A Vector Space Model for Ranking Entities and Its Application to Expert Search

Gianluca Demartini, Julien Gaugaz, and Wolfgang Nejdl
L3S Research Center
Many users search for specific entities instead of just any type of documents
  – In the web (find Harrison Ford movies)
  – In the desktop (find e-mail address of Mike)
  – In the enterprise (find an expert on IR)

Goal: going beyond document search
Example INEX XER 2008 Topics

• Countries that have hosted FIFA Football World Cup tournaments: *countries; football world cup*

• Formula 1 drivers that won the Monaco Grand Prix: *racecar drivers; formula one drivers*

• Italian nobel prize winners: *nobel laureates*

  ...

Many examples on
http://www.ins.cwi.nl/projects/inex-xer/topics/
• A general model for ranking entities in a document collection
  – Allowing integration of known techniques
  – For any type of entity

• An application to the expert finding task
• The model for Entity Ranking
  – Basic Model
  – Extensions for including several evidences
• Application to Expert Search
  – Adaptation of the model
  – Experimental proof of concept
• Conclusions
The Model

• Documents $D = d_1, \ldots, d_m$
• Entities $E = e_1, \ldots, e_n$
• Topics $T = t_1, \ldots, t_l$
• Query $q$

• Rank $e_i \in E$ by degree of relevance to $q$
• Documents as vectors in the VS
  \[ d_i = d_{1,i}t_1 + \ldots + d_{l,i}t_l \]

• Relationship between documents and entities
  \[ f : D \times E \rightarrow R : (d_i, e_j) \rightarrow r_{ij} \]
$e_j = \sum_{k=1}^{l} \left( \sum_{i=1}^{m} d_{k,i} r_{i,j} \right) t_k$
• Query \( q = q_1 t_1 + \ldots + q_n t_n \)

• Cosine similarity

\[
\text{sim}(q, v) = \frac{q \cdot v}{\|q\| \|v\|}
\]

– Where \( v \in \{d_i, e_j\} \)
• Document dependent
  
  $E = D \times (\text{diag}(x) \times R)$
  
  – diag(x) is m x m with $x_{ii}$ is the weight for $d_i$

• Entity and Topic dependent
  
  $E' = E \circ W$
  
  – W is l x n with $w_{kj}$ is weight for $e_j$ on $t_k$

• Entity dependent
  
  $E'' = E' \times \text{diag}(cf)$
  
  – diag(cf) is n x n and $cf_{jj}$ is the cost of $e_j$
• The model for Entity Ranking
  – Basic Model
  – Extensions for including several evidences
• Application to Expert Search
  – Adaptation of the model
  – Experimental proof of concept
• Conclusions
• We adapt the model to Expert Search task
  – We fix the entity type to people
  – The query describes desired expertise

• TRECent 2006
  – W3C web sites
  – 300k documents
  – 1092 (official) candidate experts
• Cosine sim does not favour long documents
• We should favour experts with more expertise

\[
\text{projSim}(q, v) = \cos \theta \|v\|
\]

• The longer the expert vector the higher sim
• Projection similarity for Expert Search

• Explore
  – document dependent extensions
  – different space dimensions
  – relationships

• Pruning (most frequent k basis)
  – for efficiency
$E = D \times (\text{diag}(x) \times R)$
### Vector Space Dimensions

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Term</th>
<th>LSA</th>
<th>LexComp</th>
<th>LexComp Pruned</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP ($p$-value)</td>
<td>0.3370</td>
<td>0.0894 ($p = 0.0$)</td>
<td>0.3586 ($p = 0.5927$)</td>
<td>0.3625 ($p = 0.5374$)</td>
</tr>
</tbody>
</table>

On the pruned dimensions: \{ *adjective? noun+* \}
## Relationship weights

<table>
<thead>
<tr>
<th>Author/Text weights</th>
<th>1/0</th>
<th>1/0.1</th>
<th>1/0.25</th>
<th>1/0.5</th>
<th>1/0.75</th>
<th>1/1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.2246</td>
<td>0.3149</td>
<td>0.3306</td>
<td>0.3378</td>
<td>0.3365</td>
<td>0.3370</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0</td>
<td>0.0183</td>
<td>0.1559</td>
<td>0.6803</td>
<td>0.5528</td>
<td>1</td>
</tr>
</tbody>
</table>
## Pruning

<table>
<thead>
<tr>
<th></th>
<th>Pruned</th>
<th>Not Pruned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only Letters</td>
<td>0.3370</td>
<td>0.3854 (p = 0.0091)</td>
</tr>
<tr>
<td>All Chars</td>
<td>0.3716 (p = 0.0112)</td>
<td>0.4024 (p = 0.0035)</td>
</tr>
</tbody>
</table>
Related Work

• Entity Search
  – Link structure [Pehcevski et al. ECIR08]
  – Ontology based [Demartini et al. WISE08]
  – Model + NLP [Demartini et al. LA-WEB08]
Related Work

• Expert Finding
  – P@noptic Expert [Craswell et al. Ausweb01]
  – Balog’s model 1 [Balog et al. SIGIR06]
  – Voting Model [Macdonald and Ounis CIKM06, ECIR07, ECIR08]
  – Experise evidence [Macdonald et al. ECIR08]
  – Topic drift: ProjSim allows multiple expertises
Conclusions

• We presented a model for Entity Ranking
  – It is based on the VSM
  – Can be applied where entities are available
  – Can be extended with different types of evidence

• We applied to the task of Expert Finding
  – By use of a custom similarity measure
  – Exploring different extensions

• Next steps:
  – Perform the Entity Ranking task in a web collection
Thank you

• Questions
  – demartini@L3S.de