Knowledge-based Graph Document Modeling

- Michael Schuhmacher, Simone Paolo Ponzetto
Idea: Modeling Documents as Graph in DBpedia for Computing Document Similarity

Bob Dylan played [...] right before Johnny Cash.

Doc A

Desire, a key folk music album from the 70’s, is mostly known for Mozambique.

Doc B
Idea: Modeling Documents as Graph in DBpedia for Computing Document Similarity

Bob Dylan played [...] right before Johnny Cash.

Doc A

Desire, a key folk music album from the 70’s, is mostly known for Mozambique.

Doc B

Entity Linker
Desire, a key folk music album from the 70’s, is mostly known for Mozambique.

**Doc A**

Bob Dylan played [...] right before Johnny Cash.

**Doc B**

Idea: Modeling Documents as Graph in DBpedia for Computing Document Similarity
Idea: Modeling Documents as Graph in DBpedia for Computing Document Similarity

Bob Dylan played [...] right before Johnny Cash.

Desire, a key folk music album from the 70’s, is mostly known for Mozambique.
Idea: Modeling Documents as Graph in DBpedia for Computing Document Similarity

$$\text{sim}(\text{Doc A}, \text{Doc B})$$
Two key questions arise for computing document similar in a semantic network with multiple, labeled edges

Q1 Which properties/edge(s) to use without making assumptions about them?

Q2 How to compare subgraphs in a network with unique concepts?
Dylan played [...] right before Cash.

Graph Space Model: The general framework

1. Entity Linker
2. Graph Construction

DBpedia

3. Edge Weighting
4. Document Graph Matching

Doc Similarity

Q1
Q2
Q1: How to exploit the properties/edge?

Select most specific paths to capture similarity

- db:Bob Dylan
- db:Mozambique (Song)
- db:United States
Q1: How to exploit the properties/edge?

Select most specific paths to capture similarity
Q1: How to exploit the properties/edge?

Select most specific paths to capture similarity
Q1: How to exploit the properties/edge?

Select most specific paths to capture similarity

- **Concepts**: db:Bob Dylan, db:Desire (Bob Dylan album), db:Duluth, Minnesota, db:Mozambique (Song), db:United States
  - dbp:artist
  - dbo:album
  - dbo:country
  - dbo:birthPlace

- **Relation Specificity**

- **Information Content (IC) of Relation**
Q1: How to exploit the properties/edge?

Select most specific paths to capture similarity

\[ w_{combIC}(e) = IC(\omega_{Pred}) + IC(\omega_{Obj}) \]
Q1: How to exploit the properties/edge?

Select most specific paths to capture similarity

1. Invert weights by maximum weight

Max = 6.0
Q1: How to exploit the properties/edge?

Select most specific paths to capture similarity

1. Invert weights by maximum weight

Max = 6.0
Q1: How to exploit the properties/edge?

Select most specific paths to capture similarity

1. Invert weights by maximum weight
2. Entity Dist := Sum of cheapest path
Q1: How to exploit the properties/edge?

Evaluation on DBpedia Entity Ranking Dataset (*Hoffart’12*)

Dataset (*Hoffart’12*):
- Human-annotated Gold Standard
- 21 Task with 20 entities each

Measure:
- Pearson correlation

Weighting (combIC) improves performance
Two key questions arise for computing document similar in a semantic network with multiple, labeled edges

Q1 Which properties/edge(s) to use without making assumptions about them?

Q2 How to compare subgraphs in a network with unique concepts?
Q2: How to compare subgraphs?

Desire, a key folk music album from the 70’s, is mostly known for Mozambique.

Bob Dylan played [...] right before Johnny Cash.
Q2: How to compare subgraphs?

Desire, a key folk music album from the 70’s, is mostly known for Mozambique.

Bob Dylan played [...] right before Johnny Cash.

Doc B

Doc A

Knowledge-based Graph Document Modeling, Michael Schuhmacher and Simone Paolo Ponzetto
Q2: How to compare subgraphs?

Adapt graph edit distance to our network

Doc A: Bob Dylan played [...] right before Johnny Cash.

Doc B: Desire, a key folk music album from the 70’s, is mostly known for Mozambique.
Q2: How to compare subgraphs?

Adapt graph edit distance to our network

Bob Dylan played [...] right before Johnny Cash.

Desire, a key folk music album from the 70’s, is mostly known for Mozambique.

Max = 6.0

Doc A

Doc B

Knowledge-based Graph Document Modeling, Michael Schuhmacher and Simone Paolo Ponzetto
Q2: How to compare subgraphs?

Adapt graph edit distance to our network

**Bob Dylan** played [...] right before **Johnny Cash.**

**Desire**, a key **folk music** album from the 70’s, is mostly known for Mozambique.

Max = 6.0

**db:**Desire (Bob Dylan album)  
**db:**Folk music
Q2: How to compare subgraphs?

Adapt graph edit distance to our network

Doc A

- db:Bob Dylan
  - 1.3
  - db:Johnny Cash

Doc B

- db:Desire (Bob Dylan album)
  - 3.5
  - db:Folk music
  - 1.7
Q2: How to compare subgraphs?

Adapt graph edit distance to our network

Approx. solution for GED = Bipartite Node Matching
Q2: How to compare subgraphs?

Adapt graph edit distance to our network

Approx. solution for GED = Bipartite 1:1 Node Matching
Q2: How to compare subgraphs?

Results on Doc Similarity benchmark (Lee‘05):

- Token Cosine
- Concept Jaccard
- GED unwghted
- GED combIC
- LSA
- ESA

Matching w/ weighted DBpedia Graph is competitive
Contributions

1. A **document model** using wide-coverage **semantic network**

2. An **edge weighting scheme**

3. A **document similarity technique**
Related Papers @ WSDM’14

Today’s session:
• **WebChild: Harvesting and Organizing Commonsense Knowledge from the Web** → *Building triples from text*

Yesterday’s Recommender systems and networks session:
• Personalized Entity Recommendation: A Heterogeneous Information Network Approach → *MetaPath between Entities/Items*
Knowledge-based Graph Document Modeling

Michael Schuhmacher, Simone Paolo Ponzetto

Visit http://dws.informatik.uni-mannheim.de

Thanks to Heiner Stuckenschmidt and Christian Meilicke for supporting this work