Impact of the Query Set on the Evaluation of Expert Finding Systems

Robin Brochier[1,2] Adrien Guille[1]
Benjamin Rothan[2] Julien Velcin[1]

[1] Université de Lyon, Lyon 2, ERIC EA 3083, France
{robin.brochier,julien.velcin, adrien.guille}@univ-lyon2.fr
https://eric.ish-lyon.cnrs.fr/

[2] Digital Scientific Research Technology, Lyon, France *{robin, benjamin}@peer.us*
https://peer.us/

07/12/2018

Table of Contents

Introduction

The Usual Evaluation (topic-query)

Another Evaluation (document-query)

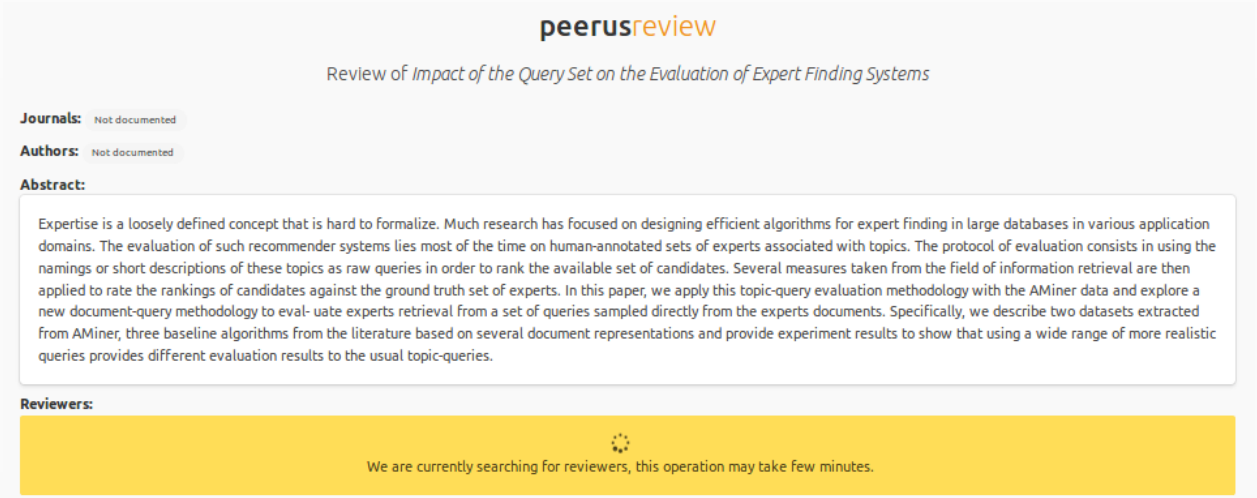
Results and Comparisons

Ongoing Works

Références

Introduction

- PhD in collaboration with Digital Scientific Research Technology [1].
- Peerus Review : a tool for publishers to find potential reviewers [2].



1. https://peer.us/
2. https://review.peer.us/

Motivations

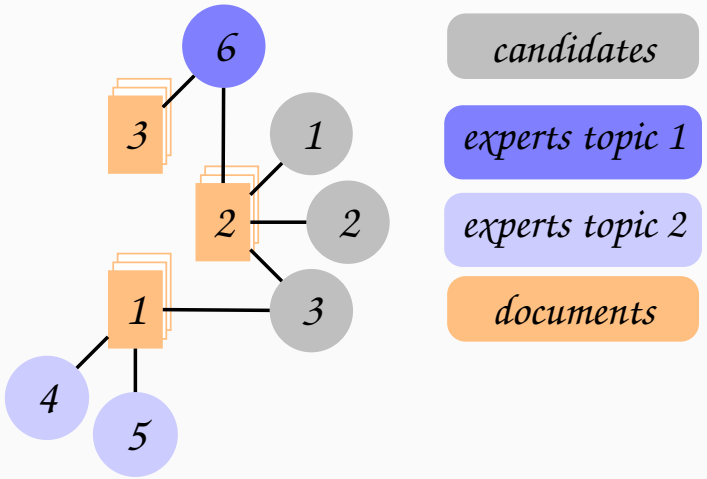
Topic-Query : "Natural Language Processing"
Document-Query : "Learning Subjective Language : Subjectivity in natural language refers to aspects of language used to express opinions, evaluations, and speculations. There are numerous natural language processing applications for which subjectivity analysis is relevant, including information extraction and text categorization..."

Topic-Query : "Backend Engineer"
Document-Query : "You write well designed, testable and reliable code as you ship features that help support amazing causes world-wide ; maintain a highly responsive system by ensuring code quality, stability and scalability ; support fellow developers by providing advice, encouraging best practices and performing design and code reviews..."

Topic-Query : "Python Tensorflow Keras"
Document-Query : "My input is a merge of datasets from few different experiments. My custom loss function needs firstly to split that data and perform different calculations in each subset, then sum results. Unfortunately, I can't split it inside custom loss. I have prepared a piece of code that works with tf.boolean_mask()..."

Is the usual topic-query evaluation statistically significant ?
Does an algorithm perform the same way for both types of queries ?

An Example of a Dataset



(a) Bipartite graph linking candidates and documents.

$$G = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}_{|D| \times |C|}$$

$$X = \begin{pmatrix} 0.4 & \dots & 0.8 \\ 0.9 & \dots & 0.1 \\ 0.3 & \dots & 0.3 \end{pmatrix}_{|D| \times |X|}$$

(b) Adjacency matrix G of the bipartite graph and the features X matrix of the documents.

Figure – Hypothetical example of a dataset for expert finding.

AMiner Dataset

AMiner provides a dataset of scientific authors and their publications (titles+abstracts).

Citation V1 [3] + Expert Finding (human-annotated) [4] :

- 532,968 candidates
- 480,630 documents
- 1,100,698 candidates-documents links
- 210 experts
- 7 topics

---

3. https://aminer.org/citation
4. https://aminer.org/lab-datasets/expertfinding/#expert-list

Usual Expert Finding Evaluation.

Algorithm 1 The usual way of evaluating.

```
Require: Ranking_Algorithm
  scores ← [ ]
  for all topics do
      candidates_ranking = Ranking_Algorithm(current_topic_textual_expression)
      current_score ← Evaluate(candidates_ranking, ground_truth_topic_experts_set)
      scores.append(current_score)
  end for
  return Mean(scores), STD(scores)
```

Topic-queries in AMiner dataset :

- "intelligent agents"
- "planning"
- "machine learning"
- "natural language processing"
- "information extraction"
- "semantic web"
- "support vector machine"

Metrics used :

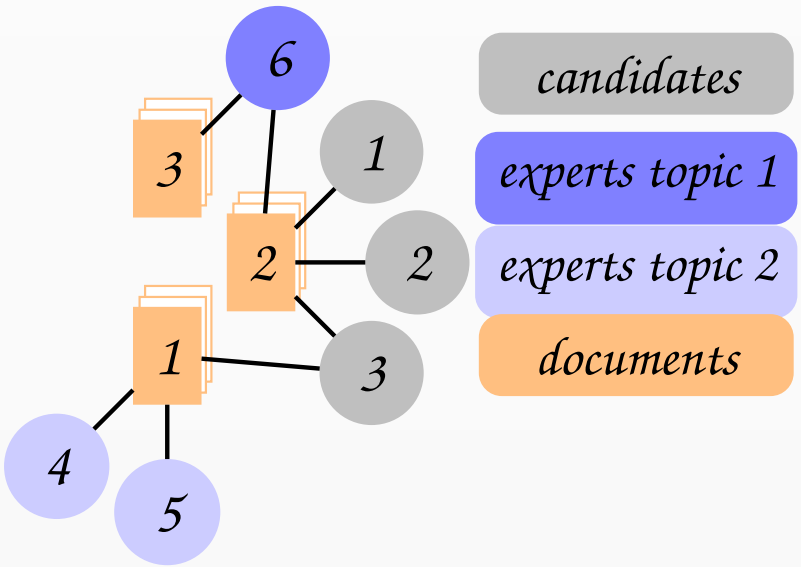
- precision at k
- average precision
- reciprocal rank
- area under the roc curve

Limitations ?

- Only 7 queries : might be statistically not significant.
- Queries are keywords : bias due to their short length ?
- Real case applications : users have different behaviors.

Another Evaluation (document-query)

Sampling documents to use them as queries :



for topic 1 :
    for expert 6 :
        we sample document 3.
        we sample document 2.
for topic 2 :
    for expert 4 :
        we add document 1.
    for expert 5 :
        we add document 1.

Sampling Documents-Queries from Ground Thruth Experts

Table – Document-Queries Sampled

Topic	Exps	Docs
intelligent agents	29	685
planning	33	510
machine learning	38	628
natural language processing	38	453
information extraction	18	290
semantic web	41	521
support vector machine	29	323
TOTAL	**210**	**3410**

Baseline Algorithms

- **P@noptic :** Craswell et al., "P@ noptic expert : Searching for experts not just for documents"
  => A simple algorithm creating meta-documents for each candidate by aggregating their documents **(only text)**.
- **Vote :** Macdonald et Ounis, "Voting for candidates : adapting data fusion techniques for an expert search task"
  => A voting system where each document votes for its candidates weighted by their similarity to the query **(text and first-order proximity)**.
- **Propagation :** Serdyukov, Rode et Hiemstra, "Modeling multi-step relevance propagation for expert finding"
  => PageRank-like algorithm intialized with query-documents similarities **(text and N-order proximity)**.

Topic-query Evaluation Results

		TF	TF-IDF	LSI
P@noptic	AUC	0.777±0.099	0.778±0.102	0.832±0.083
	P@10	0.662±0.262	0.685±0.260	0.615±0.301
	AP	0.398±0.193	0.415±0.204	0.395±0.233
	RR	1.615±0.923	1.538±0.746	1.769±0.890
Vote	AUC	0.714±0.137	0.714±0.138	0.800±0.107
	P@10	0.608±0.312	0.608±0.287	0.538±0.325
	AP	0.373±0.212	0.381±0.209	0.390±0.247
	RR	2.308±2.398	1.538±0.929	2.769±3.765
Prop ($\eta = 0.1$)	AUC	0.834±0.093	0.834±0.096	0.824±0.085
	P@10	0.669±0.270	0.677±0.264	0.615±0.298
	AP	0.458±0.239	0.473±0.243	0.389±0.230
	RR	1.462±0.929	1.538±1.082	1.692±1.136
Propagation ($\eta = 0.5$)	AUC	0.842±0.086	0.842±0.088	0.833±0.083
	P@10	0.677±0.255	0.685±0.274	0.615±0.296
	AP	0.457±0.232	0.472±0.242	0.395±0.231
	RR	1.308±0.606	1.385±0.738	1.692±1.136

Table – Results of the Topic-query evaluations.

Document-query Evaluation Results

(a) Scores and STD.

		TF	TF-IDF	LSI
P@noptic	AUC	0.593±0.131	0.618±0.134	0.620±0.139
	P@10	0.255±0.266	0.335±0.294	0.302±0.298
	AP	0.150±0.117	0.181±0.133	0.174±0.133
	RR	9.153±16.028	6.210±13.323	8.663±16.262
Vote	AUC	0.606±0.131	0.630±0.137	0.637±0.142
	P@10	0.284±0.263	0.322±0.280	0.275±0.276
	AP	0.169±0.131	0.193±0.145	0.187±0.152
	RR	6.819±15.421	5.783±13.983	8.438±16.987
Propagation ($\eta = 0.1$)	AUC	0.591±0.140	0.617±0.143	0.612±0.145
	P@10	0.256±0.253	0.336±0.292	0.300±0.288
	AP	0.152±0.114	0.183±0.132	0.173±0.127
	RR	9.035±16.718	6.773±14.641	8.948±17.422
Propagation ($\eta = 0.5$)	AUC	0.598±0.141	0.623±0.142	0.621±0.147
	P@10	0.255±0.245	0.333±0.283	0.298±0.282
	AP	0.155±0.115	0.185±0.133	0.177±0.130
	RR	8.994±17.224	6.419±14.088	9.089±18.358

(b) Inter-topic STD.

TF	TF-IDF	LSI
0.114	0.112	0.119
0.174	0.191	0.195
0.097	0.102	0.104
6.063	3.676	6.088
0.109	0.109	0.111
0.166	0.181	0.174
0.102	0.110	0.115
3.400	2.789	4.168
0.128	0.126	0.131
0.156	0.187	0.184
0.093	0.102	0.101
9.175	6.139	9.245
0.125	0.122	0.128
0.147	0.177	0.175
0.093	0.101	0.102
8.488	5.059	8.944

Table – Results of the Document-query evaluations.

Lessons Learned

- Not the same algorithms that perform best for both evaluations.
- Document-query inter-topic standard deviations are lower than the overall query standard deviations.
- The (short) namings of the topics and their low number have a too strong influence on the evaluation.
- Worse results for the document-query evaluation highlights the need of annotations for documents.

Ongoing Works

- Human annotations of AMiner document-queries.
- Expert finding evaluation python librairy based on AMiner and StackExchange[5] data.
- Online evaluations based on volunteers of `peer.us`.
- Algorithm to embed experts and documents in the same vector space with the ability to embed previously unseen documents.

5. `https://stackexchange.com/`

Thank You

Questions ?

GitHub : github.com/brochier/impact_query_expert_finding

References I

Nick Craswell et al. “P@ noptic expert : Searching for experts not just for documents”. In : *Ausweb Poster Proceedings, Queensland, Australia*. T. 15. 2001, p. 17.

Craig Macdonald et Iadh Ounis. “Voting for candidates : adapting data fusion techniques for an expert search task”. In : *Proceedings of the 15th ACM international conference on Information and knowledge management*. ACM. 2006, p. 387–396.

Pavel Serdyukov, Henning Rode et Djoerd Hiemstra. “Modeling multi-step relevance propagation for expert finding”. In : *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM. 2008, p. 1133–1142.