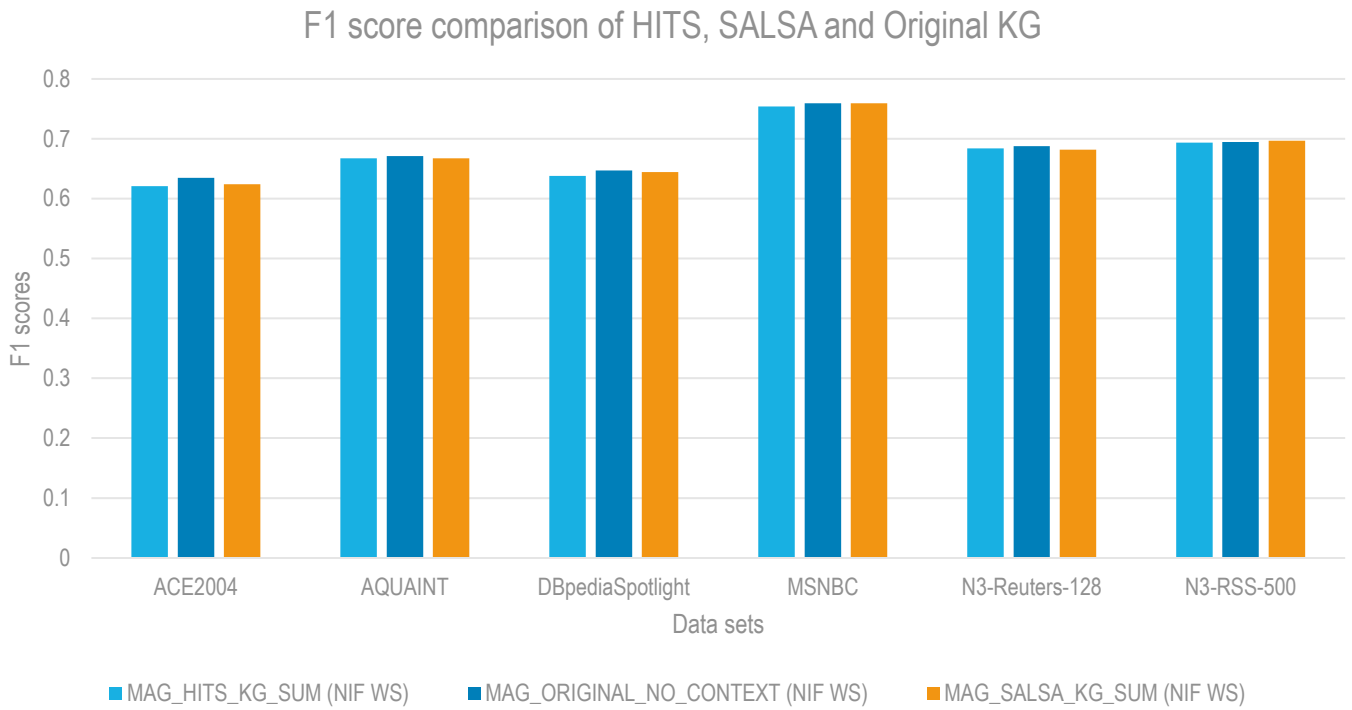
UNIVERSITÄT PADERBORN

HITS and SALSA - Evaluation

- **Evaluation was based on entity linking.**
  - E.g.: Angelina and her ex-husband Brad never played together in movie with her father Jon.
- **For evaluation purpose Indexes were created using summarized KG.**
- **AGDISTIS/MAG was used for purpose of our evaluation.**
- **We used original KG and summarized KG from SALSA, HITS as knowledge base and evaluated the results.**
- **Gerbil – Evaluation Framework.**

UNIVERSITÄT PADERBORN

HITS and SALSA (Continued….)



UNIVERSITÄT PADERBORN

Entity Summarization

- **The Idea:** There are thousands of triples describing many entities, many of which are not relevant.
- **Solution:** Summarization of those single entities (Target entity).
- **Applications:** Knowledge Graph Panels in Search Engines and Info boxes.
- **Two variants Possible:** Diversity-Centered and Relevance-Oriented.
- **LinkSUM:** a Relevance-Oriented link-based approach. **(Thalhammer et al., 2016)**

UNIVERSITÄT PADERBORN

LinkSUM - Approach

- **Stage 1: Resource Selection:**
- **Combination of Two link-measures:**
  - **PageRank**: one that accounts for the importance of the connected resource.

$$pr(r_0) = (1-d) + d \cdot \sum_{r_n \in \{r | link(r, r_0); r \in R\}} pr(r_n)/c(r_n)$$

  - **Backlink:** one that accounts for the strength of the connection.

$$bl(e) = \{r | link(r, e) \wedge link(e, r) \wedge r \in res(e), r \in R\}$$

  - Therefore, **Combined Score:**

$$score(e, r) = \alpha \cdot \frac{pr(r)}{max\{pr(a) : a \in res(e)\}} + (1-\alpha) \cdot \mathbf{1}_{bl(e)}(r)$$

UNIVERSITÄT PADERBORN

LinkSUM - Approach

- **Stage 2: Predicate Selection:**
  - In a KG, two resources can be linked through multiple semantic connections. However, in many cases, one relation is more relevant than others.

  - Could have three factors:
    - **FRQ**: The relation that is used the most is selected.
    - **EXC**: For both resources, target entity e and related resource r, we add up the number of times the relation is used with each (N+M). We use the inverse of this number 1/(N + M), in order to get the exclusivity score (the more exclusive, the better).
    - **DSC**: The sum |labels|+|ranges|+|domains| forms a basic method for estimating the quality of the description of the predicate.

UNIVERSITÄT PADERBORN

LinkSUM - Evaluation

- **Evaluated against the work “ESBM Benchmark (v1.2)”.**
- **Contained 125 Dbpedia entities over 6 different classes.**
- **6 gold standard summarizations provided by the experts.**
- **Some Considerations done:**
  - Changes from the work “LinkSUM”:
    - only outgoing links were considered in LinkSUM, we consider incoming links as well.
    - relations such as "purl.org/dc/terms/subject" were not considered in LinkSUM.
  - Changes from the work “ESBM”:
    - Relations such as the following which were included in significant numbers in ESBM, we thought otherwise:
      - http://xmlns.com/foaf/0.1/name
      - http://xmlns.com/foaf/0.1/homepage
      - http://xmlns.com/foaf/0.1/depiction
      - http://dbpedia.org/ontology/thumbnail
      - http://dbpedia.org/ontology/termPeriod

UNIVERSITÄT PADERBORN

LinkSUM - Evaluation

		FRQ mode	EXC mode	DSC mode		RELIN (Cheng et.al, 2011)	DIVERSUM (Sydow et.al.,2013)	FACES-E (Gunaratna et al., 2016)	LinkSUM	BAFREC (Kroll,H et al., 2018)	KAFCA (Kim ,E.K., et al., 2018)	MPSUM (Wei,D., et al., 2011)
Dbpedia	K = 5	0.237*	0.230	0.220		0.242	0.249	0.280	0.287	0.335	0.314	0.314
	K = 10	0.377	0.369	0.355		0.455	0.507	0.488	0.486	0.503	0.509	0.512

* Scores are F1 Measures

UNIVERSITÄT PADERBORN

References:

- **Thalhammer et al., 2016.** LinkSUM: using link analysis to summarize entity data. A Thalhammer, N Lasierra, A Rettinger - International Conference on Web Engineering, 2016.
- **Ehrlinger et al., 2016.** Towards a Definition of Knowledge Graphs. L Ehrlinger, W Wöß - SEMANTiCS (Posters, Demos, SuCCESS), 2016.
- **Cheng et al, 2011.** Cheng, G., Tran, T., Qu, Y.: RELIN: relatedness and informativeness-based cen-trality for entity summarization. In: ISWC 2011, Part I. pp. 114–129 (2011).https://doi.org/10.1007/978-3-642-25073-6 8.
- **Sydow et al, 2013.** Sydow, M., Pikula, M., Schenkel, R.: The notion of diversity in graphical entity summarisation on semantic knowledge graphs. J. Intell. Inf. Syst. 41(2), 109–149 (2013).https://doi.org/10.1007/s10844-013-0239-6.
- **Gunaratna et al., 2016**. Gunaratna, K., Thirunarayan, K., Sheth, A.P., Cheng, G.: Gleaning types for lit-erals in RDF triples with application to entity summarization. In: ESWC 2016.pp. 85–100 (2016). https://doi.org/10.1007/978-3-319-34129-3 6.

UNIVERSITÄT PADERBORN

References:

- **Kroll,H et al., 2018.** Kroll, H., Nagel, D., Balke, W.T.: BAFREC: Balancing frequency and rarity for entity characterization in linked open data. In: EYRE 2018 (2018).

- **Kim ,E.K., et al., 2018.** Kim, E.K., Choi, K.S.: Entity summarization based on formal concept analysis. In: EYRE 2018 (2018).

- **Wei,D., et al., 2011.** Wei, D., Gao, S., Liu, Y., Liu, Z., Huang, L.: MPSUM: Entity summarization with predicate-based matching. In: EYRE 2018 (2018).
- Lempel, Ronny, and Shlomo Moran. "SALSA: the stochastic approach for link-structure analysis." ACM Transactions on Information Systems (TOIS) 19.2 (2001): 131-160.
- J. M. Kleinberg. Authoritative sources in ahyperlinked environment. Journal of the ACM,46(5):604–632, 1999.

UNIVERSITÄT PADERBORN

Thank You!