

Integrating Contextual Factors into Topic-centric Retrieval Models for Finding Similar Experts

Katja Hofmann¹ Krisztian Balog¹ Toine Bogers²
Maarten de Rijke¹

¹ISLA, University of Amsterdam

²ILK, Tilburg University

Future Challenges in Expertise Retrieval
SIGIR 2008 Workshop, 24 July 2008, Singapore

Outline

- 1 Motivation
 - Expertise Seeking and Retrieval
 - Setting and Research Questions
- 2 Topic-centric Models
 - Approach
 - Results
- 3 Contextual Factors
 - Identifying Contextual Factors
 - Integration with Topic-centric Models
- 4 Summary

Outline

1 Motivation

Expertise Seeking and Retrieval
Setting and Research Questions

2 Topic-centric Models

Approach
Results

3 Contextual Factors

Identifying Contextual Factors
Integration with Topic-centric Models

4 Summary

Expertise Seeking



- More and more documents online contain evidence of expertise
- ⇒ use IR methods to find matches in documents associated with experts

Expertise Retrieval

- Current approaches based on document retrieval (e.g. TrecEnt)
 - Retrieve relevant documents (e.g. vector space model, language modeling)
 - Find experts associated with most relevant documents (e.g. author)
 - Good at finding topical matches
 - *But:* people searching for experts consider additional (contextual) factors, e.g. reliability
- ⇒ Identify contextual factors and integrate them with topic-centric approaches



Setting

- Pilot study with communication advisors at Tilburg University
 - Get requests for experts from the media
 - If the requested expert is not available recommend a similar expert
- Available tool: WebWijs
<http://www.tilburguniversity.nl/webwijs/>

Information for About Tilburg University Faculties Research

Home » Experts & Expertise

Experts & Expertise

Many researchers, scientists and support staff are working at the Tilburg University. Search for a certain individual, or within a certain expertise to find the right person.

Search experts
Which researcher or scientist are you looking for?

 Also search for support staff

Search expertise
In which field are you looking for a researcher of scientist?

Press representatives
• www.tilburguniversity.nl/ijr

Editorial staff
• webwijs@tue.nl

Experts
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Expertise
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

© 2007 Tilburg University Tel +31 (0) 13 468 9 111 - E-mail webwijs@tue.nl

 **A.M. (Toine) Bogers**
Current candidate
• a.m.bogers@tue.nl
Faculty Humanities

Expertise
Key words

- Artificial intelligence
- Computer linguistics
- Folksonomy
- Information retrieval
- Language and artificial intelligence
- Language technology
- Recommendation systems
- Search engine
- Social classification

Publications
Most recent publications

- Bekog, K., Bogers, T., Azzopardi, L., Rijke, M. de, & Bosch, A. van den (2007). Broad expertise retrieval for sparse data environments. *SIGIR'07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 551-558). New York: ACM Press. [Further information](#)
- Shoen, T., & Bekog, K. (2007). *Int'net Conference on Knowledge Discovery Research: 2007 Technical Report*

Information retrieval

- J. (Sjaak) Nouwt
- A.M. (Toine) Bogers
- B.H.M. (Bart) Cueters

See also:

- [Digital library](#)
- [Documentary information](#)
- [Fulltextsearch](#)
- [Search engine](#)

Research Questions

- RQ1 What contextual factors play a role in the specific setting?
- RQ2 Can we integrate these factors with topic-centric retrieval algorithms?
- RQ3 Can integrating contextual factors improve retrieval performance?

Outline

- 1 Motivation
 - Expertise Seeking and Retrieval
 - Setting and Research Questions
- 2 Topic-centric Models
 - Approach
 - Results
- 3 Contextual Factors
 - Identifying Contextual Factors
 - Integration with Topic-centric Models
- 4 Summary

Retrieval Task

- Finding similar experts
 - Given candidate expert e return a ranked list of similar experts f_i
- Data: UvT collection (WebWijs + public pages of UvT)
 - Database of 1168 experts
 - Publication lists
 - Topics of expertise (self-assigned by experts)

Topic-centric Similarity Measures

Table: Approaches for measuring topic-centric similarity between two experts.

method	expert representation	$sim_T(e, f)$
DOCS	set: $D(e)$	$\frac{ D(e) \cap D(f) }{ D(e) \cup D(f) }$
TERMS	vector: $\vec{t}(e)$	$\cos(\vec{t}(e), \vec{t}(f))$
TOPICS	set: $T(e)$	$\frac{ T(e) \cap T(f) }{ T(e) \cup T(f) }$

$D(e)$ Set of documents (course descriptions and publications) associated with expert e

$\vec{t}(e)$ Vector of term frequencies extracted from documents associated with e (weighted using TF.IDF)

$T(e)$ Set of topics expert e manually selected as expertise areas

Relevance Assessment

Subjects 6 communication advisors from Tilburg University

Method Printed questionnaire

Topics 44 experts that appear in the media most frequently (+12 overlapping topics)

- Relevance judgement procedure
 - 1 List similar experts (open-ended)
 - 2 Rank a list of experts

Inter-judge Agreement

- At highest rank 75%
 - Agreement at lower ranks is difficult to establish but appears lower
- ⇒ People are good at finding the one most similar expert

Topic-centric Retrieval Results

Table: Results, topic-centric similarity methods.

Method	ExCov	Jaccard	MRR	NDCG
DOCS	52%	0.1987	0.4348	0.3336
TERMS	100%	0.2143	0.2177	0.3708
TOPICS	84%	0.3129	0.4470	0.5747

ExCov Topics covered by the method

MRR Mean Reciprocal Rank, inverse of the rank of the first retrieved element

Combinations of Topic-centric Approaches

Table: Results, combinations topic-centric similarity methods.
Significant improvements over individual methods are marked with *.

Method	ExCov	Jaccard	MRR	NDCG
DOCS+TOPICS (S)	89%	0.3235	0.4529	0.5694
TERMS+TOPICS (S)	100%	0.3913	0.4789*	0.6071*
DOCS+TOPICS (R)	89%	0.3678	0.5422*	0.6064*
TERMS+TOPICS (R)	100%	0.4475	0.4317	0.6213*

(S) Linear combination by score

(R) Linear combination by rank

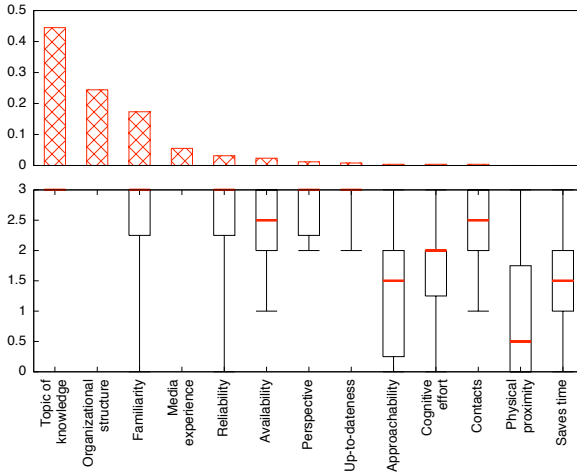
Outline

- 1 Motivation
 - Expertise Seeking and Retrieval
 - Setting and Research Questions
- 2 Topic-centric Models
 - Approach
 - Results
- 3 Contextual Factors**
 - Identifying Contextual Factors
 - Integration with Topic-centric Models
- 4 Summary

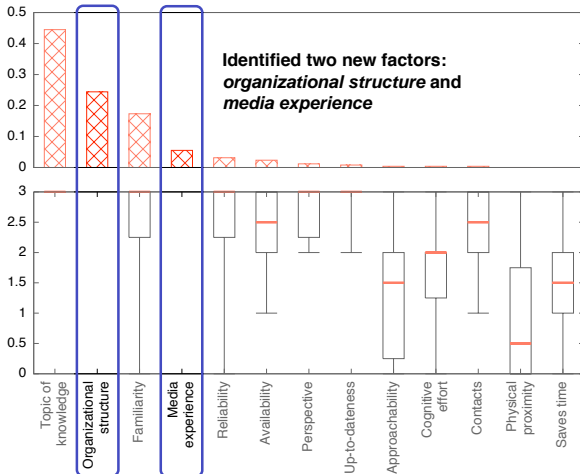
Data Collection

- Collect additional data with relevance assessments
 - For every topic: ask why decision was made (open-ended)
 - At the end: ask subjects to rate factors that they think play a role when they recommend experts (factors from [Woudstra and van den Hooff(2007)])
- Analysis
 - Open-ended responses coded according to existing coding scheme
 - Compare frequencies of mentioned factors to explicit ratings

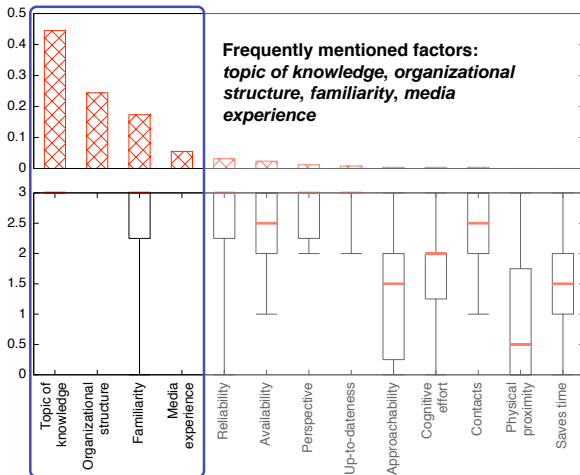
Contextual Factors for Finding Similar Experts



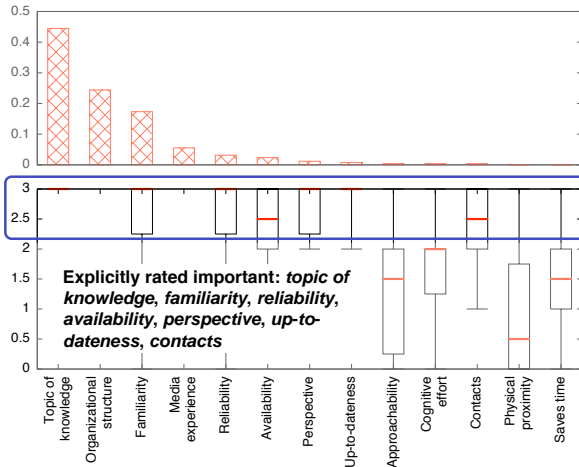
Contextual Factors for Finding Similar Experts



Contextual Factors for Finding Similar Experts



Contextual Factors for Finding Similar Experts



Modeling Contextual Factors

- 6 factors can be modeled with available data
- *Organizational structure* applied as a filter (only recommend people from the same faculty)
- 5 remaining factors
 - Modeled as prior $p(f)$
 - Linear combination with topic-centric model

$$\text{sim}(e, f) = p(f) \cdot \text{sim}_T(e, f)$$

Table: Contextual factors modeled as priors

Factor	$p(f)$ based on
<i>Media Experience</i>	Frequency of appearances in the media
<i>Reliability</i>	Number of publications
<i>Position</i>	Position within the organization
<i>Up-to-dateness</i>	Number of recent publications
<i>Contacts</i>	Number of co-authors

Retrieval Results: Baseline 1

Table: Combination of topic-centric similarity methods and contextual factors. Significant differences against the baseline are marked with *.

Method	ExCov	MRR
BASELINE 1: TERMS+TOPICS(R)	100%	0.4317
Media experience	100%	0.4749
Reliability	100%	0.5105*
Position	100%	0.4317
Up-to-dateness	100%	0.5123*
Contacts	100%	0.4517
Organizational structure	98%	0.4604*

Retrieval Results: Baseline 2

Table: Results, combination of contextual factors and content-based similarity methods.

Method	ExCov	MRR
BASELINE 2: DOCS+TOPICS (R)	89%	0.5422
Media experience	89%	0.4989
Reliability	89%	0.5801
Position	89%	0.5422
Up-to-dateness	89%	0.5823
Contacts	89%	0.5557
Organizational structure	89%	0.5393

Outline

- 1 Motivation
 - Expertise Seeking and Retrieval
 - Setting and Research Questions
- 2 Topic-centric Models
 - Approach
 - Results
- 3 Contextual Factors
 - Identifying Contextual Factors
 - Integration with Topic-centric Models
- 4 Summary

Summary

- Contextual factors such as *reliability* and *up-to-dateness* play a role in finding similar experts in the studied setting
- Extending topic-centric approaches with contextual factors can significantly improve retrieval performance
- Outlook
 - Study other tasks and settings
 - Develop more principled approaches for integrating contextual factors with topic-centric models

References I



K. Balog and M. de Rijke.

Finding similar experts.

In *SIGIR '07*, pages 821–822. ACM Press, 2007.



K. Balog, L. Azzopardi, and M. de Rijke.

Formal models for expert finding in enterprise corpora.

In *SIGIR '06*, pages 43–50. ACM Press, 2006.



L. Woudstra and B. van den Hooff.

Inside the source selection process: Selection criteria for human information sources.

Information Processing and Management, 2007.

Topic-centric Similarity (All Measures)

Table: Results, topic-centric similarity methods.

Method	ExCov	Jaccard	MRR	NDCG
DOCS	52%	0.1987	0.4348	0.3336
TERMS	100%	0.2143	0.2177	0.3708
TOPICS	84%	0.3129	0.4470	0.5747

ExCov Topics covered by the method

Jaccard Measure of set overlap irrespective of rankings

MRR Mean Reciprocal Rank, inverse of the rank of the first retrieved element

NDCG Normalized Discounted Cumulative Gain, rewards retrieving highly relevant results at high ranks

Combinations of Topic-centric Methods (All Measures)

Table: Results, combinations topic-centric similarity methods.
Significant improvements over individual methods are marked with *.

Method	ExCov	Jaccard	MRR	NDCG
DOCS+TOPICS (S)	89%	0.3235	0.4529	0.5694
TERMS+TOPICS (S)	100%	0.3913	0.4789*	0.6071*
DOCS+TOPICS (R)	89%	0.3678	0.5422*	0.6064*
TERMS+TOPICS (R)	100%	0.4475	0.4317	0.6213*

(S) Linear combination by score

(R) Linear combination by rank

Coding Scheme

Table: Example statements, frequency distribution, and explicit importance ratings (0 = *no influence*, 3 = *strong influence*) of factors mentioned.

Factor (with example statements)	Frequency (total)	Frequency (# subjects)	Median rating
Topic of knowledge (“academic record”, “has little overlap with the required expertise”, “is only in one point similar to X’s expertise”, “topically, they are close”, “works in the same area”)	44.5%	100%	3.0
* Organizational structure (“position within the faculty”, “project leader of PROJECT”, “work for the same institute”)	24.4%	100%	n/a
Familiarity (“know her personally”, “I don’t know any of them”)	17.3%	83%	3.0
* Media experience (“experience with the media”, “one of them is not suitable for talking to the media”)	5.5%	33%	n/a
Reliability (“least overlap and experience”, “seniority in the area”, “is a university professor (emeritus)”)	3.1%	33%	3.0
Availability (“good alternative for X and Y who don’t work here any more”, “he is an emeritus (even though he still comes in once in a while)”)	2.4%	66%	2.5
Perspective (“judicial instead of economic angle”, “different academic orientation”)	1.2%	33%	3.0
Up-to-dateness (“recent publications”, “[he] is always up-to-date”)	0.9%	33%	3.0
Approachability (“accessibility of the person”)	0.4%	17%	1.5
Cognitive effort (“language skills”)	0.4%	17%	2.0
Contacts (“[would] walk by the program leader for suggestions”)	0.4%	17%	2.5
Physical proximity	0.0%	0%	0.5
Saves time	0.0%	0%	1.5

Modeling Contextual Factors

Media Experience	$p(f) = 1 + \log \left(1 + \sum_y media_y(f) \right)$
Reliability	$p(f) = 1 + \log(1 + \sum_y pub_y(f))$
Position	$p(f)$ manually chosen based on position
Up-to-dateness	$p(f) = 1 + \log \left(1 + \sum_y w(y) \cdot pub_y(f) \right)$
Contacts	$p(f) = 1 + \log(1 + coauth(f))$

Contextual Factors Combined with BASELINE 1

Table: Combination of contextual factors and topic-centric similarity methods. Significant differences against the baseline are marked with *.

Method	ExCov	Jaccard	MRR	NDCG
BASELINE 1: TERMS+TOPICS(R)	100%	0.4475	0.4317	0.6213
(1) Media experience	100%	0.3929	0.4749	0.5967
(2) Reliability	100%	0.3568	0.5105*	0.6113
(3) Position	100%	0.4505	0.4317	0.6222
(4) Up-to-dateness	100%	0.3689	0.5123*	0.6193
(5) Contacts	100%	0.3871	0.4517	0.5956
(6) Organizational structure	98%	0.3607	0.4604*	0.5954*
(1) + (4)	100%	0.3330	0.4831	0.5558*
(1) + (5)	100%	0.3378	0.4817	0.5517*
(4) + (5)	100%	0.3040	0.5260	0.5756*
(1) + (4) + (5)	100%	0.2754	0.5150	0.5162*
(1) + (4) + (5) + (6)	98%	0.2827	0.5034	0.5277*

Contextual Factors Combined with BASELINE 2

Table: Results, combination of contextual factors and content-based similarity methods. Significant differences against the baseline are marked with *.

Method	ExCov	Jaccard	MRR	NDCG
BASELINE 2: DOCS+TOPICS (R)	89%	0.3678	0.5422	0.6064
(1) Media experience	89%	0.3725	0.4989	0.5881
(2) Reliability	89%	0.3508	0.5801	0.6002
(3) Position	89%	0.3678	0.5422	0.6064
(4) Up-to-dateness	89%	0.3648	0.5823	0.6119
(5) Contacts	89%	0.3621	0.5557	0.5930
(6) Organizational structure	89%	0.3363	0.5393	0.5857
(4) + (5)	89%	0.3281	0.5923	0.5686*