It Pays to be Picky: An Evaluation of Thread Retrieval in Online Forums

Jonathan L. Elsas, Jaime G. Carbonell
Language Technologies Institute, Carnegie Mellon University
Pittsburgh, PA 15213
{jelsas, jgc}@cs.cmu.edu

ABSTRACT

Online forums host a rich information exchange, often with contributions from many subject matter experts. In this work we evaluate algorithms for *thread retrieval* in a large and active online forum community. We compare methods that utilize thread structure to a naïve method that treats a thread as a single document. We find that thread structure helps, and additionally *selective* methods of thread scoring, which only use evidence from a small number of messages in the thread, significantly and consistently outperform *inclusive* methods which use all the messages in the thread.

Categories and Subject Descriptors

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

General Terms

Algorithms, Measurement, Experimentation

Keywords

online forums, message boards, thread search

1. INTRODUCTION

Online forums contain a wealth of user generated content over a wide range of topics: from computer hardware to movies reviews and commentary to specific health issues. The contributors to online forums are often domain experts and these social information spaces host in some cases many millions of archived messages. Access to forum archives, however, is often rudimentary — these services provide basic browsing interfaces and simple keyword message-searching facilities. In this paper we explore the problem of information access in online forum data.

In online forums, messages are typically grouped into *message threads*, representing a conversation between a group of contributors. A message thread has a single *start message* and zero or more *response messages*. Message threads are frequently displayed chronologically in a "flat" structure where each message in the thread has at most one response.

There has been little previous research dealing specifically with search in online forums. Several studies have looked at knowledge extraction, identifying question-answer pairs [1] or responses that relate to a previous question in the thread [2]. Similar to a retrieval task, Feng et al. [4] developed a "discussion-bot", which responds to new forum posts with automatically identified related questions and answers. The question-matching component of this system retrieves likely answers with a vector-space TF-IDF ranking formula.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
Copyright 2009 ACM 978-1-60558-483-6/09/07 ...$5.00.

2. DATASET DESCRIPTION

The work presented here uses a recent crawl of an active, technically-oriented online forum, the MacRumors Forum (http://forums.macrumors.com), which hosts discussion relating to the computer manufacturer, Apple, Inc. The crawl was conducted in March of 2008, and contains over 3 million messages organized into almost 375,000 message threads.

We make the assumption that the typical useful unit of retrieval for online forum search is the message *thread*. Although this is certainly not always the case — a single message may fully answer an information need — the thread provides useful conversation context and discussion. When viewing the forum, a thread-view is typically most convenient. Additionally, 86% of the intra-forum links in this dataset refer to other threads, rather than posts or other possible units of retrieval.

Because of the difficulty in accessing the forum archives, users often post questions to messages boards that may have been answered in previous threads in the archive. When this happens, another user occasionally responds to that question with a link to a previous discussion possibly containing the answer. We can leverage this interaction between users to build an information retrieval test collection — the original question can be considered a query and the linked-to thread a relevant document.

To build a thread-retrieval test collection, we identified 48 instances of question/answer-link pairs in this collection. A thread contains a question-answer pair when (1) a response message provides a hyperlink to a previous thread in the forum, (2) the start message in this thread contains a question that is answered by the linked-to thread and (3) subsequent response messages in the thread do not indicate the linked-to thread is irrelevant. Of the identified answers, 78% contain only a single linked-to thread. We generated keyword queries from the question messages by manually extracting terms and phrases from the question text.

3. MODELS FOR THREAD SEARCH

We apply a variety of retrieval models to thread search in online forums, all based on the language modeling framework. We assume term independence and Dirichlet smoothing with the following estimation [6]:

$$P(Q|D) = \prod_{q \in Q} \left(\frac{n(q, D) + \mu P(q|C)}{|D| + \mu} \right)^{n(q,Q)} \quad (1)$$

where $q \in Q$ are the query terms, $n(q, D)$ is the term count of q in D, $|D|$ is the document length, and $P(q|C)$ is the collection language model. In all of the following models, query likelihoods are estimated with Equation 1. The models differ in their definition of the document D and how evidence from the messages is combined into a thread score. All of these models assume a uniform thread prior $P(T)$.

The first of these models are taken from previous work in blog feed search [3], a similar task to thread ranking in that the unit of retrieval is collections of documents. We refer to these as *inclusive* retrieval models which consider evidence from each of the messages when producing a thread score. The *large document* (LD) and *small document* (SD) models are given below.

$$P_{LD}(T|Q) \stackrel{\text{rank}}{=} P(Q|T)$$
$$P_{SD}(T|Q) \stackrel{\text{rank}}{=} \sum_{M \in T} P(Q|M)P(M|T)$$

where T is the thread, Q is the query, and $M \in T$ are the messages. In the large document model, the thread language model is estimated through a concatenation of the thread's constituent messages. This model represents how a general-purpose web search engine may retrieve message threads, indexing the entire thread content as a single document and ignoring the message-thread structure. In the small document model the thread language model is estimated through a weighted mixture of the messages' language models. This weighting is controlled via the message likelihood, $P(M|T)$. We consider two message likelihood formulations, both shown to be effective in blog feed search [3]: a uniform likelihood $P_U(M|T) \propto 1$ and a "centrality" likelihood, favoring messages that more closely resemble the overall thread $P_C(M|T) \propto \prod_{t \in M} P(t|T)^{P(t|M)}$. Both are normalized to form a probability distributions over messages in a thread.

The next set of retrieval models we consider *selective* in that they use evidence only from a few of the messages in the thread. These models are given by:

$$P_{PCS}(T|Q) \stackrel{\text{rank}}{=} \prod_{i=1}^{k} P(Q|M_i)^{1/k} \quad (2)$$

$$P_{MAX}(T|Q) \stackrel{\text{rank}}{=} \max_{M \in T} P(Q|M) \quad (3)$$

$$P_{START}(T|Q) \stackrel{\text{rank}}{=} P(Q|M^0) \quad (4)$$

where M^0 represents the thread's start message. The first model, *pseudo-cluster selection* (Equation 2, PCS), has also effectively been applied to blog search [5]. In this model, threads are scored by the geometric mean of the top-k message scores. We fix $k = 5$, which has been shown to perform well in a variety of other applications. The second model (Equation 3, MAX) also selects messages based the message rank with respect to the query. Note that this model is equivalent to the PCS model when $k = 1$. The third model (Equation 4, $START$) favors messages earlier in the thread. In this case the thread score is equal to the score of the start message.

4. EXPERIMENTS & ANALYSIS

Due to the limited relevance information available in this test collection, we must avoid treating all un-judged documents as non-relevant. For this reason, we focus on a recall-oriented evaluation, reporting recall at various cutoffs (R@N), as well as Mean Reciprocal Rank (MRR) which is well suited for retrieval tasks when only a single relevant document is known. In these results (Table 1), we can

Model	MRR	R@10	R@20	R@30	R@100
LD	0.0928	0.1553	0.2436	0.3113	0.5598
$SD + U$	0.0987	0.2283	0.3481	0.4379	0.6448
$SD + C$	0.0922	0.1867	0.2993	0.4170	0.6239
$START$	0.1491	0.3095	0.3920	0.4587	0.5984
MAX	0.1570	0.3253	**0.4675**	**0.5561**	0.6180
PCS	**0.1902**	**0.3308**	0.4031	0.4158	**0.6491**

Table 1: Performance results, thread ranking.

clearly see that the large-document model, which ignores the message-thread structure, does not perform as well as the models that leverage this structure by scoring the messages individually. Additionally, we see that the *selective* models consistently outperform the *inclusive* models. In all cases except R@100, the best performing selective model significantly outperforms the LD and $SD + C$ retrieval models using a 2-tailed paired t-test. These results demonstrate that

- Thread structure is important in thread ranking;
- Message selection in the thread is also important;
- Together these yield 60-100% improvement across performance measures over the naïve LD method.

5. CONCLUSION

In this paper explore the problem of thread search in online forums. We apply several state-of-the-art retrieval models from blog feed search to this task. We show that recognizing the message-structure helps and selective methods, which only look at a few of the messages in the thread, significantly and consistently outperform inclusive methods, which use evidence from all the messages in the thread.

6. REFERENCES

[1] G. Cong, L. Wang, C.-Y. Lin, Y.-I. Song, and Y. Sun. Finding question-answer pairs from online forums. In *Proc. 31st SIGIR*, pages 467–474, 2008.

[2] S. Ding, G. Cong, C. Lin, and X. Zhu. Using conditional random fields to extract contexts and answers of questions from online forums. In *Proc. of ACL-08: HLT*, 2008.

[3] J. L. Elsas, J. Arguello, J. Callan, and J. G. Carbonell. Retrieval and feedback models for blog feed search. In *Proc. 31st SIGIR*, pages 347–354, 2008.

[4] D. Feng, E. Shaw, J. Kim, and E. Hovy. An intelligent discussion-bot for answering student queries in threaded discussions. In *Proc. 11th IUI*, pages 171–177, 2006.

[5] J. Seo and W. B. Croft. Blog site search using resource selection. In *Proc. 17th CIKM*, pages 1053–1062, 2008.

[6] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM TOIS*, 22(2):179–214, 2004.