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

Gianluca Demartini, Julien Gaugaz,
and Wolfgang Nejdl

L3S Research Center

Entity Ranking

- 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/

Our Contribution

- 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

Outline

- 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

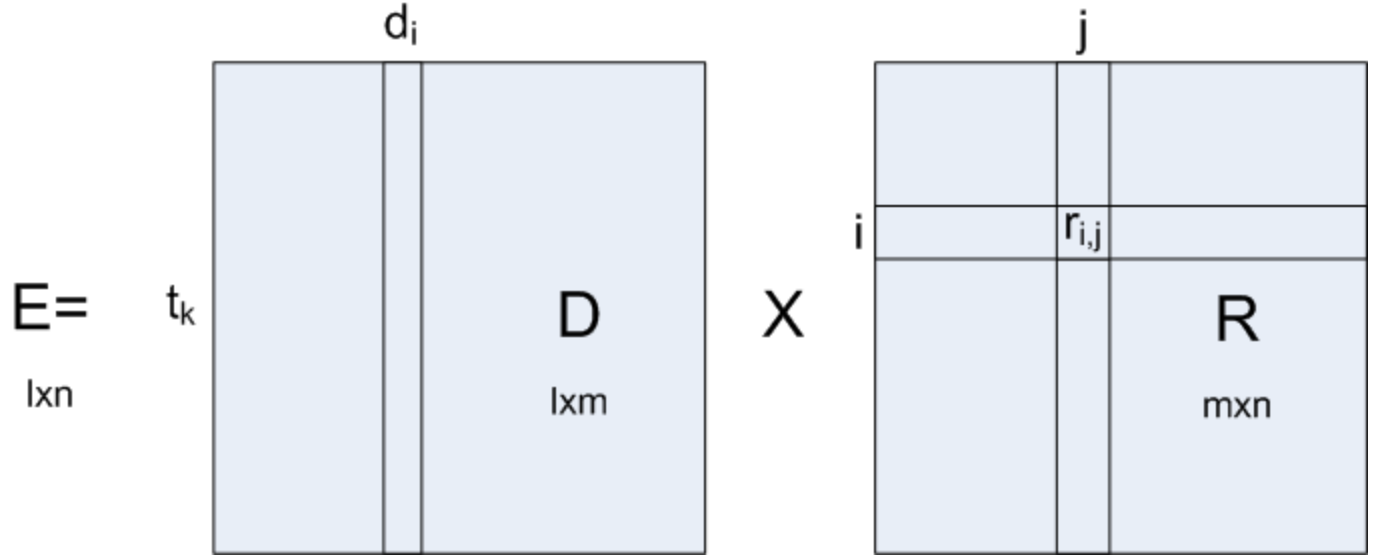
- Rank $e_i \in E$ by degree of relevance to q

Documents as vectors in the VS

- Documents as vectors in the VS
  - $d_i = d_{1,i}t_1 + \ldots + d_{l,i}t_l$
- Relationship between documents and entities
  - $f : D \times E \rightarrow R : (d_i, e_j) \rightarrow r_{ij}$

Entities as vectors in the VS

- $e_j = \sum_{k=1}^{l} \left(\sum_{i=1}^{m} d_{k,i} r_{i,j} \right) t_k$



Query

- Query $q=q_1t_1+\ldots+q_nt_n$
- Cosine similarity $$sim(q,v)=\frac{q\cdot v}{\|q\|\|v\|}$$
  - Where $v\in\{d_i,e_j\}$

Extensions

- Document dependent
  - $E = D \times (diag(x) \times R)$
  - 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
  - $E'' = E' \times diag(cf)$
  - diag(cf) is n x n and cf_{jj} is the cost of e_j

Outline

- 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

An application: Expert Search

- We adapt the model to Expert Search task
  - We fix the entity type to people
  - The query describes desidered expertise

- TRECent 2006
  - W3C web sites
  - 300k documents
  - 1092 (official) candidate experts

Projection Similarity

- Cosine sim does not favour long documents
- We should favoud experts with more expertise

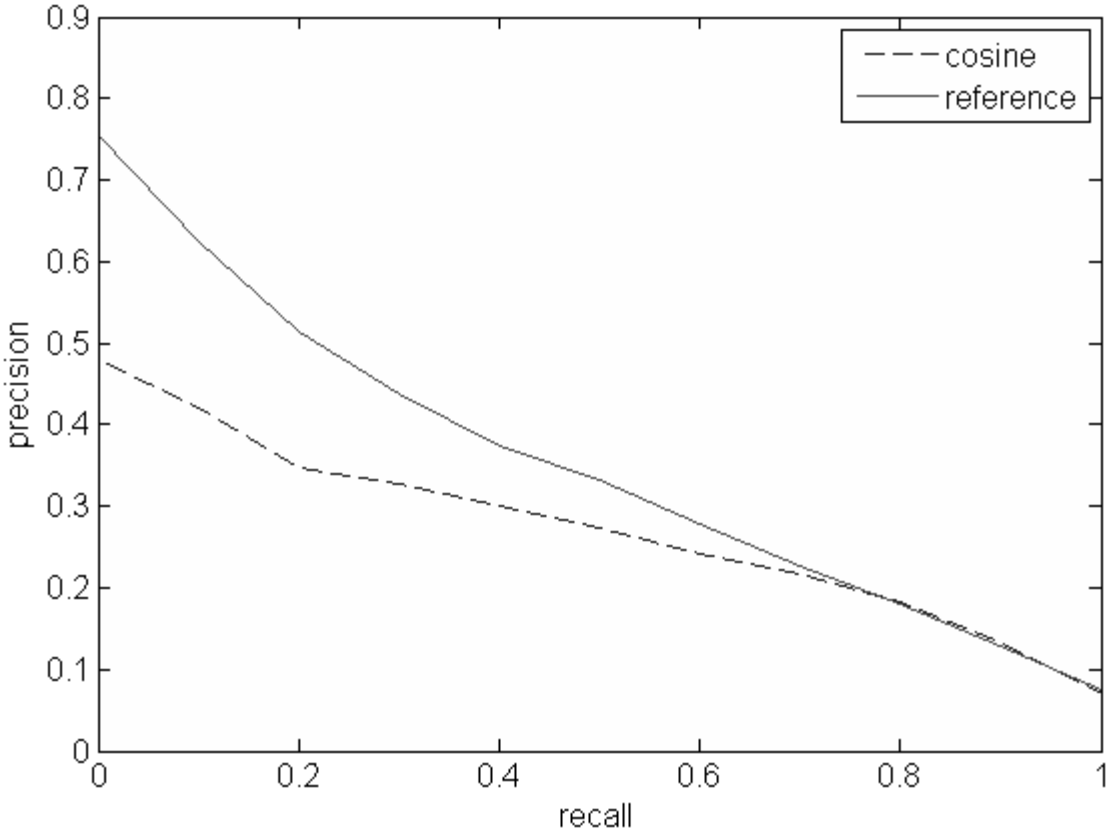
$$projSim(q, v) = \cos\theta \|v\|$$

- The longer the expert vector the higher sim

Experiments

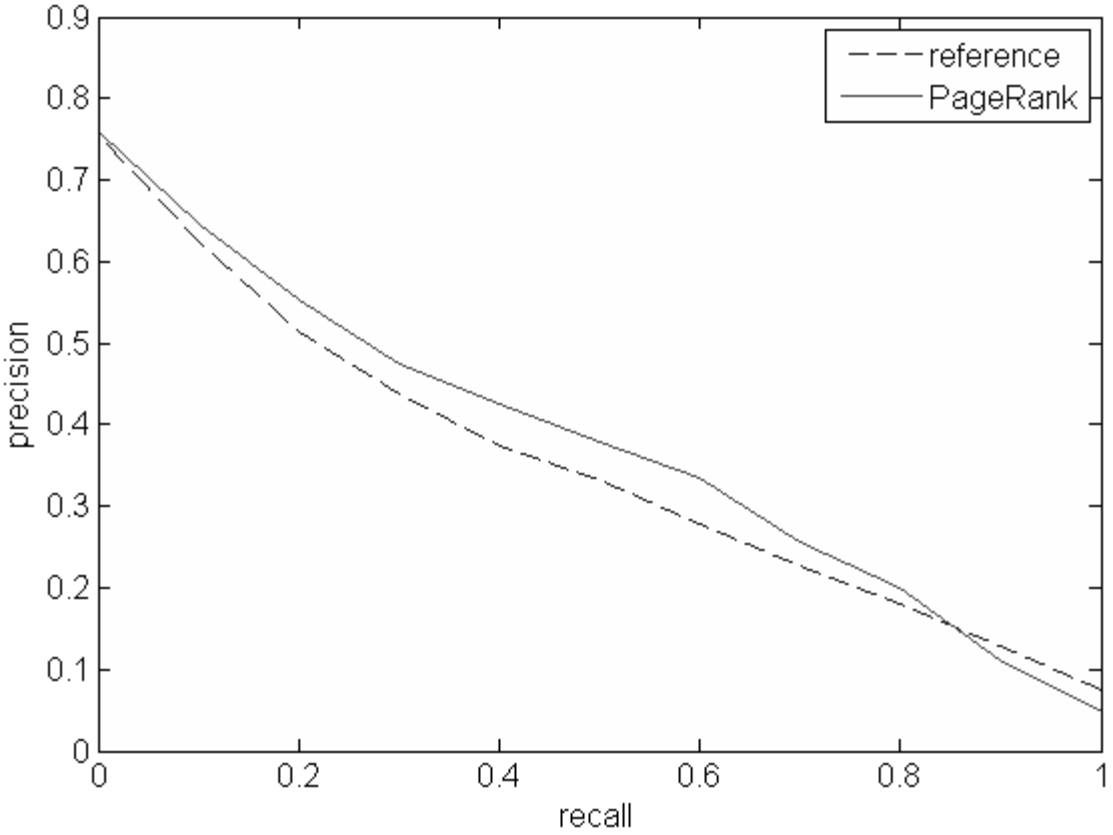
- Projection similarity for Expert Search
- Explore
  - document dependent extensions
  - different space dimensions
  - relationships
- Pruning (most frequent k basis)
  - for efficiency

ProjSim vs CosineSim



Document dependent extension

$$E = D \times (diag(x) \times R)$$



Vector Space Dimensions

Dimension	Term	LSA	LexComp	LexComp Pruned
MAP (p-value)	0.3370	0.0894 (p =0.0)	0.3586 (p =0.5927)	0.3625 (p =0.5374)

On the pruned dimensions $\{ adjective?\ noun+ \}$

Relationship weights

Author/Text weights	1/0	1/0.1	1/0.25	1/0.5	1/0.75	1/1
MAP	0.2246	0.3149	0.3306	0.3378	0.3365	0.3370
p-value	0.0	0.0183	0.1559	0.6803	0.5528	1

Pruning

	Pruned	Not Pruned
Only Letters	0.3370	0.3854 (p =0.0091)
All Chars	0.3716 (p =0.0112)	0.4024 (p =0.0035)

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