On the Evaluation of Entity Profiles

Maarten de Rijke[1], Krisztian Balog[1], Toine Bogers[2], and Antal van den Bosch[3]

[1] ISLA, University of Amsterdam, The Netherlands
derijke,k.balog@uva.nl
[2] IIIA, Royal School of Library & Information Science, Denmark
tb@db.dk
[3] ILK/Tilburg Centre for Creative Computing, Tilburg University, The Netherlands
antal.vdnbosch@uvt.nl

Abstract. Entity profiling is the task of identifying and ranking descriptions of a given entity. The task may be viewed as one where the descriptions being sought are terms that need to be selected from a knowledge source (such as an ontology or thesaurus). In this case, entity profiling systems can be assessed by means of precision and recall values of the descriptive terms produced. However, recent evidence suggests that more sophisticated metrics are needed that go beyond mere lexical matching of system-produced descriptors against a ground truth, allowing for graded relevance and rewarding diversity in the list of descriptors returned. In this note, we motivate and propose such a metric.

1 Introduction

Entity retrieval is concerned with the identification of information relevant to information needs that concern entities (people, organizations, locations, products, . . .) [4]. Entity finding systems return ranked lists of entities in response to a keyword query. Entity profiling systems return a ranked list of descriptions that (together) describe an entity. The profiling task can be viewed as a summarization or question answering type task for which a set of "information nuggets" needs to be extracted from a collection of documents [15]. Appropriate evaluation methodology has been defined, and later refined, by a number of authors; see, e.g., [13]. Alternatively, entity profiling systems can be viewed as systems that need to select a set of descriptors (from a knowledge source) that accurately describe a given input entity. E.g., when the type of entity of interest is people, the descriptors can be taken from an ontology describing a scientific discipline and the profiling system's task could be interpreted as expert profiling: for every individual, to identify the areas in which he or she is an expert [2]. This second, descriptor-based reading of entity profiling is the one on which we focus.

Evaluation of descriptor-based entity profiling is usually done in terms of precision and recall of the lists of descriptors produced by a system. This has several shortcomings. Returning a ranked list of descriptors for an entity is challenging. When descriptors are to be taken from a large knowledge source, near misses are likely. But not all mistakes are equally important, depending, in part, on the envisaged users. For users that are relatively new to the area described by the knowledge source, near misses that are too specific may be more problematic than ones that are too general; for expert users, this may be the other way around. Also, some descriptors may be more fitting than others, thus naturally leading to graded relevance values. Finally, the presence of closely related descriptors in a result set at the cost of omitting descriptors that highlight different aspects of an entity would certainly be viewed unfavorably by all users.

Building on work on novelty and diversity in information retrieval evaluation [1, 8, 11], we develop a scoring method for entity profiles that addresses many of the shortcomings of today's prevalent evaluation method. Our method allows for weighted non-exact matches between system-produced descriptors and ground truth descriptors; it systematically rewards more highly relevant descriptors and more diverse lists.

2 Motivation

To motivate the need for more sophisticated evaluation methods than straightforward precision/recall of descriptors, we build on a recent evaluation of an expert profiling system. We focus on the topical expert profiling task and use the UvT Expert collection [6] as our experimental platform; it is based on the Webwijs ("Web wise") system developed at Tilburg University (UvT) in the Netherlands. Webwijs is a database of UvT employees who are involved in research or teaching; each expert can self-assess his/her skills by selecting expertise areas from a hierarchy of descriptors.

Recently, a university-wide experiment was conducted at UvT in which expertise profiles were automatically generated and, subsequently, each employee was given the opportunity to assess the profile that was generated for him or her. Specifically, participants were given a list of descriptors proposed by the profiling system. For each descriptor, participants were asked to indicate whether it correctly describes one of his or her areas of expertise. Optionally, for a given descriptor participants could indicate their level of expertise on scale of 1 (lowest) to 5 (highest). Finally, they could leave behind any comments they wished to share. A total of 246 people self-assessed their (automatically generated) profiles. Of these, 226 indicated their levels of expertise on a scale of 1 to 5. Also, 89 participants supplied comments on the proposed profiles. In a separate study [5], we transcribe and analyze these comments through content analysis. Rather than reproducing the outcomes here, we share a selection of assessors' comments that support our proposed evaluation framework.

First, the feedback we received from our participants does signal a need for more than simply performing lexical matching. Users higher up in the organizational hierarchy, such as full professors, tend to prefer more specific expertise descriptors. One commonly mentioned reason for this is that narrower expertise descriptors tend to streamline communication and enable interested parties to directly contact the relevant expert. A specific example of this is a professor in psychopharmacology at UvT, who did not want to select the expertise keyword 'Drugs' as an expertise area, because it would result in a whole range of questions that are not part of his expertise. In contrast, teachers and research assistants at UvT tend to prefer broader terms to describe their expertise.

In the feedback we received, only one participant doubted the usefulness of rating one's expertise; 226 (out of 246) used multiple values on the rating scale. This lends credit to the idea of using graded relevance values for describing someone's expertise.

As to the importance of diversity of recommendations, several participants signaled a need for minimizing overlap in the recommended expertise descriptors. E.g., one person complained about being recommended both ‘international public law,’ ‘international law,’ and ‘international private law,’ which are all near-synonyms. A profiling system that focuses more on diversity could help avoid such problems.

3 Scoring Profiles

We present our evaluation framework in five steps, using the following notation:

- d: a descriptor (i.e., label, thesaurus term, …) that may or may not be relevant to an information need e; the output of an entity profiling system is a ranked list of descriptors $d_1, \ldots, d_k$;
- $\mathit{\Delta} = \{d_1, \ldots, d_m\}$: the set of all possible descriptors;
- e: an entity for which a profile is being sought (the user’s information need); we model e as a set of descriptors $e \subseteq \mathit{\Delta}$;
- R: a binary random variable representing relevance.

We follow Clarke et al. [8] in using the probability ranking principle as the starting point for the definition of a scoring method to be used for assessing the output of an entity profiling system. Our aim, then, is to estimate $P(R = 1|e, d)$, the probability of relevance given information need e and descriptor d.

3.1 Baseline approach

The standard way of assessing a ranked list of descriptors output by an entity profiling system is in terms of precision and recall of the descriptors retrieved [2]:

$$P(R = 1|e, d) = P(d \in e). \tag{1}$$

Traditionally, the probabilities are estimated to be 0 or 1 for particular choices of information need e and descriptor d; $P(d \in e)$ indicates that d is known to be a valid descriptor for e and $P(d \in e) = 0$ indicates that d is known not to be a valid descriptor. This traditional model only allows exact lexical matches with ground-truth descriptors.

3.2 Beyond lexical matching

We generalize (1) to allow for a more relaxed matching between a system-produced descriptor d and descriptors d_j contained in the ground truth: rather than requiring that $d = d_j$ (for some j), we ask that some d_j that is known to be relevant “provides support” for d. More precisely, we assume independence of $d_j \in e$ and $d_k \in e$ (for $j \neq k$) and reward absence of non-relevance:

$$P(R = 1|e, d) = 1 - \left(\prod_{j=1}^{m} (1 - P(d_j \in e) \cdot P(d|d_j)) \right). \tag{2}$$

Here, $P(d_j \in e)$ denotes the probability that e is correctly described by d_j and $P(d|d_j)$ denotes the probability that d_j supports d. We turn to $P(d_j \in e)$ in Section 3.3; for $P(d|d_j)$ there are several natural estimations. E.g., it could be corpus-based or a probabilistic semantic measure derived from the structure of Δ, the space of all descriptors, based on conceptual relationships.

3.3 Assessments

How should we estimate $P(d_j \in e)$? We adopt a model inspired by the way in which (topical) profiles are often determined for humans [2, 3]. We assume that a human assessor presented with information about entity e reaches a graded decision regarding each descriptor $d_j \in \Delta$. We write $grade(d_j, e) = x$ $(0 \leq x \leq 1)$ to denote that the assessor has decided to assign the value x to relevance of descriptor d_j for entity e. In the simplest case, a binary choice is made for x: $grade(d_j, e) = 0$ indicates that descriptor d_j does not apply to e, while $grade(d_j, e) = 1$ signifies that it does apply. If we assume $P(d_j \in e) = grade(d_j, e)$—a natural estimation—, then (2) becomes

$$P(R = 1|e, d) = 1 - \left(\prod_{j=1}^{m} (1 - grade(d_j, e) \cdot P(d|d_j)) \right). \tag{3}$$

3.4 Novelty

We now consider ranked lists of descriptors instead of single descriptors. Using (3) we can assign a score to the descriptor ranked first in the output of an entity profiling system. For descriptors returned at rank two and later, we view relevance conditioned on the descriptors ranked higher. We assume that relevance estimations have already been obtained for the first $k-1$ descriptors in a ranked list $d_1, \ldots, d_{k-1}$ and aim to define the relevance score of descriptor d_k returned at rank k. Let the random variables associated with relevance at each rank 1, …, k be $R_1, \ldots, R_k$. We need to estimate

$$P(R_k = 1|e, d_1, \ldots, d_k).$$

First, we estimate the degree to which support for d_k has already been provided at earlier ranks. That is, the probability that d_k contributes new information is

$$\prod_{l=1}^{k-1} \left(1 - P(R_l = 1|e, d_l) \cdot P(d_k|d_l)\right). \tag{4}$$

Here, for each descriptor d_l ranked before d_k, we determine its relevance score and use that to weight the support (if any) that d_l provides for d_k. We use (4), to replace (3) by

$$\begin{aligned} & P(R_k = 1|e, d_1, \ldots, d_k) \qquad (5) \\ & = 1 - \prod_{j=1}^{m} \left(1 - grade(d_j, e) \cdot P(d_k|d_j) \cdot \prod_{l=1}^{k-1} (1 - P(R_l = 1|e, d_l) \cdot P(d_k|d_l)) \right). \end{aligned}$$

In case the descriptors d_l (for $l \leq k-1$) in a system-produced ranking are either non-relevant or provide no supporting evidence for the descriptor d_k returned at rank k, the terms $P(R_l|e, d_l)$ or $P(d_k|d_l)$ all attain the value 0, so that (5) reduces to (3).

3.5 Aggregating

Finally, we aggregate scores of individual descriptors into a score for ranked lists of descriptors. Discounted cumulative gain has become a standard evaluation measure when graded scores are available [11]. "Standard" notions such as gain vector, cumulative gain (CG) vector and discounted cumulative gain (DCG) vector can easily be defined, using the score produced by (5) as elements in the gain vector. I.e., the k-th element of the gain vector G is $G[k] = P(R_k = 1|e, d_1, \ldots, d_k)$. And the cumulative gain vector is $CG[k] = \sum_{j=1}^{k} G[k]$, while the discounted cumulative gain is defined as $DCG[k] = \sum_{j=1}^{k} G[j]/(\log_2(1+j))$. Producing the ideal cumulative gain vector (needed for computing the normalized DCG) is more complex; various approximations have been proposed. Clarke et al. [8] propose a variant α-nDCG, where α reflects the possibility of assessor error. This extension can be incorporated by modifying (3).

4 Related Work

Nugget-based evaluation methodologies have been used in a number of large-scale evaluation efforts. For the "other" questions considered at TREC 2004 and 2005, systems had to return text snippets containing important information about a topic; if the topic at hand is an entity, this boils down to entity profiling. System responses consist of passages extracted from a document collection. To evaluate responses, human assessors classify passages into essential, worthwhile (but not essential) and non-relevant [9, 15, 16]. Several authors have examined this methodology; see, e.g., [13]. A variation was considered at CLEF 2006 [12], where evaluation was again nugget-based.

Balog and de Rijke [2] considered a descriptor-based version of a specific entity profiling task, viz. expert profiling, where characteristic expertise descriptors have to be returned. Evaluation was carried out in terms of precision and recall computed against a gold standard set of descriptors for each test topic. This type of entity profiling may be viewed as a generalization of the keyphrase extraction task (assign keyphrases, typically taken from a fixed source, to a document) [10] or a variation of the query labeling task (assign labels, typically taken from a fixed knowledge source, to a query) [14].

Our evaluation framework is based on insights from [11, 12] and especially [8]. Our proposal differs from that of Clarke et al. [8] in that the items we need to assess are descriptors (not documents) and that we allow matches between system produced descriptors with gold standard descriptors to be non-exact and weighted by a semantic distance measure that fits the domain or knowledge source at hand. The theme of combining relevance assessment with semantic distance is an old one, however, going back at least 15 years [7]; to our knowledge it has so far not been applied to entity profiling.

5 Conclusion

We have developed a new scoring method for descriptor-based entity profiles that addresses many of the shortcomings of today's prevalent evaluation method. Our method allows for weighted non-exact matches between system-produced and ground-truth descriptors and it systematically rewards more highly relevant descriptors and diverse lists of descriptors. Aggregation of individual descriptor scores was done using the discounted cumulative gain measure. In future work we will explore the creation of a test collection based on the metric introduced here, taking the graded assessments in [5] as our starting point, with comparisons of different implementations of $P(d|d_j)$.

Acknowledgements This research was 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, by the Radio Culture and Auditory Resources Infrastructure Project (LARM) funded by the Danish National Research Infrastructures Program (project nr. 09-067292), by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, and by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.066.512, 612.061.814, 612.061.815, 640.004.-802.

Bibliography

[1] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectivenss measures and user satisfaction. In *SIGIR '07*, pages 773–774. ACM, 2007.

[2] K. Balog and M. de Rijke. Determining expert profiles (with an application to expert finding). In *IJCAI-2007*, pages 2657–2662, 2007.

[3] 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.

[4] K. Balog, L. Azzopardi, and M. de Rijke. A language modeling framework for expert finding. *Inf. Proc. & Manag.*, 45(1):1–19, January 2009.

[5] K. Balog, T. Bogers, A. van den Bosch, and M. de Rijke. Expertise profiling in a knowledge-intensive organization. *Submitted*, 2010.

[6] T. Bogers and K. Balog. UvT Expert Collection documentation. Technical report, ILK Research Group Technical Report Series no. 07-06, July 2007.

[7] T. A. Brooks. Topical subject expertise and the semantic distance model of relevance assessment. *Journal of Documentation*, 51:370–387, 1995.

[8] C.L.A. Clarke, M. Kolla, G.V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR '08*, pages 659–666. ACM, 2008.

[9] H.T. Dang, J. Lin, and D. Kelly. Overview of the TREC 2006 question answering track. In *TREC 2006*. NIST, 2006.

[10] K. Hofmann, E. Tsagkias, E. J. Meij, and M. de Rijke. The impact of document structure on keyphrase extraction. In *CIKM 2009*, Hong Kong, November 2009. ACM.

[11] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[12] V. Jijkoun and M. de Rijke. Overview of the WiQA task at CLEF 2006. In *Evaluation of Multilingual and Multi-modal Information Retrieval*, pages 265–274. Springer, 2007.

[13] J. Lin and D. Demner-Fushman. Will pyramids built of nuggets topple over? In *HLT'06*, pages 383–390, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[14] E. J. Meij, M. Bron, B. Huurnink, L. Hollink, and M. de Rijke. Learning semantic query suggestions. In *ISWC 2009*. Springer, October 2009.

[15] E.M. Voorhees. Overview of the TREC 2004 question answering track. In *TREC 2004*. NIST, 2005.

[16] E.M. Voorhees and H.T. Dang. Overview of the TREC 2005 question answering track. In *TREC 2005*. NIST, 2006.