Ranking Documents based on Relevance of Semantic Relationships

Boanerges Aleman-Meza, I. Budak Arpinar, Mustafa V. Nural, Amit P. Sheth
Goal

• Provide a ranking algorithm for documents with no structure or links between them
• Traditional methods may not work well (Pagerank etc.)
Methodology

- Annotate documents with named entities
- Exploit relationships between the query and the entities using Ontology
- Rank the relationships with the Relevance Measure
Overview: Schematic Diagram

- Semantic Annotation
  - Named Entities
- Indexing/Retrieval
  - Using UIMA
- Ranking Documents
  - Relevance Measure
Semantic Annotation

- Spotting appearances of named-entities from the ontology in documents

Relevance Measure of Entities

• Finds Relevant Neighboring Entities
• Keyword Query -> Entity Results
• Ranked by Relevance of Interconnections among Entities (a.k.a. relationships)
Determining Relevance (first try)

“Closely related entities are more relevant than distant entities”

\[ E = \{ e \mid \text{Entity } e \in \text{Document} \} \]

\[ R = \{ f \mid \text{type}(f) \in \text{user-request} \]

and distance(f, e \in E) \leq k \}

- **Good for grouping documents w.r.t. a context**
  (e.g., insider-threat)

- **Not so good for precise results**
... Measuring what is relevant

Many relationships connecting one entity ...
Few Relevant Entities

*From Many Relationships . . .

- **very few** are relevant paths
Defining Relevant Relationships

- Relevance is determined by considering:
  - type of *next* entity (from ontology)
  - *type* of connecting *relationship*
  - *direction* of the connection
  - *length* of discovered path so far
    (short paths are preferred)
... Defining Relevant Relationships

- Does the ‘ticker’ symbol of a Company make a document more relevant?
  ... yes?

- Does the ‘industry focus’ of a company make a document more relevant?
  ... no?
Relevance Measure

- **Input:** Entity e

Relevant Sequences (defined by a domain-expert)

Find: relevant neighbors of entity e

- Entity-neighborhood expansion delimited by the ‘relevant sequences’
Relevance Measure, Relevant Sequences

• Ontology of Bibliography Data
Relevance Score for a Document

1. User Input: keyword(s)

2. Keywords match a semantic-annotation
   An annotation is related to one entity \( e \) in the ontology

3. Find relevant neighborhood of entity \( e \)
   Using the populated ontology

4. Increase the score of a document w.r.t.
   the other entities in the document that belong to
   \( e \)’s relevant neighbors
   (Each neighbor’s relevance is either low, med, or high)
Evaluation

• Used SwetoDBLP as the domain ontology
  – Built from DBLP database
  – Contains more than ½ million authors and 900K publications, more than 1.5M relationships

• 150 randomly selected queries containing authors
Evaluation

Precision for top 5, 10, 15 and 20 results ordered by their precision value for display purposes
Findings from Evaluation

• Average precision for top 5, top 10 is above 77%
  • Precision for top 15 is 73%; for top 20 is 67%

• Low Recall was due to queries involving first-names that are common (unintentional input in the evaluation)
  • Examples: Philip, Anthony, Christian
Conclusions

- **Relationship-based document ranking**
  - Relevance-score is based on appearance of relevant entities to input from user
  - Does not require link-structure among documents
Conclusions

• Challenges
  - Keeping ontology up to date and of good quality
  - Make it work for unnamed entities such as events.

• Future Work
  - Usage of ontology + documents in other domains
Thank You