Finding Experts in Community-Based Question-Answering Services

Xiaoyong Liu, W. Bruce Croft
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst, MA 01003
{xliu, croft}@cs.umass.edu

Matthew Koll
Wondir, Inc.
7735 Old Georgetown Rd., Ste. 1200
Bethesda, MD 20814
matt@wondir.com

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process.*

H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *User profiles and alert services.*

General Terms Experimentation, Measurement

Keywords Expert finding, information retrieval, expertise modeling, community-based Web information service, digital reference, information systems, language models

1. INTRODUCTION

There have been a growing number of Web information services that bring together a network of self-declared "experts" to answer other people's questions. This started as digital reference services such as the Mad Scientist Network[1], but has now become a popular part of several Web search services, including Google Answers[2] and All Experts[3]. One such service, called Wondir[4], is a free, publicly available, live question and answer engine that connects people with questions to people with answers. People using such services are like a community – anyone can ask, anyone can answer, and everyone can share, since all of the questions and answers are public and searchable immediately. We refer to this type of services as community-based question-answering (QA) services. There are hundreds of questions asked each day but some portion of them may not be answered or there may be a lag between the time when a question is asked and when it is answered. To get fast, relevant answers, the key is getting the right question in front of the right person. The goal of our work is to investigate how the expertise of users, or "experts", can be captured, and when combined with state-of-the-art information retrieval techniques whether the system is able to identify the group of "experts" who are likely to provide answers to given questions.

The expert finding problem has been explored in the research communities of Digital Reference [1,2] and Knowledge Management [3]. Our work is different from previous efforts in that we focus on automatically finding experts in an open-domain community-based QA service, and the expert finding task is evaluated on large-scale, real data.

[1] http://www.madsci.org
[2] http://answers.google.com/answers/
[3] http://www.allexperts.com
[4] http://www.wondir.com

CIKM'05, October 31-November 5, 2005, Bremen, Germany.
ACM 1-59593-140-6/05/0010.

2. METHODOLOGY

2.1 Data

The data for this study comes from the log of the questions and answers submitted to the Wondir QA service between Oct. 2, 2002 and Feb. 15, 2005. We created a pool of 852,316 QA pairs, and derived 5 data sets each having a different requirement on the minimum number of questions each user must be associated with[5]. For example, the set D5 is all the QA pairs in the pool that are associated with users who answered at least 5 questions. For each data set, we created a test set by randomly selecting one question for each expert included in that set. The remaining questions and corresponding answers form the training set. Questions in the test set are queries for experiments. The QA pairs in the training set are used to create different expert profiles. Relevance judgments are generated by taking the users who actually answered the questions in the test set. Statistics of the data sets are given in table 1.

2.2 Experimental Method

We cast the expert finding problem as an IR problem. Given a question, we define an "expert" as a person who has answered similar questions in the past in the system. The expertise of a person is characterized using a profile that has been derived from the previously answered questions. The given question can be viewed as query and the expert profiles can be viewed as documents. These profiles are ranked using language models which are representative of state-of-the-art information retrieval techniques. More specifically, the language models we used in this work are: the query likelihood model [4], the relevance model [5], and the cluster-based language model [6]. People whose profiles are ranked higher are considered more likely to be experts for answering the given question.

Depending on the text that is used, expert profiles can be built from: 1) all previously answered questions by a user, both question and answer texts (i.e. "All QA pairs" in table 2); 2) all previously answered questions, question texts only (i.e. "All Qs"); 3) one of the previously answered questions, both question and answer texts (i.e. "Single QA pair"); Or, 4) one of the previously answered questions, question text only (i.e. "Single Q"). Note that by using 3) and 4) we could have multiple profiles for each expert – they can be viewed as different versions of the profile. At the time of retrieval, a language model is computed for each version, and experts are ranked based on the score of their best profile version.

[5] An expert is associated with a question if he/she provided an answer.

Table 1. Statistics of data sets

Data Set ID	Total # of QA pairs	# of "experts"	Avg. # of questions per expert on entire set	Avg. # of answerers per question on entire set	# of test questions	Avg. query length (in # of words) after stemming and stopping	# of QA pairs in training data
D2	805,898	37,723	21.4	2.0	23,949	9.5	778,667
D5	752,381	17,525	43.0	1.9	14,795	9.8	736,490
D20	639,233	5,025	127.2	1.7	4,900	9.9	633,897
D50	547,668	2,017	271.7	1.6	1,997	9.9	545,650
D100	474,185	958	495.2	1.6	954	10.1	473,225

Table 2. Results for using the query likelihood (QL) model to rank experts. Evaluation measure is MRR.

Data Set ID	Expert Profiles			
	All QA pairs	All Qs	Single QA pair	Single Q
D2	0.1152	0.1228	0.1285	0.1300
D5	0.1115	0.1193	0.1266	0.1282
D20	0.1002	0.1079	0.1037	0.1041
D50	0.0885	0.0936	0.0855	0.0832
D100	0.0889	0.0907	0.0871	0.0849

In all experiments, both the queries and documents are stemmed, and stopwords are removed. The Mean Reciprocal Rank (MRR) measure [7] is used for evaluation.

3. EXPERIMENTS AND RESULTS

The first set of experiments investigates how well experts can be ranked when each of the four different profile configurations is used. The query likelihood model is applied to produce the ranking. Results are given in table 2. We observe that, on each data set, for runs with profiles considering single questions, the performance is very similar between using QA pair and Q only. For runs with profiles considering all previously answered questions, using Qs only gives better performance than using QA pairs, with an average of 6.1% difference in MRR score. The results of using "All Qs" are comparable to those of using "Single QA pair" or "Single Q". In general, performance tends to go up when the requirement on the minimum number of questions each expert should have answered in the past drops. The best-performing single run is on the D2 set with the "Single Q" profile configuration, which has a MRR score of 0.13. The profile configuration "All Qs" seems to give the most consistent performance across different data sets. The next set of experiments compares the performance of different language models in ranking experts. The results are shown in table 3. All three models can rank the true answerer within rank 9 (with over 0.11 in MRR score). The performances of QL and CBDM are comparable and they are both better than that of RM. Across all data sets, the performance of all three language models improves as the minimum number of questions each expert must be associated with decreases. The best performance is achieved on the D2 set.

At first sight, the MRR scores are not as high as some of those reported in the TREC QA track. Considering the task at hand, however, we feel that the results obtained in these experiments are very reasonable, because ranking experts is very different from ranking answers in a typical QA system. For example, in the TREC QA track, there are straightforward correct answers for most test questions, and the number of correct answers to each question is typically small. In the expert finding task that we discussed in this paper, however, there is no such thing as a "correct" expert. All we know is who actually answered a question, but not who possesses the knowledge for that question. Therefore the relevance judgment set that considers relevant only the true answerers of a question suffers from serious incompleteness as there are possibly many experts that possess the knowledge about a given topic but only a very small number of them actually answered the question. We started investigating a possible solution to this problem, which is to boost the relevance judgment set by exploiting hierarchical clustering methods to group experts based on their profiles.

Table 3. Results for different retrieval models. Expert profiles are "All Qs". Evaluation measure is MRR.

Data Set ID	Retrieval Methods		
	Query likelihood (QL)	Relevance model (RM)	Cluster-based language model (CBDM)
D2	0.1228	0.1172	0.1253
D5	0.1193	0.1126	0.1198
D20	0.1079	0.0982	0.1082
D50	0.0936	0.0837	0.0928
D100	0.0907	0.0779	0.0900

4. CONCLUSIONS

We have experimented with state-of-the-art information retrieval (IR) techniques and different ways of building profiles for finding experts in an open-domain community-based QA service. Language models have been chosen as representative of state-of-the-art IR techniques in this work but other retrieval techniques can also be applied. Among the four different profile configurations, the one that considers all previously answered questions with question texts only seems to give the most consistent performance across different data sets. Results have shown that reasonable performance for ranking experts can be achieved when language models are combined with this type of profiles. For future work, we plan to carry out more experiments with boosted relevance judgment set and possibly other techniques to expand expert profiles.

5. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by Advanced Research and Development Activity and NSF grant #CCF-0205575, and in part by NSF grant number DUE-0226144.

6. REFERENCES

[1] McClennen, M., and Memmott, P. (2001). Roles in digital reference. *Information Technology and Libraries 20*, pp. 143-148.

[2] Pomerantz, J., Nicholson, S., and Lankes, R. D. (2003). Digital reference triage: Factors influencing question routing and assignment. *The Library Quarterly*, 73(2), pp. 103-120.

[3] Yimam Seid, D., and Kobsa, A. (2003). Expert finding systems for organizations: Problem and domain analysis and the DEMOIR approach. *Journal of Organizational Computing and Electronic Commerce*, 13(1):1-24.

[4] Miller, D., Leek, T., and Schwartz, R. (1999). A hidden Markov model information retrieval system. In *Proceedings of SIGIR'99*.

[5] Lavrenko, V. and Croft, W.B. (2001). Relevance-based language models. In W.B. Croft, D.J. Harper, D.H. Kraft, & J. Zobel (Eds.), In *Proceedings of SIGIR 2001,* pp.120-127.

[6] Liu, X., and Croft, W.B. (2004). Cluster-based retrieval using language models. In *Proceedings of SIGIR 2004*.

[7] Voorhees E. M. (1999). The TREC-8 Question Answering Track Report. In *Proceedings of TREC-8*.