On the Biased Assessment of Expert Finding Systems

Jens-Joris Decorte[1,2,*], Jeroen Van Hautte[2], Chris Develder[1] and Thomas Demeester[1]

[1]*Ghent University – imec, 9052 Gent, Belgium*

[2]*TechWolf, 9000 Gent, Belgium*

Abstract

In large organisations, identifying experts on a given topic is crucial in leveraging the internal knowledge spread across teams and departments. So-called enterprise expert retrieval systems automatically discover and structure employees' expertise based on the vast amount of heterogeneous data available about them and the work they perform. Evaluating these systems requires comprehensive ground truth expert annotations, which are hard to obtain. Therefore, the annotation process typically relies on automated recommendations of knowledge areas to validate. This case study provides an analysis of how these recommendations can impact the evaluation of expert finding systems. We demonstrate on a popular benchmark that system-validated annotations lead to overestimated performance of traditional term-based retrieval models and even invalidate comparisons with more recent neural methods. We also augment knowledge areas with synonyms to uncover a strong bias towards literal mentions of their constituent words. Finally, we propose constraints to the annotation process to prevent these biased evaluations, and show that this still allows annotation suggestions of high utility. These findings should inform benchmark creation or selection for expert finding, to guarantee meaningful comparison of methods.

1. Introduction

As organisations grow in size, effectively leveraging internal expertise becomes harder, as it is harder to locate experts on specific matters. This need for efficiently locating expertise in large organisations has been recognized for over two decades, with early systems such as *P@NOP-TIC Expert* that automatically identify experts based on information in an organisation's intranet [1]. This task, known as expert finding, is a specialized form of information retrieval (IR) where the focus is on identifying individuals with relevant expertise rather than on retrieving documents. Another related task, expert profiling, focuses on retrieving all areas of expertise for a given individual [2]. The evaluation of these tasks, grouped under *expertise retrieval* [3], requires developing a comprehensive gold standard of expertise annotations. A setup where a list of experts is annotated for a given topic proves difficult, resulting in either just one or two experts linked per topic [4, 5], or in the case of more extensive expert lists, limited to a mere seven topics overall [6]. As a result, annotation efforts have shifted towards asking the experts themselves to list their areas of expertise [7]. As

expert profiling and expert finding are strongly related, there is a trend of using expert profiling benchmarks for the task of expert finding as well [7, 8, 9]. If a complete and accurate gold standard of expertise profiles is available, it can be inverted to identify relevant experts for a specific topic. However, achieving such comprehensive and accurate profiles is often unrealistic. The sheer number of topics typically precludes exhaustive consideration during annotation. Secondly, self-selected topics rely on the expert's recollection and understanding of the system's taxonomy, often resulting in sparse profiles subject to cognitive biases, such as recency bias. To address this, the annotation process often includes an automated system recommending additional, likely topics for each expert [8, 9]. These system-validated topics yield more comprehensive and varied expert profile annotations.

Note that these personalized recommendations can greatly influence which topics each expert might consider during annotation. We argue that this can, under certain circumstances, preclude meaningful comparisons of annotations across experts, and therefore the use of these benchmarks for expert finding systems. Specifically, this work addresses these research questions:

- **RQ1:** Do system-validated and self-selected annotations exhibit significantly different characteristics that could bias the evaluation of expert finding systems?

- **RQ2:** How do system-validated annotations impact perceived performance of term-based versus neural retrieval systems in expert finding tasks?

- **RQ3:** Can we establish constraints for a new annotation setup that ensures the evaluation of expert finding systems remains representative and unbiased?

We address these questions based on an analysis of the popular TU Expert Collection [8], which makes available multiple sets of ground truth that nicely facilitate our analysis. The nature of self-selected and system-validated expertise profiles is compared, specifically examining the properties of system-validated annotations and their impact on the validity of the expert finding task in section 3. We implement both traditional term-based retrieval systems and more recent neural IR methods in section 4. Additionally, the section covers a procedure to augment all test queries in the TU Expert Collection with synonyms, allowing to further analyze the effect of any term-based biases in the annotations. We also propose constraints for system-validated annotations and demonstrate the potential of a new annotation suggestion system in this section. Finally, section 5 discusses all results and provides answers to the research questions.

2. Related Work

Expert finding systems The development of expert finding systems has a rich history, starting with the introduction of *P@NOPTIC Expert*, an early systems designed to automatically identify experts based on textual documents available within an organisation's intranet [1]. To this day, expert finding remains an important and challenging topic for large organisations across different niches like the medical domain [10]. Most expert finding methods are formalized as one of two prominent models: the candidate model and the document model, referred to as the query-independent and query-dependent models, or Model 1 and Model 2, respectively [11]. The candidate model regards each candidate expert as the set of their linked documents, to directly retrieve relevant experts given a query. The document model operates on two steps, first determining the relevance of individual documents towards a query, and afterwards aggregating this document ranking into a candidate ranking. Of these, the document model (Model 2) has generally been shown to be more effective [11]. Subsequent research has largely built upon these models, with some studies exploring expert finding as a voting problem, utilizing data fusion techniques in a metasearch framework [12]. Other works have an extended input data scope to the expert finding system, such as by incorporating prior topic distributions [13] or by leveraging document structure to enhance retrieval performance [14]. Our work relies on the presence of a textual corpus linked to employees without further properties like document structure.

Expert finding benchmarks The first large-scale benchmarks for expert finding were developed as part of the TREC Enterprise track, which ran from 2005 to 2008. It introduced the W3C expert corpus in 2005, alongside an *expert search* task consisting of 50 knowledge areas, annotated with experts from a total of 1,092 candidates [4]. In 2007, the CERC dataset was introduced [5]. Similar to the W3C dataset, it included 50 topics, which were developed by nine science communicators at CSIRO. These communicators were then tasked with identifying one or two CSIRO staff members as experts on each topic, contributing to a robust dataset for expert finding research. Another notable dataset is derived from DBLP biographical data, augmented with abstracts from Google Scholar [6]. This dataset contains 953,774 papers in total and 574,369 valid authors, with 2,498 topics sourced from a research events website. However, only seven topics have been annotated with expert lists, each containing between 20 and 45 experts [6]. The dataset development process, both for TREC and the DBLP-based dataset, highlights the difficulty of identifying experts on a certain topic within large organisations, as annotators often lack detailed knowledge of the topics themselves.

Expert profiling benchmarks A more distributed approach to gathering annotations involves asking employees to fill in their own expertise profiles, as seen in the UvT dataset [7]. This new annotation scheme became more prominent shortly after the introduction of the task of expert *profiling* (rather than *finding*), where the goal is to retrieve all areas of expertise for a given individual [2]. The UvT dataset was the first large-scale benchmark for expertise profiling relying on self-reported expertise. However, this approach often results in sparse profiles due to the difficulty of recalling all areas of expertise. To address this sparsity, semi-automated annotation procedures have been proposed. Berendsen et al.[8] extended the self-selected profiles from the UvT dataset by presenting up to 100 high-probability topics for further annotation, re-releasing the new annotation sets under the name of the TU Expert Collection. Similarly, Mangaravite et al. [9] employed a content-based tag recommendation system to suggest annotations for approval. These enriched annotations, while originally intended for expert profiling, have been increasingly used to evaluate expert finding tasks as well [15, 16].

The use of personalized annotation suggestions raises concerns about the validity of these annotations for evaluating expert finding systems, which this study aims to address. Specifically, because these suggestions are different for each employee, the impact of this mechanism may introduce properties in the annotations that forego their comparability across employees. Our work is closely related to that of Berendsen et al. [8], who conducted an extensive study on the impact of different expert profile annotation schemes on the evaluation of expert profiling tasks. However, our focus diverges in that we specifically

investigate the impact of these benchmarks on expert finding, uncovering significant challenges in their usability under system-validated setups.

3. Analysis of Annotation Schemes

We analyze the differences of self-selected versus system-validated expertise annotations, and how they may influence the perceived performance of expert finding systems when used as a benchmark. Section 3.1 introduces the TU Expert Collection, which is the dataset used in this analysis, as introduced in [8]. We perform initial analysis of the annotation suggestions in section 3.2, showing their utility for expanding profile annotations, but also their high false negative rate. Finally, section 3.3 analyzes the mechanism behind these false negatives, exposing a large positive bias towards literal mentions of the knowledge topics' constituent words in the corpus.

3.1. TU Expert Collection

The TU expert collection is an expertise retrieval benchmark focused on a knowledge-intensive organisation, namely the Tilburg University [8]. It is an updated version of the earlier UvT dataset [7]. The dataset contains a variety of documents, being academic publications, supervised student dissertations, course descriptions, and research summaries. These documents are primarily in Dutch and English, and are explicitly linked to experts in the university's Webwijs system, indexes over 2,000 unique knowledge areas and 761 employees. The TU dataset provides several ground truth (GT) sets of graded expert profile annotations, labeled GT1 through GT5. These annotations are the result of experts indicating their expertise areas on a scale of 1 (lowest) to 5 (highest). Note that in this work, we consider all annotations as binary and ignore the attached grades, due to inconsistencies in how different annotators may interpret and apply these grades, compromising the comparability of grades across experts [8]. We leave the analysis of graded relevance annotations for future work. GT1 contains the self-selected knowledge areas of 761 TU employees. To further expand these annotations, a system was developed to recommend up to 100 highly likely knowledge areas to each expert, which they could easily validate or discard through a user interface. The extended expertise profiles of all 239 participating employees is bundled as GT5, accompanied by GT4 which simply binarizes the annotations by dropping the graded relevance scores. For ease of use, GT2 is provided as the subset of GT1 for those 239 employees. Finally, GT3 is an alteration of self-selected profiles in GT2, reduced to only those topics that were also present in the personalized suggested annotations.

3.2. Distribution of system-validated annotations

As reported in [8], the average self-selected profile in GT1 contains 6.4 knowledge areas. The average size of the extended profiles in GT5 expands to 8.6 areas. Notably, we find that the percentage of employees with three or fewer expertise areas decreases from 19.7% in GT1 to 10.5% in GT5. Additionally, system-generated profiles capture 81% of the final knowledge areas compared to 65% for self-selected profiles, and unique topics in the annotations grows from 937 in GT1 to 1,266 in GT5. This shows the sparsity of self-selected expertise profiles, and how the situation improves through personalized system recommendations for annotation. However, because these recommendations are personal to each expert, the underlying recommendation method may compromise the comparability of the annotated knowledge areas across experts. To explore this, we focused on niche topics, present in more than one but no more than three self-selected profiles in GT2, identifying 290 such topics. Examples are sub-saharan africa, policy evaluation, cognitive linguistics, nonprofit organisations and extreme value theory. By contrasting GT2 with GT3, we know whether a self-selected topic was also part of the system annotation suggestions, allowing us to estimate the recall of this system. We find that only 125 out of the 290 niche topics – around 43% – were recommended for annotation to all experts who had self-selected it. It is this low recall that can compromise the comparability of annotations across experts: if it is caused by a certain weakness of the annotation recommendation system, expert finding systems with a similar weakness will produce the same recall patterns and therefore may appear stronger than they are.

3.3. Term-bias in system-validated annotations

To construct GT5, up to 100 knowledge areas were suggested for further annotation to each expert, produced by an ensemble of eight expert profiling systems [8]. These systems vary in retrieval models (Model 1 or Model 2), the query language (English or Dutch), and whether they consider relationships between topics in the Webwijs system. All these systems have in common that they model the probability of a topic for a document or expert based on the literal textual occurrences of their constituent words. This approach is prone to false negatives due to its inability to account for synonyms and other semantic nuances, leading to a low recall and a strong bias towards literal mentions of the topic's constituent words.

We aim to quantify the presence of this bias towards literal mentions in system-validated topics or their constituent words. To this end, we construct a corpus with

one long document per expert, being the concatenation of all original documents linked to the expert. We then calculate tf-idf scores of queries with respect to these concatenated expert documents to express the degree to which their constituent words are being literally mentioned. Whenever both an English and Dutch name are available for a topic, we consider the largest of the tf-idf scores of both versions. By contrasting GT2 and GT3, we are able to determine for each topic in the self-selected profiles whether it was also included in the suggested annotation list for the corresponding expert. Figure 1 shows the distribution of tf-idf scores for both groups of self-selected topics throuh a boxplot, with commonly used whiskers drawn to the farthest datapoint within 1.5 times the interquartile range from the nearest hinge. It shows a large shift in scores for the subset of self-selected topics that are also included in the annotation recommendations versus those that are not included.
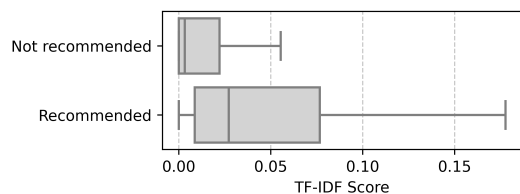

Figure 1: Distribution of tf-idf scores for self-selected topics that were part of the annotation suggestions versus those that were not. Significantly higher scores are observed for those that were part of the annotation suggestions.

Given this observation, we expect this bias to be extended into the additional topics that are added to the profiles through system-validation. Figure 2 shows the distribution of tf-idf scores of the topics in the original self-selected expertise profiles, compared to the topics added after system-validation. As expected, a clear positive bias in tf-idf scores is observed for the system-validated topics. These findings validate our concerns about system-driven annotation biases.
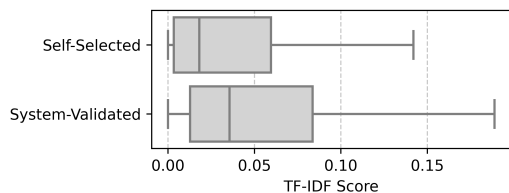

Figure 2: Distribution of tf-idf scores for self-selected topics versus additional topics added through system validation.

In conclusion, our analysis reveals a strong term frequency bias in the annotation recommendations. As a result, there are likely false negatives in the system-validated annotations that systematically favor systems with similar term frequency-based biases. This suggests that evaluations of expert finding systems using these annotations may overestimate the performance of systems that share these biases.

4. Assessment of Expert Finding

4.1. Expert Finding Models

We implement two different expert finding systems and evaluate them against different ground truths. Both systems operate under the query-dependent Model 2, first determining the relevance of each document in the corpus to the query and then aggregating document ranks into a final ranking of experts. Two different information retrieval models are implemented to rank the documents against a query. We implement a popular term-based retrieval model, as well as a more modern neural IR technique, aiming to maximally expose potential term-based biases in the annotations.

Term-based retrieval: we use BM25 [17] to rank all documents against the query topic because it is a commonly used term-based retrieval method that accounts for document lengths. In order to support the multilinguality of the dataset, we simply concatenate the English and Dutch translations of the topic when both are present.

Late-interaction neural retrieval: We opt for the ColBERT retrieval technique [18], due to its unique combination of efficiency and strong IR capabilities. To facilitate the multilingual nature of the corpus, we use the multilingual ColBERT-XM model [19]. The corpus is split into chunks of up to 256 tokens, with an overlap of 64 tokens. We make use of the *RAGatouille*[1] implementation to chunk and index the whole corpus, which took close to eight hours on one NVIDIA P100 GPU. As with the BM25 model, we simply concatenate the English and Dutch translations of the topic when present. The documents are ranked according to the average retrieval score of all their constituent chunks.

Following [16], for both IR methods, we use the same *rr* function to aggregate the document ranking into an expert ranking, as defined by:

$$rr(c) = \sum_{j=1}^{|D_{c,q}|} \frac{1}{\text{rank}(d_j)}$$

---

[1]https://github.com/AnswerDotAI/RAGatouille

In this equation, $D_{c,q}$ is the subset of documents linked to candidate c that are retrieved for query q. The rank of the retrieved documents is indicated by rank(d_j), starting at 1 for the highest ranked document.

4.2. Query Augmentation

With the goal of further studying the impact of term-based biases in the annotations, we extend the Webwijs knowledge areas with synonyms. Qualitative expert finding system should surface relevant experts on a topic, even if provided by a synonym of the query instead of the original query. Because of this, query synonyms provide an opportunity to study term-based annotation biases. We manually annotate 109 randomly selected knowledge areas with both English and Dutch synonyms. To facilitate the annotation process, the topic up for annotation is contextualized by providing the annotator with all its relations to other topics in the Webwijs inventory. For further contextualization, we gather the corresponding wikidata and wikipedia page on the topic if they exist. Whenever available, good synonyms are selected from the *"Also known as"* table in wikidata. If not, we scan the first paragraph of the wikipedia page for synonyms. Finally, we manually provide a synonym if none are available in the wiki pages. An example result is shown here:

```
Topic EN: auction theory
Topic NL: veilingstheorie

Webwijs links:
Makes use of: auctions / veilingen

Annotation:
wikidata: https://www.wikidata.org/wiki/Q771334
wikipedia: https://en.wikipedia.org/wiki/Auction_theory

Synonym EN: Bidding Theory
Synonym NL: Biedingstheorie
```

Based on the initial set of synonym annotations, we automatically generate synonyms for all remaining queries using OpenAI's GPT4o model, selected for its high accuracy on common benchmarks. We randomly select 35 annotated queries for training and the remaining 74 for validation, and automatically optimize a chain-of-thought prompt [20] using DSPy's BootstrapFewShot prompt optimization technique [21]. We release the full set of synonym annotations.[2] Limited manual quality checks were performed on these synonyms. We note that this process can be more qualitatively performed in future studies, however the main reasoning of using topic synonyms to indicate a bias towards literal mentions of their constituent words still holds, irrespective of suboptimal quality of the synonyms.

4.3. Alternative Annotation Suggestions

We analyze an alternative approach to suggest additional knowledge areas for annotation. Specifically, we introduce the constraint that the suggestion mechanism has no access to the document corpus, and may only recommend additional knowledge areas based on the available self-selected topics. This setup prevents any systematic bias stemming from the annotation procedure with respect to the textual corpus. The downside of this setup is that the rich information in the corpus cannot be utilised, and it instead relies on at least a small number of manually annotated expertise areas per expert. We argue that experts should be sufficiently engaged in the annotation process such that this is a reasonable requirement. Additionally, there is an opportunity for these annotation suggestions to dynamically adapt throughout the validation procedure, although we leave this out of the current scope.

We develop two item-to-item recommendation systems (where the Webwijs topics serve as items). The first system learns from the full set of self-selected profiles in GT1, and recommends topics that have high pointwise mutual information (PMI) with the given topic. PMI is a simple but effective measure of association between items. Topics that occur three times or less are excluded from this system to ensure a minimum level of robustness. The second system operates on all Webwijs topics, and recommends topics that are semantically similar to a given topic, as measured by the cosine similarity of semantic neural embeddings of their respective names. We use a generic multilingual sentence-transformer model (*paraphrase-multilingual-mpnet-base-v2*)[3] that was pre-trained on over 1B English sentence pairs [22] and then adapted to over 50 languages through a knowledge distillation process [23]. We embed English and Dutch names separately and always consider the highest similarity between either versions. For a given self-selected profile, we compile a list of up to 100 annotation recommendations by pooling the top recommendation of both systems, looping over the self-selected expertise topics in a round-robin fashion.

5. Results and Discussion

Impact of the annotation on expert finding evaluations   We evaluate system performance using precision at 5 (P@5), mean average precision (MAP), normalized discounted cumulative gain (nDCG) and mean reciprocal rank (MRR). These metrics provide a comprehensive view of the retrieval system's ability to rank relevant experts at the top of the list. All evaluations are conducted using the

---

	GT2				GT5				GT5 with synonym queries			
	P@5	MAP	nDCG	MRR	P@5	MAP	nDCG	MRR	P@5	MAP	nDCG	MRR
BM25	15.47	37.81	52.39	46.06	**21.94**	**56.56**	**67.53**	**64.02**	12.16	31.68	46.41	38.12
ColBERT	**16.29**	**39.78**	**54.07**	**48.43**	18.44	46.46	59.39	54.41	**13.98**	**34.21**	**48.68**	**40.42**

Table 1

Performance metrics for both the term-based (BM25) and the late-interaction neural retrieval based (ColBERT) methods under different annotation schemes. GT2, containing just self-selected expertise topics, shows stronger performance for the ColBERT method, while the addition of system-validated topics in GT5 leads to a strong edge for the BM25 system. Using the same GT5 annotations, but swapping test queries for their synonyms, yields a system ranking that is again in accordance with GT2.

official TREC evaluation software[4], ensuring standardized and comparable results. We report the performance of the BM25 and the ColBERT system on both the self-selected profiles (GT2) as well as their larger counterparts extended with system-validated topics (GT5) in table 1. We find that the ColBERT-based expert finding system outperforms its BM25-based counterpart on all metrics when using the self-selected profiles (GT2) as ground truth. The performance of both systems increases considerably when assessing them against the system-validated profile (GT5). However, the increase in performance is much more drastic for the BM25-based system, leading it to strongly exceed ColBERT. We hypothesize that it is the term-frequency bias in the annotation procedure of GT5 that leads to a strongly overestimated performance measurement for the term-based BM25 approach. When swapping the test queries in GT5 for their synonyms, a significant drop in performance is observed, especially for the BM25 system which drops over 20 %-points in MAP, nDCG and MRR. We also observe that system ranking in this scenario corresponds to that under the GT2 evaluation.

Statistics of the alternative annotation suggestions
We perform the same analysis as in section 3 with respect to the recall of 290 specific topics in the annotation recommendations. This requires the specific topic for which recall is measured to first be removed from the self-selected profile. The results show that 235 out of the 290 specific topics (around 81%) is being recommended to every expert that had self-selected the topic. Compared to the 43% of the content-based annotation recommendations in the original study, this is a considerable increase in recall, which should further improve the comparability of annotations across experts. For completeness, we also report how well these new recommendations cover the topics that where added in GT5 compared to GT2. We find that a total of 1,059 topic additions are made, and 505 (around 48%) of these topics are also present in our proposed annotation recommendation method. Because this proposed method has no access to the documents in

the organisation, we do not expect strong overlap, and we consider 48% to be relatively high. Apart from this overlap measurement, it is difficult to assess the true precision of these recommendations because we did not have access to the candidate experts to facilitate manual validation. However, examples of the topic recommendations is provided in appendix A.

In conclusion, our analysis of the TU Expert Collection allows us to answer *yes* to all three research questions. Section 3 provides an answer to **RQ1**, showing that the underlying mechanism used for the system-validated topics is subject to a high false negative rate, and that it introduces a significant bias towards literal mentions of knowledge areas' constituent words. With respect to **RQ2**, as shown above, the perceived performance of expert finding systems is indeed strongly impacted by these system-validated topics, and they even lead to significant differences in the ranking of these systems. Finally, we have proposed an annotation suggestion procedure that is independent of the document corpus, and we have developed such a system accordingly. It exhibits strong utility for the annotation process while significantly reducing the false negative rates observed in the original benchmark, leading us to answer **RQ3** positively as well. Our analysis should help future work on expert finding – or evaluation thereof – make more informed decisions with respect to the selection or creation of these benchmarks.

Acknowledgments

---

[4]https://trec.nist.gov/trec_eval/

References

[1] N. Craswell, D. Hawking, A.-M. Vercoustre, P. Wilkins, P@ noptic expert: Searching for experts not just for documents, in: Ausweb Poster Proceedings, Queensland, Australia, volume 15, Citeseer, 2001, p. 17.

[2] K. Balog, M. De Rijke, Determining expert profiles (with an application to expert finding), in: Proceedings of the 20th International Joint Conference on Artifical Intelligence, IJCAI'07, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007, p. 2657–2662.

[3] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, L. Si, Expertise retrieval, Found. Trends Inf. Retr. 6 (2012) 127–256. URL: https://doi.org/10.1561/1500000024. doi:10.1561/1500000024.

[4] N. Craswell, A. P. De Vries, I. Soboroff, Overview of the trec 2005 enterprise track., in: Trec, volume 5, 2005, pp. 1–7.

[5] I. Soboroff, P. Bailey, N. Craswell, A. de, Overview of the trec 2007 enterprise track, 2008. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=152187.

[6] H. Deng, I. King, M. R. Lyu, Formal models for expert finding on dblp bibliography data, in: 2008 Eighth IEEE International Conference on Data Mining, 2008, pp. 163–172. doi:10.1109/ICDM.2008.29.

[7] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, A. van den Bosch, Broad expertise retrieval in sparse data environments, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, Association for Computing Machinery, New York, NY, USA, 2007, p. 551–558. URL: https://doi.org/10.1145/1277741.1277836. doi:10.1145/1277741.1277836.

[8] R. Berendsen, M. De Rijke, K. Balog, T. Bogers, A. Van den Bosch, On the assessment of expertise profiles, Journal of the American Society for Information Science and Technology 64 (2013) 2024–2044.

[9] V. Mangaravite, R. L. Santos, I. S. Ribeiro, M. A. Gonçalves, A. H. Laender, The lexr collection for expertise retrieval in academia, in: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 721–724. URL: https://doi.org/10.1145/2911451.2914678. doi:10.1145/2911451.2914678.

[10] S. Borna, B. Barry, S. Makarova, Y. Parte, C. Haider, A. Sehgal, B. Leibovich, A. Forte, Artificial intelligence algorithms for expert identification in medical domains: A scoping review, European Journal of Investigation in Health, Psychology and Education 14 (2024) 1182–1196. doi:10.3390/ejihpe14050078, publisher Copyright: © 2024 by the authors.

[11] K. Balog, L. Azzopardi, M. de Rijke, Formal models for expert finding in enterprise corpora, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, Association for Computing Machinery, New York, NY, USA, 2006, p. 43–50. URL: https://doi.org/10.1145/1148170.1148181. doi:10.1145/1148170.1148181.

[12] C. Macdonald, I. 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, CIKM '06, Association for Computing Machinery, New York, NY, USA, 2006, p. 387–396. URL: https://doi.org/10.1145/1183614.1183671. doi:10.1145/1183614.1183671.

[13] H. Fang, C. Zhai, Probabilistic models for expert finding, in: G. Amati, C. Carpineto, G. Romano (Eds.), Advances in Information Retrieval, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 418–430.

[14] J. Zhu, D. Song, S. Rüger, M. Eisenstadt, E. Motta, The open university at trec 2006 enterprise track expertsearch task, in: Fifteenth Text REtrieval Conference (TREC2006), 2006. URL: http://trec.nist.gov/pubs/trec15/papers/openu.ent.final.pdf, proc. of The Fifteenth Text REtrieval Conference (TREC 2006), Gaithersburg, Maryland USA, National Institute of Standards and Technology, USA.

[15] C. Van Gysel, M. de Rijke, M. Worring, Unsupervised, efficient and semantic expertise retrieval, in: Proceedings of the 25th International Conference on World Wide Web, WWW '16, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2016, p. 1069–1079. URL: https://doi.org/10.1145/2872427.2882974. doi:10.1145/2872427.2882974.

[16] P. Cifariello, P. Ferragina, M. Ponza, Wiser: A semantic approach for expert finding in academia based on entity linking, Information Systems 82 (2019) 1–16. URL: https://www.sciencedirect.com/science/article/pii/S0306437918302515. doi:https://doi.org/10.1016/j.is.2018.12.003.

[17] S. E. Robertson, S. Walker, Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, in: B. W. Croft, C. J. van Rijsbergen (Eds.), SIGIR '94, Springer London, London, 1994, pp. 232–241.

[18] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interac-

tion over bert, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 39–48. URL: https://doi.org/10.1145/3397271.3401075. doi:10.1145/3397271.3401075.

[19] A. Louis, V. Saxena, G. van Dijck, G. Spanakis, Colbert-xm: A modular multi-vector representation model for zero-shot multilingual information retrieval, CoRR abs/2402.15059 (2024). URL: https://arxiv.org/abs/2402.15059. doi:10.48550/arXiv.2402.15059. arXiv:2402.15059.

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[21] O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam, H. Miller, M. Zaharia, C. Potts, Dspy: Compiling declarative language model calls into self-improving pipelines, arXiv preprint arXiv:2310.03714 (2023).

[22] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.

[23] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4512–4525. URL: https://aclanthology.org/2020.emnlp-main.365. doi:10.18653/v1/2020.emnlp-main.365.

A. Example annotation recommendations

Given the knowledge area COMPUTER LINGUISTICS, the top five item-to-item recommendations according to the PMI-based system are:

1. TALKING COMPUTER
2. AUTOMATIC LANGUAGE ANALYSIS
3. MAN-MACHINE INTERACTION
4. ALGEMENE TAALWETENSCHAP (NL)
5. SPEECH TECHNOLOGY

The top five recommendations according to the semantic embedding similarity method are:

1. LANGUAGE AND COMPUTERS
2. TAALPRODUCTIE DOOR COMPUTERS (NL)
3. TAALTECHNOLOGIE EN COMPUTERS (NL)
4. LANGUAGE TECHNOLOGY AND COMPUTERS
5. COMPUTER AND GRAMMAR

Note that we use the Dutch topic name if no English name is available for the topic in Webwijs. Now consider a self-selected expertise profile consisting of the following three topics:

- CITIZENSHIP
- EUROPEAN LAW
- VREEMDELINGENRECHT (NL)

The top 20 recommended annotations for this expert, as a result of the round-robin aggregation of both types of item-to-item recommenders, are:

1. VOLUNTARY WORK
2. NATIONALITY
3. CONSTITUTIONAL ISSUES
4. THEORY OF EUROPEAN LAW
5. MIGRATION LAW
6. COMPETITION LAW
7. FOUNDATIONS OF DEVELOPMENTS IN EUROPEAN LAW
8. REFUGEE LAW
9. VRIJWILLIGERSORGANISATIES (NL)
10. INBURGERING VAN ALLOCHTONEN (NL)
11. LEGAL PRINCIPLES
12. EUROPEAN ADMINISTRATIVE LAW
13. ADMINISTRATIVE LAW
14. INTERNATIONAL LAW
15. URBAN GOVERNANCE
16. VREEMDELINGENBELEID (NL)
17. RECHTSBESCHERMING (NL)
18. EUROPEAN SOCIAL LAW
19. MUNICIPAL LAW
20. MAATSCHAPPELIJKE ORGANISATIES (NL)