Information Processing and Management 52 (2016) 529–549

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Formal language models for finding groups of experts☆

Shangsong Liang[1,*], Maarten de Rijke

Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 24 December 2014
Revised 10 October 2015
Accepted 25 November 2015
Available online 2 February 2016

Keywords:
Group finding
Entity retrieval
Enterprise search

ABSTRACT

The task of finding groups or teams has recently received increased attention, as a natural and challenging extension of search tasks aimed at retrieving individual entities. We introduce a new group finding task: given a query topic, we try to find knowledgeable groups that have expertise on that topic. We present five general strategies for this group finding task, given a heterogenous document repository. The models are formalized using generative language models. Two of the models aggregate expertise scores of the experts in the same group for the task, one locates documents associated with experts in the group and then determines how closely the documents are associated with the topic, whilst the remaining two models directly estimate the degree to which a group is a knowledgeable group for a given topic. For evaluation purposes we construct a test collection based on the TREC 2005 and 2006 Enterprise collections, and define three types of ground truth for our task. Experimental results show that our five knowledgeable group finding models achieve high absolute scores. We also find significant differences between different ways of estimating the association between a topic and a group.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

A major challenge within any organization is managing the expertise of formal or informal groups of people within the organization such that groups with expertise in a particular area can be identified (Balog, Azzopardi, & de Rijke, 2006; Juang, Huang, & Huang, 2013). Rather than finding knowledgeable individuals, sometimes locating a *group* with appropriate skills and knowledge in an organization is of great importance to the success of a project being undertaken (Lappas, Liu, & Terzi, 2009; Li, Shan, & Lin, 2013; Neshati, Beigy, & Hiemstra, 2014). For instance, an engineering organization may want to find a group of scientists who have expertise for dealing with technical problems when constructing a long high-speed railway without having to trawl through descriptions of individuals or groups (if there are any). A group of doctors in a hospital may have to be found immediately so as to perform an operation for a seriously-ill patient. Identifying the right groups of experts with specific knowledge for a task at hand may reduce costs and save the lives of people.

Finding a group or a team that harbors expertise is different from first finding an expert and then sorting out to which team the expert belongs. Conceptually, the difference is that finding a group mainly focuses on how to collect evidence so

☆ This paper is a substantially revised and extended version of Liang and de Rijke (2013). New models have been added, the formal details of all models considered are included, and we report on substantially more elaborate experiments than in Liang and de Rijke (2013). The code to support the experiments in this paper is available from http://ilps.science.uva.nl/resources.

* Corresponding author. Tel.: +31684987399.
E-mail addresses: s.liang@uva.nl (S. Liang), derijke@uva.nl (M. de Rijke).
[1] Now at University College London.

http://dx.doi.org/10.1016/j.ipm.2015.11.005
0306-4573/© 2015 Elsevier Ltd. All rights reserved.

as to make a decision on whether the group is knowledgeable on the topic, whilst the approach of first locating an expert is mainly focused on collecting evidence to make a decision on whether the expert has expertise on the topic. Technically, there are important differences too. For instance, as we will see below, a group finding model that finds knowledgeable groups via documents directly is significantly outperformed by models that decompose the problem differently.

Traditional approaches to finding knowledge, whether in individuals or in groups within an organization, usually include two main steps. For a given task the expertise of the experts in a group is recorded and then the expertise of a group is computed by aggregating the expertise values of all group members. Both steps are traditionally done manually and require considerable effort to set up and maintain. In addition, this approach is usually restricted to a fixed set of expertise areas, making it hard to find knowledgeable groups in areas not explicitly coded (Pryor, Myles, Williams, & Anand, 1988).

To reduce the effort of recording and evaluating the expertise of people from their representations, many automatic approaches have been proposed. There has been an increasing move to automatically extract such representations for evaluating expertise from heterogeneous document collections, such as conference papers, corporate intranets and community question answering collections (Balog, Fang, de Rijke, Serdyukov, & Si, 2012).

To compute the expertise values of a group, in principle, many aggregation operators are available, such as maximum, sum, or average. These can simply be employed to combine the expertise values of each expert within a given group. There are at least 90 families of aggregation operators (Zhou, Chiclana, John, & Garibaldi, 2011); they have been put to use in a range of applications, e.g., in clustering (Beliakov, James, & Li, 2011), image segmentation (Ghosh, Kothari, Halder, & Ghosh, 2009), and control (Senge & Hullermeier, 2011). However, a solution to the problem of how to aggregate expertise values of all experts within a group so that the expertise scores of different groups can easily be compared and ranked by using a suitable aggregation operator, is still unknown.

We treat the problem of finding a knowledgeable group differently. Five distinct models are proposed. Our models are based on probabilistic language modeling techniques, which have been successfully applied in a range of related Information Retrieval (IR) tasks, such as ad hoc retrieval (Ponte & Croft, 1998; Zhai & Lafferty, 2004), expert finding (Balog et al., 2006; 2009; Balog et al., 2012; Fang, Si, & Mathur, 2010), similar people finding (Weerkamp et al., 2011), and republished article finding (Tsagkias, de Rijke, & Weerkamp, 2011). Language models are attractive because of their foundations in statistical theory, the great deal of complementary work on language modeling in speech recognition and natural language processing, and the fact that very simple language modeling applied to retrieval problems tends to perform very well empirically (Balog, Azzopardi, & de Rijke, 2009). Each group finding model that we consider ranks groups according to the probability of a group being a knowledgeable group given the query topic, but the models differ in how this is performed. Three types of variables play a key role in our estimations: groups (G), queries (Q) and documents (D). The order in which we estimate these is reflected in our naming conventions. E.g., the model named GDQ proceeds by first collecting evidence of whether a group is knowledgeable about the topic via the experts in the group (G), and then determining whether each expert in the group has expertise on the topic via documents (D), and finally whether a document is talking about the given query (Q) topic. In our Group-Query-Document (GQD) model and Group-Document-Query (GDQ) model, the expertise scores of each expert in a group are computed first and then aggregated into an overall score. These two models differ in the way in which they compute expertise scores for individual experts; in both cases, the experts in a group act as a latent variable between the group and the query. In our Document-Group-Query (DGQ) model, documents are ranked according to the query, and then we determine how likely a group is a knowledgeable group by considering the set of documents associated with them. Here, the documents act as a latent variable between the query and the group. Our last two models, the Query-Group-Document (QGD) model and the Query-Document-Group (QDG) model, rank groups according to the query, and then we determine how likely a person in the group is a knowledgeable expert by considering the set of documents associated with the expert; in these two models, it is the query that acts as a latent variable between the group and the experts in the group; we find that the QGD model actually yields the same ranking as the GQD model. Unlike early automatic group finding systems that tended to focus on specific document genres only, such as email (Campbell, Maglio, Cozzi, & Dom, 2003) or software and software documentation (Mockus & Herbsleb, 2002) to build profiles and find the entities, e.g., experts, our group finding algorithms can work on heterogeneous document genres and the profiles of groups and experts are not required to be given in advance.

For evaluation purposes, we use data from both the TREC 2005 and 2006 Enterprise tracks to create our test sets. As the data sets were created for expert finding (as opposed to knowledgeable group finding), some additional work is needed to turn them into a test set for group finding. We define three types of ground truth for our knowledgeable group finding task, implementing three readings of what makes a group a knowledgeable group. Familiar retrieval metrics such as NDCG, NDCG@k, MAP, and p@k are applied as evaluation metrics in our experiments. We perform a range of experiments to analyze our proposed knowledgeable group finding models, and find that some of our models perform similarly according to one metric but not according to another. E.g., GDQ and DGQ models are not statistically significantly different when using NDCG as a performance metric on our datasets; but when using MAP as our metric, the observed differences are statistically significant. Our main research goals in experimentation are to understand how the five models listed above compare.

In summary, the contributions of this paper are the following:

(i) We introduce a new information retrieval task: given a topic, find knowledgeable groups that have expertise on the topic.
(ii) We propose five language modeling approaches to tackle the challenge of automatically finding knowledge groups in heterogeneous document collections.

Download English Version:

https://daneshyari.com/en/article/514942

Download Persian Version:

https://daneshyari.com/article/514942

Daneshyari.com