Finding Experts and their Details in E-mail Corpora

Krisztian Balog Maarten de Rijke
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
kbalog,mdr@science.uva.nl

ABSTRACT

We present methods for finding experts (and their contact details) using e-mail messages. We locate messages on a topic, and then find the associated experts. Our approach is unsupervised: both the list of potential experts and their personal details are obtained automatically from e-mail message headers and signatures, respectively. Evaluation is done using the e-mail lists in the W3C corpus.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.4 [**Information Systems Applications**]: H.4.2 Types of Systems; H.4.m Miscellaneous

General Terms

Algorithms, Management, Experimentation

Keywords

Expert search, Expert finding, E-mail processing

1. INTRODUCTION

E-mail has become the primary means of communication in many organizations. It is a rich source of information that could be used to improve the functioning of an organization. Hence, search and analysis of e-mail messages has drawn significant interest from the research community [5, 2].

Specifically, e-mail messages can serve as a source for "*expertise identification*" [1], since they capture people's activities, interests, and goals in a natural way.

While early approaches to *expert finding* (i.e., identifying experts on a given topic) employed manually maintained databases, there has been a move towards unsupervised methods that use *expertise indicators* from documents produced within an organization; the resulting evidence of expertise is then used to build an employee's expertise profile.

Our main aim in this paper is to study the use of e-mail messages for mining expertise information. Our main findings are that (i) the fielded structure of e-mail messages can be effectively exploited to find pieces of evidence of expertise, which can then be successfully combined in a language modeling framework, and (ii) e-mail signatures are a reliable source of personal contact information.

The rest of the paper is structured as follows. In Section 2 we detail and assess our model of expert search. In Section 3 we harvest contact details for candidates by mining e-mail signatures. We conclude in Section 4.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

2. FINDING EXPERTS

We model the expert finding task as follows: *what is the probability of a candidate ca being an expert given the query topic q?* Instead of computing this probability $p(ca|q)$ directly, we can use Bayes' Theorem to rank candidates in proportion to $p(q|ca)$, the probability of the query given the candidate. Below, we first detail our model, and then evaluate its effectiveness.

2.1 Modeling Expert Search

We first find documents (i.e., e-mail messages) which are relevant to the query topic and then score each candidate by aggregating over all documents associated with that candidate. That is, $p(q|ca) \propto \sum_d p(q|d)p(ca|d)$. To determine $p(q|d)$, the probability of a query given a document, we use a standard language modeling for IR approach. To estimate the strength of the association between document d and candidate ca, $p(ca|d)$, we assume that an association score $a(d, ca)$ has been calculated for each document d and for each candidate ca. To turn these associations into probabilities, we put $p(ca|d) = a(d, ca)/(\sum_{d_i \in D} a(d_i, ca))$, where D is a set of e-mail messages.

To compute the associations $a(d, ca)$ we exploit the fact that our documents are e-mail messages. A list of candidate experts is created by extracting names and e-mail addresses from message headers. We introduce four binary association methods for deciding whether a document d and candidate ca are associated:

- A_0 `EMAIL_FROM` returns 1 if the candidate appears in the *`from`* field of the e-mail
- A_1 `EMAIL_TO` returns 1 if the candidate appears in the *`to`* field of the e-mail
- A_2 `EMAIL_CC` returns 1 if the candidate appears in the `cc` field of the e-mail
- A_3 `EMAIL_CONTENT` returns 1 if the candidate's name appears in the content of the e-mail message. The first and last names are obligatory; middle names are not.

Since A_0–A_3 are likely to capture different aspects of the relation between a document and a candidate expert, we also

consider (linear) combinations of their outcomes. Hence, we put $a(d, ca) := \sum_{i=0}^{3} \pi_i A_i(d, ca)$, where the π_i are weights.

2.2 Experimental Evaluation

We carried out experiments to answer the following question: how effective is our modeling approach for finding experts? The document collection we use is part of the W3C corpus [4], which was used at the 2005 TREC Enterprise track [3] and comes with a list of candidate experts, expert finding topics, and relevance assessments for these topics. For the purposes of our experiments, we restrict ourselves to the e-mail lists in the corpus, omitting other types of documents from the W3C corpus and candidate expert names that do not occur in the e-mail lists.

We conducted two sets of experiments: comparing the impact of the association methods on expert finding effectiveness, and examining the impact of combinations of these association methods. Table 1 contains the expert finding results for different association methods. The most effective association method is A_0 (`EMAIL_FROM`), on all measures.

association	%rel	map	P@5	P@10	P@20	RR1
`EMAIL_FROM`	**62.2**	**0.233**	**0.270**	**0.241**	**0.180**	**0.447**
`EMAIL_TO`	61.8	0.211	0.262	0.229	0.177	0.424
`EMAIL_CC`	53.4	0.157	0.220	0.202	0.155	0.376
`EMAIL_CONTENT`	61.1	0.174	0.175	0.173	0.152	0.272

Table 1: Finding experts in the W3C e-mail lists. Columns: association method, fraction of relevant experts found, Mean Average Precision, Precision at 5, 10, 20 candidates found, and reciprocal rank of the top relevant result. Best scores in boldface.

Assuming that different associations perform in complementary ways, we explored linear combinations of association methods; Table 2 reports a sample of results. Briefly, the main findings are (i) using the `EMAIL_CONTENT` method improves on the number of retrieved candidates, but hurts on other measures; (ii) extra weights on a single header field improves, but only on a subset of the measures; (iii) our best found combination (bottom row) improves on all measures. Surprisingly, the `cc` field has a great importance when it is used within a combination; the person being cc'd appears to be an authority on the content of the message.

combination	%rel	map	P@5	P@10	P@20	RR1
single bests	62.2	0.233	0.270	0.241	0.180	0.447
1+1+1+1	62.7	0.183	0.170	0.175	0.163	0.286
2+1+1+0	65.0	**0.242**	0.267	0.238	0.187	**0.455**
1+2+1+0	**65.6**	0.236	0.263	0.238	0.189	0.424
1+1+2+0	65.0	0.238	0.270	**0.248**	**0.193**	0.448
1.5+1+2.5+0	65.2	0.239	**0.279**	0.244	**0.193**	0.452

Table 2: Finding experts in the W3C e-mail lists. Same measures as in Table 1. First row: best result for each measure using a single association method. Rows 2–6 lists sample combinations; the first column shows the weights used for `EMAIL_FROM`, `EMAIL_TO`, `EMAIL_CC`, `EMAIL_CONTENT`, respectively. Best scores in boldface.

3. MINING CONTACT DETAILS

Once an expert has been determined, retrieving his/her contact details is a natural next component of an operational expert finder. We show that contact details can be effectively mined from e-mail signatures.

3.1 Extracting Signatures

One of the challenges of expert profiling is to maintain a database with the candidates' details. To address the issue, we mine the e-mail signatures. Many (but by no means all) contain reliable details about a person's affiliation and contact details.

Before mining signatures, we need to identify them. Our heuristics are precision-oriented; using the following heuristics we find a large number of signatures with valuable personal data: (i) signatures are placed at the end of the e-mails and separated from the message body with `"--"`; (ii) the length of a signature should be between 3 and 10 lines; (iii) it should contain at least one web address or tel/fax number; and (iv) signatures containing stop words (P.S., antivirus, disclaimer, etc.) or PGP keys are ignored.

3.2 Statistics on Experts' Details

How effective are our unsupervised methods for extracting personal information? Table 3 details the results of our signature mining experiments. ALL refers to all people found within the corpus, while W3C refers to people found that were on the list of candidate experts, provided by TREC. We restricted our identification method to find people that appear more than 5 times in e-mail headers.

	ALL	W3C
signatures extracted	54.533	15.514
unique signatures	12.544	3.447
people identified	2.708	326
personal data found in signatures	1.492	246

Table 3: Identifying people and extracting personal data from the W3C e-mail lists corpus.

4. CONCLUSIONS AND FURTHER WORK

We have presented methods for expertise identification using e-mail communications. Our expert modeling approach uses language modeling techniques and combines evidences of expertise. This method is very effective in terms of the number of relevant experts found. Possible further improvements concern determining more expertise indicators and using the thread structure of the e-mail lists. Our extraction method finds contact information for candidates using email signatures. In future work we plan to extract additional details, such as affiliation and address information.

5. REFERENCES

[1] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 528–531. ACM Press, 2003.

[2] K. Mock. An experimental framework for email categorization and management. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 392–393, New York, NY, USA, 2001. ACM Press.

[3] Enterprise track, 2005. URL: `http://www.ins.cwi.nl/projects/trec-ent/wiki/`.

[4] W3C. The W3C test collection, 2005. URL: `http://research.microsoft.com/users/nickcr/w3c-summary.html`.

[5] S. Whittaker and C. Sidner. Email overload: exploring personal information management of email. In *CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 276–283, New York, NY, USA, 1996. ACM Press.