On the Assessment of Expertise Profiles (Abstract)

Richard Berendsen
University of Amsterdam, The Netherlands
r.w.berendsen@uva.nl

Krisztian Balog
University of Stavanger, Norway
krisztian.balog@uis.no

Toine Bogers
Royal School of Library Information Science, Denmark
tb@iva.dk

Antal van den Bosch
Radboud University Nijmegen, The Netherlands
a.vandenbosch@let.ru.nl

Maarten de Rijke
University of Amsterdam, The Netherlands
derijke@uva.nl

1. INTRODUCTION

We summarize findings from [3]. At the TREC Enterprise Track [2], the need to study and understand *expertise retrieval* has been recognized through the introduction of the expert finding task. The goal of *expert finding* is to identify a list of people who are knowledgeable about a given topic. An alternative task, building on the same underlying principle of computing people-topic associations, is *expert profiling*, where systems have to return a list of topics that a person is knowledgeable about [1].

We focus on benchmarking systems performing the topical expert profiling task. We define this task as a ranking task, where knowledge areas from a thesaurus have to be ranked for an expert. We release an updated version of the UvT (Universiteit van Tilburg) expert collection [1]: the *TU* (Tilburg University) *expert collection*.[1] The TU expert collection is based on the *Webwijs* ("Webwise") system[2]: a publicly accessible database of TU employees who are involved in research or teaching. In a back-end for this database, experts can indicate their skills by selecting knowledge areas from an alphabetical list. Prior work has used these *self-selected knowledge areas* as ground truth for both expert finding and expert profiling tasks [1].

One problem with self-selected knowledge areas is that they may be sparse, since experts have to select them from an alphabetically ordered list of well over 2,000 knowledge areas. Using these self-selected knowledge areas as ground truth for assessing automatic profiling systems may therefore not reflect the true predictive power of these systems. To find out more about how well these systems perform in real-world circumstances, we have asked TU employees to judge and comment on profiles that have been automatically generated for them. We refer to this process as the *assessment experiment*. In § 2 we answer the broad research question "How well are we doing at the expert profiling task?" We do this through an error analysis and through a content analysis of free text comments that experts could give. During the assessment experiment, experts judge areas in the system-generated profiles on a five point scale. This yields a new set of graded relevance assessments, which we call the *judged system-generated knowledge areas*. In § 3 our research question is: "Does benchmarking a set of expertise retrieval systems with the judged system-generated profiles lead to different conclusions, compared to benchmarking with the self-selected profiles?" We benchmark eight state-of-the-art expertise retrieval systems with both sets of ground truth and investigate differences in completeness, system ranking, and the number of significant differences detected between systems.

2. THE ASSESSMENT EXPERIMENT

Generating profiles. We use eight expert profiling models. Each of them uses either Model 1 or Model 2 [1], either uses Dutch or English representations of knowledge areas, and either uses relations between knowledge areas extracted from the thesaurus or not. Because experts have limited time and participate in the experiment on a voluntary basis, we rank areas by their estimated probability of being part of the expert's profiles. The more traditional pooling approach would require experts to exhaustively judge the pool. We linearly combine output scores of the eight systems, giving each system equal weight. We boost the top three of each system by adding a sufficiently large constant to the top three scores, to make sure they are judged. System-generated knowledge areas that were in the original self-selected profile of the expert are ticked by default in the interface, but the expert may deselect them, thereby judging them non-relevant.

The assessment interface. Using the assessment interface, each expert can judge retrieved knowledge areas relevant by ticking them. Immediately below the top twenty knowledge areas listed by default, the expert has the option to view and assess additional knowledge areas. For the ticked knowledge areas, experts have the option to indicate a level of expertise. If they do not do this, we still include these knowledge areas in the judged system-generated profiles, with a level of expertise of three ("somewhere in the middle"). At the bottom of the interface, experts can leave any comments they might have on the generated profile.

Error analysis of system-generated profiles. Here, we aim to find properties of experts that can explain some of the variance in nDCG@100 performance. We use the self-selected profiles of all 761 experts we generated a profile for, allowing us to incorporate self-selected knowledge areas that were missing from the system-generated profiles in our analysis. Based on visual inspection, we find no correlation between the number of relevant knowledge areas selected and nDCG@100, and no correlation between the number of documents associated with an expert and nDCG@100 either. Intuitively, the relationship between the ratio of relevant knowledge areas and number of documents associated with the expert is also interesting. However this ratio does not correlate with nDCG@100 either. Looking a bit deeper into the different kinds

[1] http://ilps.science.uva.nl/tu-expert-collection
[2] http://www.tilburguniversity.edu/webwijs/

DIR 2013, April 26, 2013, Delft, The Netherlands.

of document that can be associated with an expert, we find that it matters whether or not an expert has a research description. For the 282 experts without a research description we achieve significantly lower average nDCG@100 performance than for the remaining 479 experts (Welch Two Sample t-test, $p < 0.001$). The difference is also substantial: 0.39 vs. 0.30 for experts with and without a research description, respectively. It is not surprising that these research descriptions are important; they constitute a concise summary of a person's qualifications and expertise, written by the expert himself/herself.

Content analysis of expert feedback. 239 Experts participated in the self-assessment experiment, providing graded relevance judgments. 91 Of them also left free text comments. We study what are important aspects in expert feedback by means of a content analysis. In our analysis, expert comments were coded by two of the authors, based on a coding scheme developed in a first pass over the data. A statement could be assigned multiple aspects. After all aspect types were identified, the participants' comments were coded in a second pass over the data. Upon completion, the two coders resolved differences through discussion. Micro-averaged inter-annotator agreement (the number of times a comment was coded with the same aspect divided by the total number of codings) was 0.97. The main aspects in the feedback of experts are (i) missing a key knowledge area in the generated profile (36%); (ii) only irrelevant knowledge areas in the profile (16.9%); (iii) redundancy in the generated profiles (11.2%); (iv) knowledge areas being too general (11.2%). Based on these results, it seems there is still room for improvement in the performance of expert profiling systems. Also, interesting directions for future work are to address the redundancy in generated profiles, and to take into account the specificity of knowledge areas.

3. BENCHMARKING DIFFERENCES

Completeness. To assess completeness, we estimate the set of all relevant knowledge areas for an expert with the union of the self-selected profile and the judged system-generated profile. Doing this, we find that the judged system-generated profiles are more complete. On average, a judged system-generated profile contains 81% of all relevant knowledge areas, while a self-selected profile contains only 65%.

Changes in system ranking. To better understand the differences in evaluation outcomes between using the self-selected profiles (we call this ground truth set: **GT1**) and the judged system-generated profiles (we call this set **GT5**), we construct three intermediate sets of ground truth (**GT2-4**). Each intermediate set differs from the previous set in only one aspect; in this way we can isolate the contribution each difference makes to differences in evaluation outcomes. The intermediate sets of ground thruth are: **GT2**: The 239 self-selected profiles of participants in the assessment experiment; **GT3**: For each self-selected profile of an assessor, we only use knowledge areas that were in the system-generated profile. This means that knowledge areas that are not in the system-generated profile are treated as irrelevant; **GT4**: The knowledge areas judged relevant during the assessment experiment. We only consider binary relevance; if a knowledge area was selected it is considered as relevant, otherwise it is taken to be irrelevant. We report Kendall's τ correlation between system rankings using consecutive sets of ground truth. We rank the eight systems that contributed to the generated profile, but leave out the algorithm that combined them. In this abstract, we focus on system rankings computed with nDCG@100. With eight systems, Kendall's τ correlations of 0.79 or higher are significant at the $\alpha = 0.01$ level. Correlating **GT1-GT2**, we find that evaluating on a subset of experts does not change system ranking much: $\tau = 0.86$. Correlating **GT2-GT3**, we find that regarding non-pooled knowledge areas as irrelevant does not rank our eight systems very differently: $\tau = 0.86$. Correlating **GT3-GT4** we find that new knowledge areas judged relevant during the assessment do change system ranking: $\tau = 0.56$. Contrasting **GT4-GT5** we find that considering the grade of relevance does not change system ranking: $\tau = 1.00$.

Pairwise significant differences. The final analysis we conduct concerns a high-level perspective: the sensitivity of our evaluation methodology. The measurement that serves as a rough estimate here is the average number of systems each system differs from; we compute this for each of the five sets of assessments **GT1-5**, and focus here on nDCG@100. We use Fisher's pairwise randomization test ($\alpha = 0.001$). For **GT1** we get 4.75. For **GT2** we observe 3.00, the decrease is not surprising as **GT2** has much less experts. Regarding non-pooled knowledge areas as irrelevant does not affect sensitivity much (**GT3**: 2.75). The sensitivity increases again when we evaluate with the more complete judged system-generated knowledge areas (**GT4**:3.50). Taking into account the level of expertise indicated, we see another small increase (**GT5:**4.00).

4. CONCLUSION

We released, described and analyzed the TU expert collection for assessing automatic expert profiling systems. In an error analysis of system-generated profiles, we found that it is easier to generate profiles for experts that have a research description. A content analysis of expert feedback revealed that there is room for improvement in the expert profiling task, and that an interesting direction for future work is to consider diversity in profiles. Contrasting using the self-selected profiles or using the judged system-generated profiles for evaluation, we find that the latter profiles are more complete. The two sets of ground truth rank systems somewhat differently.

Acknowledgments. This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the PROMISE Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 258191, the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, and under COMMIT project Infiniti.

References

[1] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch. Broad expertise retrieval in sparse data environments. In *SIGIR'07*, pages 551–558. ACM, 2007.

[2] K. Balog, I. Soboroff, P. Thomas, N. Craswell, A. P. de Vries, and P. Bailey. Overview of the TREC 2008 Enterprise Track. In *TREC 2008 Proceedings*. NIST, 2009. Special Publication.

[3] R. Berendsen, K. Balog, T. Bogers, A. van den Bosch, and M. de Rijke. On the assessment of expertise profiles. *JASIST*, To appear.