Ranking Archived Documents for Structured Queries on Semantic Layers

Pavlos Fafalios, Vaibhav Kasturia, Wolfgang Nejdl
fafalios@L3S.de

L3S Research Center
University of Hannover, Germany
Document Archives

- Newspaper archives
- Web archives
- Social Media archives
- ...

- Valuable sources for research in many disciplines

Semantic Models for document archives

• Semantic Layer
  • Structured data (RDF triples) describing metadata (e.g., publication date) and content-based information (e.g., mentioned entities) about the archived documents

• Benefits:
  • Advanced (entity-centric) query capabilities
  • Real-time data integration

SELECT ?article ?nylawyer WHERE {
  SERVICE <http://dbpedia.org/sparql> {
    ?nylawyer dc:subject dbc:New_York_lawyers ; dbo:birthPlace dbr:Brooklyn
  }
  ?article schema:mentions ?entity . ?entity oae:hasMatchedURI ?nylawyer
}

“Find articles of 1989 mentioning lawyers of New York born in Brooklyn”
The problem

• The query results can be numerous
• All results equally match the query
  • There is no relevance ranking!

• How to rank the results?
  • Discover (and show to the user) the most “important” documents

SELECT ?article ?nylawyer WHERE {
  SERVICE <http://dbpedia.org/sparql> {
    ?nylawyer dc:subject dbc:New_York_lawyers ; dbo:birthPlace dbr:Brooklyn
    ?article dc:date ?date FILTER(year(?date) = 1989)
    ?article schema:mentions ?entity . ?entity oae:hasMatchedURI ?nylawyer
  }
}

Result: 184 articles mentioning 44 distinct lawyers
Contributions

- We introduce and formalize the problem of “ranking archived documents for structured queries on semantic layers”

- We propose two ranking models for the problem at hand:
  - Probabilistic
  - Stochastic (Random Walk with restart)

- New ground truth dataset
  - News archive (NYT articles)
  - Queries – results – relevance judgements (+justifications)
  - Publicly available!

- Evaluation results
  - Effectiveness
  - Interesting findings
Related Work - Temporal Information Retrieval

• Surveys: Campos et al. 2015, Kanhabua et al. 2015

• Time-aware document ranking
  • Leveraging temporal expressions (Arikan et al. 2009)
  • Time-aware language model (Berberich et al. 2010)
  • Answering time-sensitive queries (Metzlere et al. 2009) (Dakka et al. 2012)
  • Diversity-aware (Singh et al. 2016)
  • Temporal archive search based on anchor texts (Holzmann et al. 2017)

Difference of our case:
• No access on the full contents of the documents
• Structured SPARQL queries, not keywords
Related Work - Ranking in Knowledge Graphs

- Survey: Roa-Valverde and Sicilia, 2014
- Ad-hoc object retrieval (Pound et al. 2010) (Tonon et al. 2012)
  - SemSearch challenge
  - TREC entity track

Difference of our case:
- The units of retrieval are textual (unstructured) documents
- Semantic Layer: special kind of Knowledge Graph containing metadata and entity annotations about a collection of textual documents
Problem Definition

SPARQL QUERY

```sparql
SELECT ?article ?nylawyer WHERE {
  SERVICE <http://dbpedia.org/sparql> {
    ?nylawyer dc:subject dbc:New_York_lawyers ;
    dbo:birthPlace dbr:Brooklyn .
    ?article schema:mentions ?entity .
    ?entity oae:hasMatchedURI ?nylawyer
  }
}
```

The query requests documents:
- Published within a specific time period, and
- Mentioning one or more entities of interest (query entities)

SEMANTIC LAYER

RANKED RESULTS

No ranking!

1. Doc_y
2. Doc_d
3. Doc_m
4. Doc_n
5. ...
Query cases

• Logical “AND” semantics:
  • The documents mention **ALL** the query entities

```
SELECT DISTINCT ?article WHERE {
  ?article dc:date ?date FILTER(year(?date) = 1990) .
  ?article oae:mentions ?entity1, ?entity2 .
  ?entity1 oae:hasMatchedURI dbr:Nelson_Mandela .
  ?entity2 oae:hasMatchedURI dbr:F._W._de_Klerk }
```

“Find articles of 1990 mentioning both Nelson Mandela and F. W. de Klerk”

• Logical “OR” semantics
  • The documents mention **AS LEAST ONE** of the query entities

```
SELECT DISTINCT ?article WHERE {
  SERVICE <http://dbpedia.org/sparql> {
    ?president dc:subject dbc:State_Presidents_of_South_Africa }
  ?article dc:date ?date FILTER(year(?date) = 1990) .
  ?article oae:mentions ?entity .
  ?entity oae:hasMatchedURI ?president }
```

“Find articles of 1990 mentioning state presidents of South Africa”
Approach

• What makes an archived document **important** given a **time period** and one or more **entities of interest**?

• Requirements:
  • Make use of the least amount of document metadata → widely applicable
  • Exploit only the contents of the semantic layer

• Aspects affecting the importance of a document to a query:
  • The **relativeness** of a document wrt the query entities
    • the document should talk about the query entities, ideally as its main topic
  • The **timeliness** of a document wrt its publication date
    • the document should have been published during an important (for the query entities) time period
  • The **relatedness** of a document wrt other important entities mentioned in it
    • The document should discuss the relation of the query entities with other important entities
Probabilistic Modeling

• Relativeness
  • The probability to pick a document based (only) on the query entities it mentions
  • Consider the frequency of the query entities in the document
  • Motivation: frequency is a classic numerical statistic reflecting how important a term (entity) is to a document (Leskovec et al. 2014)

\[
P(d|E_Q) = \frac{\text{score}_f(d, E_Q)}{\sum_{d' \in D_Q} \text{score}_f(d', E_Q)}
\]

**AND semantics**

\[
\text{score}_\land(d, E_Q) = \frac{\sum_{e \in E_Q} \text{count}(e, d)}{\sum_{e' \in \text{ents}(d)} \text{count}(e', d)}
\]

**OR semantics**

\[
\text{score}_\lor(d, E_Q) = \frac{\sum_{e \in E_Q} \text{count}(e, d)}{\sum_{e' \in \text{ents}(d)} \text{count}(e', d)} \cdot \frac{|\text{ents}(d) \cap E_Q|}{|E_Q|}
\]

Percentage of query entities discussed in the document
Probabilistic Modeling

• Timeliness
  • **The probability to pick a document based (only) on its publication date**
  • Favor time periods with large number of documents discussing about the query entities
  • **Motivation**: Considering the fraction of documents mentioning the query entities in a given time period can improve document retrieval (Singh et al. 2016)

\[
P(d|t_d) = \frac{\text{score}^t(t_d)}{\sum_{d' \in D_Q} \text{score}^t(t_{d'})}
\]

**AND semantics**

\[
\text{score}^t_{\land}(t) = \frac{|\text{docs}(t) \cap D_Q|}{|D_Q|}
\]

**OR semantics**

\[
\text{score}^t_{\lor}(t) = \frac{|\text{docs}(t) \cap D_Q|}{|D_Q|} \cdot N(E_Q, t)
\]

\[
N(E_Q, t) = \frac{\sum_{d \in \text{docs}(t) \cap D_Q} \frac{|\text{ents}(d) \cap E_Q|}{|E_Q|}}{|\text{docs}(t) \cap D_Q|}
\]

Average percentage of query entities mentioned in documents during \(t\)
Probabilistic Modeling

• Relatedness
  • The probability to pick a document based (only) on other entities mentioned in it
  • Favor entities that are co-mentioned frequently with the query entities in important time periods
  • Motivation: The co-occurrence of entities in documents of a specific time period is a strong indicator of their temporal relatedness (Zhang et al. 2016)

\[
P(d|E_{DQ}) = \frac{\sum_{e \in \text{ents}(d) \setminus E_Q} \text{score}_r(e)}{\sum_{d' \in D_Q} \sum_{e' \in \text{ents}(d') \setminus E_Q} \text{score}_r(e')}
\]

**AND semantics**

\[
\text{score}_A(e) = \text{idf}_A(e) \cdot \sum_{t \in T_Q} (\text{score}_A(t) \cdot \frac{|\text{docs}(t) \cap D_Q \cap \text{docs}(e)|}{|\text{docs}(t) \cap D_Q|})
\]

\[
\text{idf}_A(e) = 1 - \frac{|\text{docs}(e) \cap (\bigcap_{e' \in E_Q} \text{docs}(e'))|}{|\bigcap_{e' \in E_Q} \text{docs}(e')|}
\]

**OR semantics**

\[
\text{score}_V(e) = \text{idf}_V(e) N(E_Q, e) \sum_{t \in T_Q} (\text{score}_V(t) \cdot \frac{|\text{docs}(t) \cap D_Q \cap \text{docs}(e)|}{|\text{docs}(t) \cap D_Q|})
\]

\[
\text{idf}_V(e) = 1 - \frac{|\text{docs}(e) \cap (\bigcup_{e' \in E_Q} \text{docs}(e'))|}{|\bigcup_{e' \in E_Q} \text{docs}(e')|}
\]

Inverse document frequency of \(e\) (considering the entire corpus and documents that contain all the query entities)

Inverse document frequency of \(e\) (considering the entire corpus and documents that contain at least one of the query entities)

Average percentage of query entities mentioned in the documents together with \(e\) during \(t\)
Probabilistic Modeling

• Joining the models
  • Combine the different models in a single probability score

\[
P(d|E_Q, t_d, E_{DQ}) = \frac{P(d|E_Q)P(d|T_Q)P(d|E_{DQ})}{\sum_{d' \in D_Q} P(d'|E_Q)P(d'|T_Q)P(d'|E_{DQ})}
\]
Stochastic Modeling

• Random Walk on the graph (Markov chain) defined by:
  • the query entities
  • the returned documents
  • the entities mentioned in the documents

• Stochastic analysis
  • Biased PageRank-like algorithm

• Why?
  • It can be easily customized ("biased") for different types of applications
  • Allows understanding how the different aspects of the model affect the rankings
Stochastic Modeling - The State Transition Graph

- Start from a query entity:
  - Move to a doc mentioning it, or
  - Move to a related entity (co-mentioned in docs)
- From a doc:
  - Move to an entity mentioned in the doc
- From a related (no query) entity:
  - Move to a doc mentioning the entity
Stochastic Modeling - The Transition Probabilities

- From query entities to documents
  - Relativeness + Timeliness
    \[
    \text{weight}(e \rightarrow d) = \frac{\text{score}^f(d, E_Q) \cdot \text{score}^t(t_d)}{\sum_{d' \in \text{docs}(e) \cap D_Q} (\text{score}^f(d', E_Q) \cdot \text{score}^t(t_{d'}))}
    \]

- From query entities to related entities (co-mentioned in docs)
  - Relatedness
    \[
    \text{weight}(e \rightarrow e') = \frac{\text{score}^r(e')}{\sum_{e'' \in \text{out}(e)} (\text{score}^r(e''))}
    \]

- From query entities to another node (document or related entity)
  \[
  \text{weight}(e \rightarrow n) = \begin{cases} 
  p_1 \cdot \text{weight}(e \rightarrow d) & n \in D_Q \\
  (1 - p_1) \cdot \text{weight}(e \rightarrow e') & n \in E_{D_Q}
  \end{cases}
  \]

\(p_1\): probability to move to a document node.
Stochastic Modeling - The Transition Probabilities

• From documents to entities
  • Entity frequency

\[
\text{weight}(d \rightarrow e) = \frac{\text{count}(e, d)}{\sum_{e' \in \text{ents}(d)}(\text{count}(e', d))}
\]

• From no-query entities to documents
  • Entity frequency

\[
\text{weight}(e \rightarrow d) = \frac{\text{count}(e, d)}{\sum_{d' \in \text{docs}(e) \cap D_Q}(\text{count}(e, d'))}
\]
Stochastic Modeling - The Stochastic Analysis

• Random Walk with Restart:
  • Start from the query entities and either **follow an edge** or **jump back** to a query entity and start again the traversal (“restart”)

\[
r(n) = d \cdot \text{Jump}(n) + (1-d) \cdot \sum_{n' \in \text{in}(n)} \left( \text{weight}(n' \rightarrow n) \cdot r(n') \right)
\]

- **Probability to perform a restart**
- **Probability to jump to node “n”**

\[
\text{Jump}(n) = \begin{cases} 
1/|E_Q| & n \in E_Q \\
0 & n \notin E_Q 
\end{cases}
\]

- **Initial node scores:**
  - Query entities: \(1/|E_Q|\)
  - All other: zero
- **Iteratively run to convergence**
Evaluation - Ground Truth dataset

• New ground truth dataset! (due to lack of benchmark datasets for our problem)

• Semantic Layer:
  • Corpus: **New York Times**
  • Entity Linking: **Babelfy**

• 24 SPARQL queries:
  • 6 single-entity queries (Q1-Q6)
    • E.g.: articles of 1990 discussing about Nelson Mandela
  • 6 multiple-entity AND queries (Q7-Q12)
    • E.g.: articles of 1990 discussing about Nelson Mandela and F. W. de Klerk
  • 6 multiple-entity OR queries (Q13-Q18)
    • E.g.: articles of 1990 discussing about Nelson Mandela or F. W. de Klerk
  • 6 category queries (Q19-Q24)
    • E.g., articles of 1990 discussing about philanthropists
Evaluation - Ground Truth dataset

• Manual evaluation of all results (773 totally)
  • **Score 0:** The document has almost nothing to do with the query entities
  • **Score 1:** The topic of the document is not about the query entities, however the query entities are related to the document context
  • **Score 2:** The topic of the document is not about the query entities, however the query entities are important for the document context
  • **Score 3:** The topic of the document is about the query entities and discusses something important about them

Publicly available:

[http://l3s.de/~fafalios/jcdl/evaluation_dataset.zip](http://l3s.de/~fafalios/jcdl/evaluation_dataset.zip)

(semantic layer, queries, results, relevance scores + explanations)
Evaluation results

- Probabilistic modeling – all queries

<table>
<thead>
<tr>
<th>Measure</th>
<th>Random ranking</th>
<th>Relativeness [A]</th>
<th>Timeliness [B]</th>
<th>Relatedness [C]</th>
<th>[A][B]</th>
<th>[A][C]</th>
<th>[B][C]</th>
<th>[A][B][C]</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDCG@5</td>
<td>0.26</td>
<td>0.48</td>
<td>0.27</td>
<td>0.41</td>
<td>0.53</td>
<td>0.52</td>
<td>0.45</td>
<td>0.56</td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.33</td>
<td>0.52</td>
<td>0.36</td>
<td>0.50</td>
<td>0.55</td>
<td>0.56</td>
<td>0.51</td>
<td>0.58</td>
</tr>
<tr>
<td>NDCG@all</td>
<td>0.68</td>
<td>0.79</td>
<td>0.69</td>
<td>0.76</td>
<td>0.80</td>
<td>0.81</td>
<td>0.76</td>
<td>0.82</td>
</tr>
<tr>
<td>P@5</td>
<td>0.27</td>
<td>0.44</td>
<td>0.28</td>
<td>0.48</td>
<td>0.48</td>
<td>0.50</td>
<td>0.45</td>
<td>0.52</td>
</tr>
<tr>
<td>P@10</td>
<td>0.29</td>
<td>0.38</td>
<td>0.30</td>
<td>0.45</td>
<td>0.37</td>
<td>0.42</td>
<td>0.41</td>
<td>0.41</td>
</tr>
</tbody>
</table>
Evaluation results

- Probabilistic modeling

<table>
<thead>
<tr>
<th>Measure</th>
<th>[A]</th>
<th>[B]</th>
<th>[C]</th>
<th>[A][B]</th>
<th>[A][C]</th>
<th>[B][C]</th>
<th>[A][B][C]</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDCG@5</td>
<td>0.66</td>
<td>0.30</td>
<td>0.40</td>
<td>0.68</td>
<td>0.69</td>
<td>0.45</td>
<td>0.70</td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.69</td>
<td>0.38</td>
<td>0.51</td>
<td>0.70</td>
<td>0.71</td>
<td>0.51</td>
<td>0.70</td>
</tr>
<tr>
<td>NDCG@all</td>
<td>0.88</td>
<td>0.67</td>
<td>0.75</td>
<td>0.87</td>
<td>0.88</td>
<td>0.72</td>
<td>0.87</td>
</tr>
<tr>
<td>P@5</td>
<td>0.57</td>
<td>0.23</td>
<td>0.50</td>
<td>0.57</td>
<td>0.60</td>
<td>0.40</td>
<td>0.60</td>
</tr>
<tr>
<td>P@10</td>
<td>0.40</td>
<td>0.27</td>
<td>0.45</td>
<td>0.40</td>
<td>0.40</td>
<td>0.38</td>
<td>0.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure</th>
<th>[A]</th>
<th>[B]</th>
<th>[C]</th>
<th>[A][B]</th>
<th>[A][C]</th>
<th>[B][C]</th>
<th>[A][B][C]</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDCG@5</td>
<td>0.34</td>
<td>0.28</td>
<td>0.31</td>
<td>0.42</td>
<td>0.42</td>
<td>0.35</td>
<td>0.38</td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.43</td>
<td>0.33</td>
<td>0.40</td>
<td>0.46</td>
<td>0.46</td>
<td>0.45</td>
<td>0.46</td>
</tr>
<tr>
<td>NDCG@all</td>
<td>0.76</td>
<td>0.71</td>
<td>0.75</td>
<td>0.78</td>
<td>0.78</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td>P@5</td>
<td>0.43</td>
<td>0.33</td>
<td>0.30</td>
<td>0.50</td>
<td>0.50</td>
<td>0.47</td>
<td>0.43</td>
</tr>
<tr>
<td>P@10</td>
<td>0.50</td>
<td>0.32</td>
<td>0.42</td>
<td>0.48</td>
<td>0.53</td>
<td>0.47</td>
<td>0.50</td>
</tr>
</tbody>
</table>

- Very low!
  Reason: large number of disambiguation errors

<table>
<thead>
<tr>
<th>Measure</th>
<th>[A]</th>
<th>[B]</th>
<th>[C]</th>
<th>[A][B]</th>
<th>[A][C]</th>
<th>[B][C]</th>
<th>[A][B][C]</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDCG@5</td>
<td>0.68</td>
<td>0.24</td>
<td>0.44</td>
<td>0.70</td>
<td>0.72</td>
<td>0.46</td>
<td>0.71</td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.69</td>
<td>0.36</td>
<td>0.55</td>
<td>0.69</td>
<td>0.72</td>
<td>0.54</td>
<td>0.73</td>
</tr>
<tr>
<td>NDCG@all</td>
<td>0.87</td>
<td>0.69</td>
<td>0.79</td>
<td>0.87</td>
<td>0.88</td>
<td>0.79</td>
<td>0.88</td>
</tr>
<tr>
<td>P@5</td>
<td>0.60</td>
<td>0.27</td>
<td>0.57</td>
<td>0.60</td>
<td>0.63</td>
<td>0.43</td>
<td>0.63</td>
</tr>
<tr>
<td>P@10</td>
<td>0.42</td>
<td>0.30</td>
<td>0.48</td>
<td>0.40</td>
<td>0.45</td>
<td>0.42</td>
<td>0.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure</th>
<th>[A]</th>
<th>[B]</th>
<th>[C]</th>
<th>[A][B]</th>
<th>[A][C]</th>
<th>[B][C]</th>
<th>[A][B][C]</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDCG@5</td>
<td>0.22</td>
<td>0.27</td>
<td>0.48</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>0.52</td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.28</td>
<td>0.37</td>
<td>0.53</td>
<td>0.36</td>
<td>0.38</td>
<td>0.38</td>
<td>0.53</td>
</tr>
<tr>
<td>NDCG@all</td>
<td>0.66</td>
<td>0.68</td>
<td>0.78</td>
<td>0.70</td>
<td>0.71</td>
<td>0.71</td>
<td>0.77</td>
</tr>
<tr>
<td>P@5</td>
<td>0.17</td>
<td>0.27</td>
<td>0.53</td>
<td>0.23</td>
<td>0.30</td>
<td>0.30</td>
<td>0.53</td>
</tr>
<tr>
<td>P@10</td>
<td>0.20</td>
<td>0.30</td>
<td>0.45</td>
<td>0.20</td>
<td>0.28</td>
<td>0.28</td>
<td>0.37</td>
</tr>
</tbody>
</table>

[|A|]: relativeness
[|B|]: timeliness
[|C|]: relatedness
Evaluation results

- **Stochastic modeling** (using random jump probability = 0.2)

<table>
<thead>
<tr>
<th>Measure</th>
<th>$p_1=0.0$</th>
<th>$p_1=0.2$</th>
<th>$p_1=0.4$</th>
<th>$p_1=0.6$</th>
<th>$p_1=0.8$</th>
<th>$p_1=1.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDCG@5</td>
<td>0.09</td>
<td>0.27</td>
<td>0.48</td>
<td>0.63</td>
<td>0.65</td>
<td><strong>0.67</strong></td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.18</td>
<td>0.35</td>
<td>0.54</td>
<td>0.69</td>
<td>0.72</td>
<td><strong>0.74</strong></td>
</tr>
<tr>
<td>NDCG@all</td>
<td>0.60</td>
<td>0.68</td>
<td>0.77</td>
<td>0.85</td>
<td>0.87</td>
<td><strong>0.87</strong></td>
</tr>
<tr>
<td>P@5</td>
<td>0.07</td>
<td>0.23</td>
<td>0.43</td>
<td>0.47</td>
<td>0.53</td>
<td><strong>0.60</strong></td>
</tr>
<tr>
<td>P@10</td>
<td>0.12</td>
<td>0.23</td>
<td>0.33</td>
<td>0.37</td>
<td>0.42</td>
<td><strong>0.43</strong></td>
</tr>
</tbody>
</table>

**Single-entity queries**

- **Multiple-entity AND queries**

<table>
<thead>
<tr>
<th>Measure</th>
<th>$p_1=0.0$</th>
<th>$p_1=0.2$</th>
<th>$p_1=0.4$</th>
<th>$p_1=0.6$</th>
<th>$p_1=0.8$</th>
<th>$p_1=1.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDCG@5</td>
<td>0.29</td>
<td>0.29</td>
<td>0.32</td>
<td>0.31</td>
<td>0.37</td>
<td><strong>0.48</strong></td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.36</td>
<td>0.36</td>
<td>0.40</td>
<td>0.45</td>
<td>0.47</td>
<td><strong>0.52</strong></td>
</tr>
<tr>
<td>NDCG@all</td>
<td>0.72</td>
<td>0.73</td>
<td>0.74</td>
<td>0.75</td>
<td>0.78</td>
<td><strong>0.80</strong></td>
</tr>
<tr>
<td>P@5</td>
<td>0.27</td>
<td>0.27</td>
<td>0.27</td>
<td>0.30</td>
<td>0.37</td>
<td><strong>0.57</strong></td>
</tr>
<tr>
<td>P@10</td>
<td>0.30</td>
<td>0.30</td>
<td>0.38</td>
<td>0.45</td>
<td>0.47</td>
<td><strong>0.50</strong></td>
</tr>
</tbody>
</table>

**Multiple-entity AND queries**

- **Multiple-entity OR queries**

<table>
<thead>
<tr>
<th>Measure</th>
<th>$p_1=0.0$</th>
<th>$p_1=0.2$</th>
<th>$p_1=0.4$</th>
<th>$p_1=0.6$</th>
<th>$p_1=0.8$</th>
<th>$p_1=1.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDCG@5</td>
<td>0.20</td>
<td>0.30</td>
<td>0.32</td>
<td>0.38</td>
<td>0.45</td>
<td><strong>0.59</strong></td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.27</td>
<td>0.38</td>
<td>0.42</td>
<td>0.51</td>
<td>0.52</td>
<td><strong>0.65</strong></td>
</tr>
<tr>
<td>NDCG@all</td>
<td>0.66</td>
<td>0.71</td>
<td>0.72</td>
<td>0.75</td>
<td>0.76</td>
<td><strong>0.84</strong></td>
</tr>
<tr>
<td>P@5</td>
<td>0.27</td>
<td>0.33</td>
<td>0.33</td>
<td>0.37</td>
<td>0.47</td>
<td><strong>0.53</strong></td>
</tr>
<tr>
<td>P@10</td>
<td>0.30</td>
<td>0.33</td>
<td>0.35</td>
<td>0.42</td>
<td>0.42</td>
<td><strong>0.45</strong></td>
</tr>
</tbody>
</table>

**Multiple-entity OR queries**

- **Category queries**

<table>
<thead>
<tr>
<th>Measure</th>
<th>$p_1=0.0$</th>
<th>$p_1=0.2$</th>
<th>$p_1=0.4$</th>
<th>$p_1=0.6$</th>
<th>$p_1=0.8$</th>
<th>$p_1=1.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDCG@5</td>
<td>0.43</td>
<td>0.47</td>
<td><strong>0.53</strong></td>
<td>0.50</td>
<td>0.46</td>
<td>0.23</td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.48</td>
<td>0.54</td>
<td><strong>0.55</strong></td>
<td>0.52</td>
<td>0.54</td>
<td>0.36</td>
</tr>
<tr>
<td>NDCG@all</td>
<td>0.74</td>
<td>0.77</td>
<td><strong>0.79</strong></td>
<td>0.77</td>
<td>0.77</td>
<td>0.68</td>
</tr>
<tr>
<td>P@5</td>
<td>0.57</td>
<td>0.57</td>
<td><strong>0.60</strong></td>
<td>0.60</td>
<td>0.60</td>
<td>0.23</td>
</tr>
<tr>
<td>P@10</td>
<td>0.42</td>
<td><strong>0.43</strong></td>
<td>0.40</td>
<td>0.40</td>
<td><strong>0.43</strong></td>
<td>0.28</td>
</tr>
</tbody>
</table>

\( p_1 \): probability to move to a document node when being at a query-entity node

- \( p_1 = 0.0 \) → move only to related-entity nodes
- \( p_1 = 1.0 \) → move only to document nodes

**VERY LOW!**
Evaluation results

• Probabilistic vs Stochastic

<table>
<thead>
<tr>
<th>Measure</th>
<th>Random ranking</th>
<th>Relativeness [A]</th>
<th>Timeliness [B]</th>
<th>Relatedness [C]</th>
<th>[A][B]</th>
<th>[A][C]</th>
<th>[B][C]</th>
<th>[A][B][C]</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDCG@5</td>
<td>0.26</td>
<td>0.48</td>
<td>0.27</td>
<td>0.41</td>
<td>0.53</td>
<td>0.52</td>
<td>0.45</td>
<td>0.56</td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.33</td>
<td>0.52</td>
<td>0.36</td>
<td>0.50</td>
<td>0.55</td>
<td>0.56</td>
<td>0.51</td>
<td>0.58</td>
</tr>
<tr>
<td>NDCG@all</td>
<td>0.68</td>
<td>0.79</td>
<td>0.69</td>
<td>0.76</td>
<td>0.80</td>
<td>0.81</td>
<td>0.76</td>
<td>0.82</td>
</tr>
<tr>
<td>P@5</td>
<td>0.27</td>
<td>0.44</td>
<td>0.28</td>
<td>0.48</td>
<td>0.48</td>
<td>0.50</td>
<td>0.45</td>
<td>0.52</td>
</tr>
<tr>
<td>P@10</td>
<td>0.29</td>
<td>0.38</td>
<td>0.30</td>
<td><strong>0.45</strong></td>
<td>0.37</td>
<td>0.42</td>
<td>0.41</td>
<td>0.41</td>
</tr>
</tbody>
</table>

Best Stochastic

- 0.57
- 0.62
- 0.83
- 0.58
- 0.45

• For multi-entity OR queries:
  • Probabilistic model > stochastic model
  • Reason: not well-connected transition graph
    • E.g., when there are no documents mentioning all query entities

$p_1 = 0.4$ for category queries
$p_1 = 0.0$ for all other
Evaluation results - Synopsis

• **Category** queries:
  - Large number of query entities → more disambiguation errors
  - Consider **relatedness**: the association of the query entities with other entities
  - Stochastic model (with $p_1 = 0.4$) > probabilistic model

• Other query types:
  - Consider relativeness + relatedness
  - Timeliness does not seem to affect the rankings
  - Stochastic model outperforms probabilistic model for AND queries
  - Probabilistic model outperforms stochastic model for OR queries
Conclusion

• Problem: ranking archived documents for structured queries on semantic layers
• Approach: probabilistic and stochastic models that jointly consider:
  • the relativeness of the documents to the query entities
  • the timeliness of the documents
  • the temporal relatedness of other entities to the query entities
• New ground truth dataset (publicly available)
  • http://l3s.de/~fafalios/jcdl/evaluation_dataset.zip
• Evaluation results:
  • Useful insights on the effectiveness of the proposed models
  • Relatedness can limit the negative effect caused by disambiguation errors

• Future work:
  • Diversity-aware ranking methods
  • Apply the models on other types of archives (like Web or Social Media archives)
Thank you

Comments / Questions?

fafalios@L3S.de

L3S Research Center
Leibniz University of Hannover

ALEXANDRIA Project
(ERC Nr. 339233)