A Model for Ranking Entities and Its Application to Wikipedia

Gianluca Demartini,
Claudiu S. Firan, Tereza Iofciu,
Ralf Krestel, Wolfgang Nejdl
– Searching the Web you retrieve documents
– “American countries” is an entity search query
– It is easy for a classical search engine to return documents *about* American countries
– The actual countries need to be identified and extracted by the user

– Goal: is to develop a system that can find and return entities
– and not just documents on the Web.
– Formal Model for Entity Ranking
– Application Scenarios
– Wikipedia Setting
– Entity Ranking Algorithms
– Experimental Results
– Demo
– Conclusions
- Steps for finding entities:
  
  - **Indexing**
    - Entities are identified from data sources
    - Entity descriptions are built
  
  - **Searching**
    - User’s need is translated into a query
    - IRS extracts from q the *entity need*
    - IRS search the indexed entity description
– Indexing
  • Entities
  • Data Sources

“Alexandre Pato”
ID: ap12dH5a
(born in; 1989)
(playing with; acm15hDJ)
Formal Model for Entity Ranking

- Searching
  - Users' Information Need
  - Entity Ranking System

User Query
- Keywords or
- Natural Language Text

Processed Query
- List of Attribute-Value pairs

Rank Entities
- Use scoring function

Entities

\[q \left\{ (a_i, v_i) \right\} \]

\[E = \{ e^i, ..., e^f \} \]

Ranked List of Entities
Application Scenarios

– Global Entity Identifiers: OKKAM.org

– Ranking Consumer Products
– Ranking Entities on the Web
– Ranking Entities in Wikipedia
– INEX Wikipedia
  • 653338 Wikipedia articles in early 2006
  • 50 entity ranking queries
  • Relevance assessments

– Our approach uses:
  • IR search
  • Link Analysis
  • Natural Language Processing (in the query)
  • Named Entity Recognition (in the query)
Outline

– Formal Model for Entity Ranking
– Application Scenarios
– Wikipedia Setting
– **Entity Ranking Algorithms**
– Experimental Results
– Demo
– Conclusions
– INEX query
  • Keywords, Category

– Baseline
  • Search Keywords in the text and Category

– Link based
  • Links in Wikipedia are usually entities
  • Search Keywords also in anchor text of outLinks

– Synonyms and Related Words
  • Query extension: synonyms of nouns in the Keywords + Word Sense Disambiguation for the correct meaning
Entity Ranking Algorithms

– Core Characteristics
  • Clean the Keywords removing terms (and synonyms) appearing in Category
  • Keep only nouns and adjectives in Keywords

– Named Entities
  • Use only NE (i.e., organizations, locations, persons) from Keywords
<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Tom Hanks movies where he plays a leading role.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category</strong></td>
<td>Films</td>
</tr>
<tr>
<td><strong>Synonyms</strong></td>
<td>Tom &quot;Uncle Tom&quot; Hanks &quot;Thomas J. Hanks&quot; movies film flick &quot;motion picture&quot; &quot;motion-picture show&quot; &quot;moving picture&quot; pic picture &quot;picture show&quot; &quot;moving-picture show&quot; where he plays a leading role</td>
</tr>
<tr>
<td><strong>Related Words</strong></td>
<td>Synonyms plus 50 additional concepts related mainly to motion pictures</td>
</tr>
<tr>
<td><strong>Core Characteristics</strong></td>
<td>Tom Hanks leading role</td>
</tr>
<tr>
<td><strong>Named Entities</strong></td>
<td>Tom Hanks</td>
</tr>
</tbody>
</table>
Experimental Results

– INEX XER 2007 benchmark
– MAP, P10 metrics

– Baseline

<table>
<thead>
<tr>
<th>Nr</th>
<th>Method</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{text;Keywords}</td>
<td>0.19</td>
</tr>
</tbody>
</table>

– outLinks

<table>
<thead>
<tr>
<th>Nr</th>
<th>Method</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>{text;Keywords}; {outLinks;Keywords}</td>
<td>0.23</td>
</tr>
<tr>
<td>3</td>
<td>{text;Keywords}; {outLinks;CC(Keywords)}</td>
<td>0.26*</td>
</tr>
<tr>
<td>4</td>
<td>{text;Keywords}; {outLinks;NE(Keywords)}</td>
<td>0.24</td>
</tr>
</tbody>
</table>
## Experimental Results

### Synonyms (SY) and Related Words (RW)

<table>
<thead>
<tr>
<th>Nr</th>
<th>Method</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>{text;Keywords U SY(Keywords)}</td>
<td>0.23</td>
</tr>
<tr>
<td>6</td>
<td>{text;Keywords U RW(Keywords)}</td>
<td>0.20</td>
</tr>
</tbody>
</table>

### Core Characteristics (CC) and Named Entities (NE)

<table>
<thead>
<tr>
<th>Nr</th>
<th>Method</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>{text;Keywords U CC(Keywords)}</td>
<td>0.23*</td>
</tr>
<tr>
<td>8</td>
<td>{text;Keywords U NE(Keywords)}</td>
<td>0.23*</td>
</tr>
</tbody>
</table>
## Experimental Results

### – Combinations

<table>
<thead>
<tr>
<th>Nr</th>
<th>Method</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td><code>{text;Keywords U SY(Keywords) U RW(Keywords) U CC(Keywords) U NE(Keywords)}</code></td>
<td>0.28*</td>
</tr>
<tr>
<td>10</td>
<td><code>{text;Keywords U SY(Keywords) U RW(Keywords) U CC(Keywords) U NE(Keywords) }; {outLinks;CC(Keywords)}</code></td>
<td>0.29*</td>
</tr>
</tbody>
</table>
– American countries
– http://okkam.l3s.uni-hannover.de:8081/er08web/
– http://search.yahoo.com
Conclusions

– We presented a model for Entity Ranking
– The model can be applied to different scenarios
– We applied it to Wikipedia
– We defined algorithms for the Wikipedia context

– Results show that:
  • Combining Links, NLP, NER techniques we achieve 35% (MAP) and 53% (P10) improvement over normal search
  • Effectiveness is overall low: young research area

– Next we will focus on searching the Web of Entities
Thanks