Indexing Emails and Email Threads for Retrieval

Yejun Wu and Douglas W. Oard
College of Information Studies and UMIACS
University of Maryland, College Park, MD 20742
{wuyj,oard }@glue.umd.edu

ABSTRACT

Electronic mail poses a number of unusual challenges for the design of information retrieval systems and test collections, including informal expression, conversational structure, variable document granularity (e.g., messages, threads, or longer-term interactions), a naturally occuring integration between free text and structural metadata, and incompletely characterized user needs. This paper reports on initial experiments with a large collection of public mailing lists from the World Wide Web consortium that will be used for the TREC 2005 Enterprise Search Track. Automatic subject-line threading and removal of duplicated text were found to have little effect in a small pilot study. Those observations motivated development of a question typology and more detailed analysis of collection characteristics; preliminary results for both are reported.

Categories and Subject Descriptors

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

General Terms

Performance, Experimentation

Keywords

Indexing, email, threads, retrieval

1. INTRODUCTION

Electronic mail (email) was designed to support one-to-one asynchronous communications; mailing lists extend email to support one-to-many communications. Mailing lists are often archived, and those archives can conceivably serve as an important source of institutional memory if appropriate search technology is available. In 2005, the TREC Enterprise Search Track (TRECENT) will begin the process of developing the world's first information retrieval test collection for email. An information retrieval test collection includes documents, topics, and relevance judgments that are created in a way that models some actual information access task. At present, however, we do not yet know exactly what a document or a topic means in this context. In this paper, we begin to explore those questions using the archived World Wide Web Consortium (W3C) mailing lists that will be the focus of the email search task in the 2005 TRECENT.

At first glance the question "what is a document?" would seem easy to answer; a natural goal would be to find email messages. Two simple examples reveal the limitations of that view. Imagine, for example, a user wishes to know when it was agreed that the HTML 3.0 standard would require further revision. If the answer happened to be in a message for which the entire body text was "OK - lets do it!," even a human assessor might have trouble determining whether that message is relevant. In this case, conversational threading might help. Or consider the question "Who were the most active participants early in the Semantic Web discussions?" It would be hard to answer questions like that from the content of individual messages, but if all messages sent from an email address were indexed together as a "document" then active participants in discussions on any topic could be identified fairly easily. So the first question that we explore in this paper addresses document granularity: does threading the messages that have nearly identical subject lines contribute to improved retrieval effectiveness?

As the examples above indicate, the "right" document granularity depends on the nature of the question, but we are aware of no published studies that explicate a question typology for retrospective access to email collections. We have therefore started to develop such a typology by mining actual questions from the W3C mailing list archives and through discussions with historians, archivists, information system managers, and lawyers. This paper presents some initial results from that ongoing effort.

2. THE W3C COLLECTION

The W3C collection was crawled from w3c.org in June 2004; we obtained the collection from NIST. Each message in the collection was embedded in a Web page with extensive XML markup (usually generated by the hypermail utility program) to format the most important fields from the message for display to end users. We used a Java SAX parser to recover the original RFC-822 header structure and to extract the body text. This yielded 161,645 messages with a total size of 474 MB.

Email threads have the potential to organize groups of messages around a single topic [1]. We automatically identified threads (reply chains) by linking messages with identical nontrivial subject lines (after removal of any sequence of "re:", "fw:", and "fwd:" prefixes). This yielded 21,071 multi-message threads and 62,142 single messages. The mean

SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

	Q1	Q2	Q3	Avg
Msg+	4	8	6	6.0
Thd+	4	9	5	6.3
Thd-	2	10	5	5.7

Table 1: Relevant messages/threads in top 10, for messages (Msg+), threads (Thd+), and threads with included text suppressed (Thd-).

thread length (over multi-message threads) was 4.7 messages, with a median thread length of 3 messages.

Heuristics for automatically tagging text that originally appeared in another message were developed based on inspection of the collection. Two types of quoted text were tagged, (1) lines that start with > (or | >, “ >, “), and (2) lines below “Forwarded message”, “Original message”, “Mensagem original”, “Mensaje original”, or “In/On/At (time) (somebody) wrote/ writes/ said”, etc. While imperfect, these heuristics serve as a useful basis for beginning to explore the effect of subject line threading and suppression of (probably duplicated) included text within those threads.

3. INITIAL EXPERIMENTS

We created three indicies using Lucene, a text search engine library written in Java [2]: one for individual messages, one for threads (which could be single messages) with included text retained, and one for threads (which could be single messages) with included text removed. The following fields were indexed: the date and time when the email was sent, person names (parsed from the to, from, and cc fields), email addresses (parsed from the same fields), text from the subject field, and text from the body.

One possible use of archived mailing lists in technical organizations would be to recover evidence of design rationale when making future changes to standards or products. We therefore developed 3 design rationale questions by introspection after a cursory examination of the collection. In Table 1 we report the number of relevant messages in the top 10 documents, as assessed by the first author of this paper, that results from presenting as a query the full text of the following three questions:

Q1: Who participated in the discussion of HTML4.0 or HTML4 cascading style sheets?

Q2: What problems (bugs, errors) have been reported when converting HTML to XML using Tidy?

Q3: Who agreed with the decision that the reader rather than the author should have higher priority of cascade in the cascading style?

Little difference was observed across the three indices, leading us to conclude that, while threading and included text may ultimately prove to be helpful for some types of questions, we may want to initially focus our efforts in other directions. We did, however, note that searching the thread index yielded substantially longer threads than is typical for the collection as a whole. Only 10 of 30 top-ranked threads (33%) were composed of a single document, compared to 75% for the collection as a whole. For the highly ranked multi-document threads, the average thread length was 9 (compared to 4.7) and the median thread length for highly ranked threads was 5 (compared to 3). This enrichment effect is not surprising; modern weighting functions exhibit some preference for longer documents.

One beneficial effect from threads was observed in our experiments. Near-duplicate emails are not uncommon in email collections (such as the W3C collection). By near-duplicates, we mean emails sent by the same person at different times with almost the same content. Duplicates of relevant emails were judged as relevant in our experiments, reflecting the usual practice in TREC. Our results on the message index may therefore overstate somewhat the degree to which a real user would be satisfied with the highly ranked documents. This problem did not generally arise in the same way with threads, however, because near-duplicates often appeared together in the same thread. From this we conclude that some aspects of threading that may help with the design of easily navigated user interfaces may not be captured well by typical retrieval effectiveness measures. This suggests that careful consideration should be given to the selection and interpretation of evaluation measures.

We also tried one question requesting contact information: “What’s Dan Connolly’s phone number?” We used “Dan Connolly” as a query to search only person names in the “from” field. Our system found 4 relevant messages in the top 10, but 2 of them occur in a single thread, and the other 2 did not appear in the top 10 in thread retrieval. Clearly, care must be taken when comparing message-based and thread-based measures, since looking at this case as a reduction from 4 in 10 to 1 in 10 would be misleading.

4. TOWARDS A TOPIC TYPOLOGY

Here is an incomplete list of the question types we have identified that might be appropriate for the W3C collection. The question typology was developed by mining actual questions from the W3C mailing list archives and through discussions with historians, archivists, information system managers, and lawyers.

Who participated in the discussion of issue A?
Who made the decision on issue A?
What was the decision on issue A?
When did somebody start to be involved in issue A?
Where was somebody working when s/he participated in the discussion of issue A?
How was the decision made on issue A?
Why did people talk about issue A?

5. FUTURE WORK

Much remains to be done. First, we must achieve consensus on the question typology. Once that is agreed, we will need to develop some form of test collection to support formative evaluation. With that in hand, we can begin to explore some of the questions that make email retrieval particularly interesting, such as speech act classification (e.g., reducing the weight of questions that are never answered) and accommodations for informal expression (e.g., spelling correction and acronym expansion).

6. REFERENCES

[1] Derek Lam, Steven L. Rohall, Chris Schmandt, and Mia K. Stern, 2002. Exploiting e-mail structure to improve summarization. In *ACM CSCW 2002*, Interactive Posters, New Orleans, LA.

[2] Apache Lucene. http://lucene.apache.org/java/docs/