

Non-Local Evidence for Expert Finding

Krisztian Balog and Maarten de Rijke



ACM 17th Conference on Information and Knowledge Management (CIKM 2008)
Napa Valley, California Oct 26 - 30, 2008

Non-Local Evidence for Expert Finding

- Task

- Find the right person with the appropriate skills and knowledge
- Given a topic, rank expert candidates

Non-Local Evidence for Expert Finding

- Existing approaches to expert finding
- Compute associations between candidates and topics, based on their co-occurrence in
 - documents
 - text-snippets

Non-Local Evidence for Expert Finding

- Our aim
- Identify and integrate non-local sources of evidence into existing expert finding models
- Evidence that is not available from an individual page or text snippet

Outline

- Retrieval model
- Experimental setting
- Identifying and integrating non-local evidence
- Results
- Conclusions

Retrieval Model

- The problem of experts finding is stated as:
- What is the probability of a candidate ca being an expert given the query topic q ?

$$p(ca|q) = \frac{p(q|ca) \cdot p(ca)}{p(q)}$$

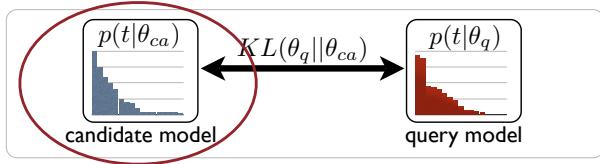
$$p(ca|q) \propto \underbrace{p(q|ca)}_{\text{How likely the candidate would produce the query}} \cdot \underbrace{p(ca)}_{\text{The } a \text{ priori belief that the candidate is an expert}}$$

How likely the candidate would produce the query

The *a priori* belief that the candidate is an expert

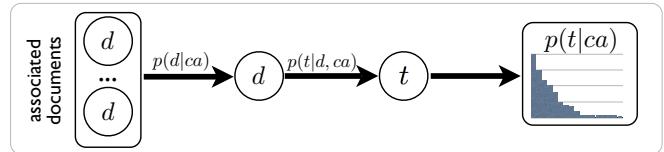
Retrieval Model (2)

- How likely the candidate would produce the query? $p(q|ca)$
- Generative language modeling approach
 - Both the candidate and the query are represented as a multinomial probability distribution over terms



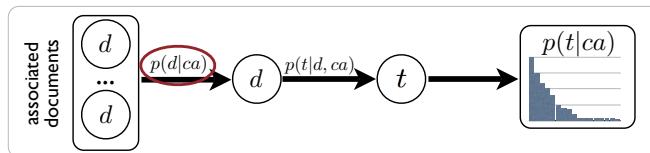
Candidate model

$$p(t|\theta_{ca}) = (1 - \lambda_{ca}) \cdot p(t|ca) + \lambda_{ca} \cdot p(t)$$



Candidate model

$$p(t|\theta_{ca}) = (1 - \lambda_{ca}) \cdot p(t|ca) + \lambda_{ca} \cdot p(t)$$



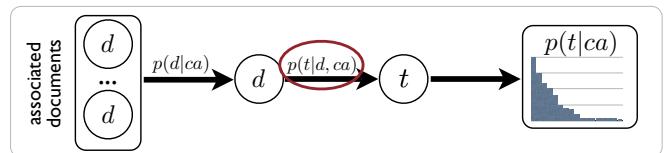
Document-candidate associations:
Boolean model

$$p(d|ca) = \begin{cases} 1, & n(ca, d) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The number of times ca is
recognized in document d

Candidate model

$$p(t|\theta_{ca}) = (1 - \lambda_{ca}) \cdot p(t|ca) + \lambda_{ca} \cdot p(t)$$



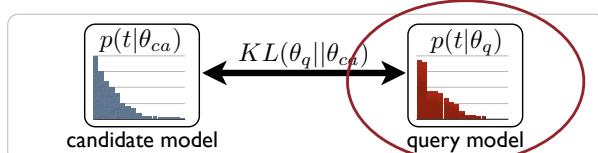
Document-based model:
Model I



Proximity-based model:
Model IB



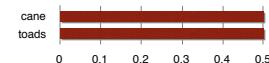
Query Model



Baseline query model (BL)

Probability mass assigned
uniformly across query terms

Example query: *cane toads*



Outline

- Retrieval model
- Experimental setting
- Identifying and integrating non-local evidence
- Results
- Conclusions

Experimental Setting

- TREC 2007 Enterprise Track
 - Document collection: web crawl of CSIRO (~370.000 docs, 4.2 GB)
 - 50 topics
 - Candidate identification
 - No canonical list is given in advance
 - E-mail addresses follow `Firstname.Lastname@csiro.au` format
 - Occurrences are replaced with a unique id

Setting the Baseline

- Boolean document-candidate associations
 - All candidates mentioned in the document are equally important, and vice versa
- Baseline query
 - All query terms are equally important
- Uniform priors
 - All candidates are equally likely to be experts

Non-Local Evidence

- Document-candidate associations
- Query model
- Candidate priors

Document–candidate Associations

- Importance of a candidate given a document $p(d|ca)$
 - So far: all candidates are equally important
 - Estimate the strength of the association based on
 - How many times the candidate is mentioned in the document
 - How many other documents the candidate is related to

Document–candidate Associations (2)

- Lean document representation
 - Document contains only candidate mentions
- Use a term weighting scheme that combines the candidate's (local) frequency in the document and its global frequency

$$p(d|ca) \propto TF.IDF(d, ca)$$

Document–candidate Associations (3)

| | ca_1 | \dots | ca_i | \dots | ca_n |
|---------|--------|---------|--------|---------|--------|
| d_1 | | | | | |
| \dots | | | | | |
| d_j | | | | | |
| \dots | | | | | |
| d_m | | | | | |

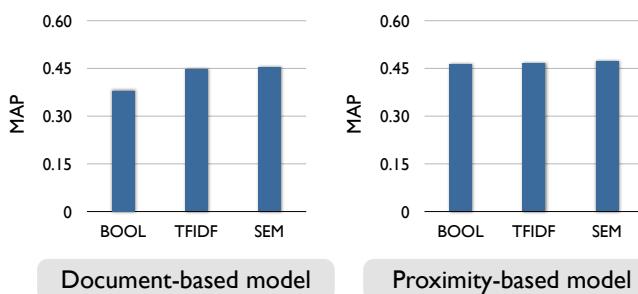
- Weight of the candidate in the document is computed in two ways
 - I) Number of occurrences $n(ca, d)$
 - 2) Semantic relatedness SEM

$$n'(ca, d) = \begin{cases} KL(\theta_{ca} || \theta_d), & n(ca, d) > 0 \\ 0, & \text{otherwise.} \end{cases}$$

TFIDF

SEM

Results



Query Models

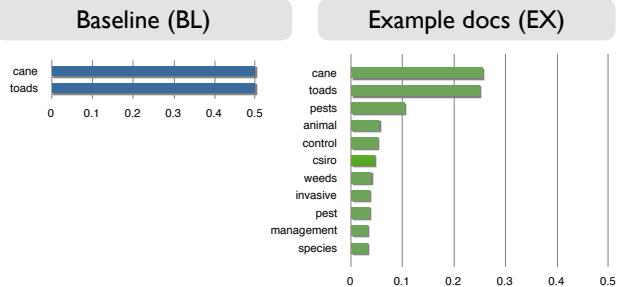
- TREC 2007 Enterprise track simulates a type of click-based system
- A few examples of key documents are provided with the topic description
- [Balog et. al., 2008] propose an effective method for constructing a query model by sampling terms from example documents

K. Balog, W. Weerkamp and M. de Rijke. A Few Examples Go A Long Way: Constructing Query Models from Elaborate Query Formulations. In: SIGIR 2008.

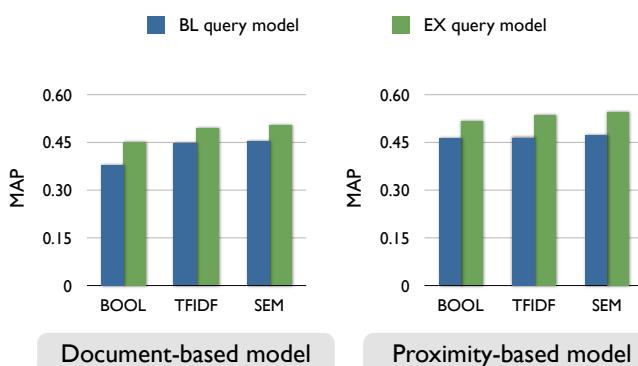
Example Topic

```
<top>
<num>CE-039</num>
<query>cane toads</query>
<narr>
Cane toads were introduced into Australia in a failed bid to control Australian native beetles. [...] Resources describing cane toads, invasive species, pest management, biological control would all be relevant to the topic.
</narr>
<page>CSIRO141-14983789</page>
<page>CSIRO139-09015831</page>
<page>CSIRO134-11651748</page>
</top>
```

Example Query Model



Results



Candidate Priors

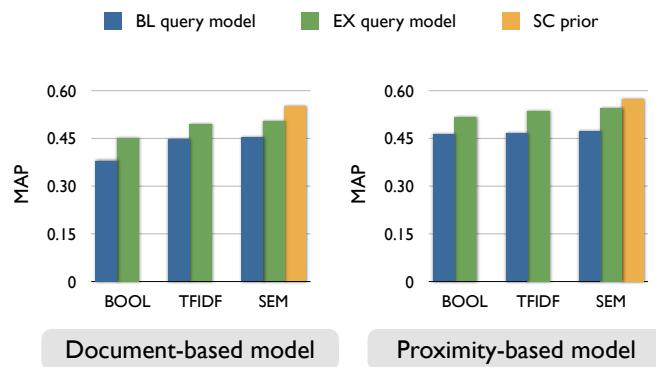
The screenshot shows the Food Science Australia website. The header includes the CSIRO logo and a search bar. The main menu has options like Home, About CSIRO, Doing Business, Media, Events, Digital & Science, Publications, Careers, CSIRO Shop, and Contact Us. The 'Food' section is highlighted. The 'OVERVIEW' section describes Food Science Australia's role in providing quality ingredients and processing technologies to support the food industry. It also mentions its consumer insights to organizations that market food products. The 'FAST FACTS' section notes it is a joint venture of CSIRO and the Victorian Government. The 'PRIMARY CONTACT' section provides details for Mr. John Smith. The 'LOCATIONS' section lists Food Science Australia - North Ryde, Riverside Corporate Park, 131 Victoria Street, North Ryde NSW 2113, Australia, PO Box 52, North Ryde NSW 1670, Australia, and Food Science Australia - Werribee, Werribee, Victoria 3030, Australia. The footer includes links to the Victorian Government and the Victorian Government's privacy statement.

Candidate priors

- Encodes organizational knowledge
- Extracted names and positions from contact boxes
- Filtering out science communicators (SC) based on position information
 - communication officer/manager/advisor
 - manager public communications

$$p(ca) = \begin{cases} 1, & ca \notin SC, \\ 0, & ca \in SC. \end{cases}$$

Results



How good is it?

| Method | Run type | MAP |
|--------------------------------|-----------|--------|
| TREC 2007 best | automatic | 0.4632 |
| TREC 2007 best | feedback | 0.3660 |
| TREC 2007 best | manual | 0.4787 |
| Voting model [1] | automatic | 0.3519 |
| Relevance prop. [2] | automatic | 0.4319 |
| Baselines in this paper | | |
| Document-based model | automatic | 0.3801 |
| Proximity-based model | automatic | 0.4633 |

[1] C. Macdonald, D. Hannah, and I. Ounis. High quality expertise evidence for expert search. In *ECIR 2008*.

[2] P. Serdyukov, H. Rode, and D. Hiemstra. Modeling relevance propagation for the expert search task. In *TREC 2007*.

How good is it? (2)

| Method | Run type | MAP |
|------------------------------|-----------|--------|
| Document-based model | | |
| Baseline | automatic | 0.3801 |
| Document-cand. assoc. | automatic | 0.4541 |
| Query model | feedback | 0.5044 |
| Candidate priors | feedback | 0.5506 |
| Proximity-based model | | |
| Baseline | automatic | 0.4633 |
| Document-cand. assoc. | automatic | 0.4735 |
| Query model | feedback | 0.5465 |
| Candidate priors | feedback | 0.5747 |

Conclusions

- Identified a number of non-local sources of evidence for expert finding
- Complemented existing document and proximity-based approaches to incorporate non-local evidence
- Showed significant improvements over a very competitive baseline
- Outperformed existing state-of-the-art

Further Work

- Non-local evidence within documents
 - Recognize and exploit internal document structure

Future Work (2)

The screenshot shows the CSIRO website's 'WATCH & LISTEN' section. A video player displays a thumbnail for a video titled 'DNA Doctor: Catalyst, ABC Interview'. The thumbnail shows two micrographs of cells: one normal and one damaged. Below the video player, there is a sidebar with navigation links for various CSIRO divisions like Agriculture & Space, Environment, Farming & Food, Health & Wellbeing, Information & Technology, Manufacturing, Mining & Minerals, Transport & Infrastructure, and Sustainable Solutions.

Non-Local Evidence for Expert Finding

K.Balog@uva.nl
<http://www.science.uva.nl/~kbalog>