# Combining Fields in Known-Item Email Search

Craig Macdonald
Department of Computing Science,
University of Glasgow, Scotland, UK
craigm@dcs.gla.ac.uk

Iadh Ounis
Department of Computing Science,
University of Glasgow, Scotland, UK
ounis@dcs.gla.ac.uk

## ABSTRACT

Emails are examples of structured documents with various fields. These fields can be exploited to enhance the retrieval effectiveness of an Information Retrieval (IR) system that searches mailing list archives. In recent experiments of the TREC 2005 Enterprise track, various fields were applied to varying degrees of success by the participants. In this work, using a field-based weighting model, we investigate the retrieval performance attainable by each field, and examine when fields evidence should be combined or not.

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

**General Terms:** Performance, Experimentation

**Keywords:** Email, retrieval, fields, structure, metadata.

## 1. INTRODUCTION

In a known-item task (KI), there is only one relevant document that must be ranked as early as possible by the retrieval system. The evaluation measure in a KI task is Mean Reciprocal Rank (MRR), which rewards retrieval systems that rank the target document as early as possible. In TREC 2005, the Enterprise track (TREC-Ent) had a known-item task for email search, using an archive of mailing lists emails of the World Wide Web Consortium (W3C).

Emails are composed of two parts: the written message of the email, and various header fields such as subject and sender information. These fields may bring evidence of different importance, which can be taken into account to enhance retrieval performance. We use a field-based weighting model to combine the fields evidence of emails. A research problem is to determine which fields to apply and combine in retrieval. Our objectives are two-fold: Firstly, to determine how useful each separate field is for retrieval purposes. Secondly, to find indications of when the combination of two fields is effective.

## 2. FIELDS IN EMAIL SEARCH

In the W3C collection, there are six fields that we apply, namely Subject, Sender, mailing List name, message Text, and finally the Unquoted and Quoted parts of the message text.

As the W3C collection is in the form of a small Web crawl we additionally apply Body (which contains all the email fields), Title (which contains a mix of subject, sender and date), and Anchor Text of the incoming hyperlinks (which mostly contains the subject and sender of the email) as fields in our experiments. We denote the field that is the concatenation of the Body, Title and Anchor Text fields by All.

We index each field individually, removing stopwords and applying the first two steps of Porter's stemming algorithm. In Table 1, the second column shows the average length of each field over the 174,311 email documents of the W3C collection. For our experiments, we use the topics and the W3C collection from the TREC-Ent 2005 KI task.

To rank email documents, we use the Divergence from Randomness field-based weighting model PL2F. This model has shown a good retrieval performance on this task [1].

For the PL2F model, the relevance score of an email document $d$ to a query $Q$ is given by:

$$score(d,Q) = \sum_{t \in Q} \frac{qtf}{qtf_{max}} \cdot \frac{1}{tfn+1} \big( tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn) \big) \quad (1)$$

where $\lambda$ is given by $\lambda = F/N$. $F$ is the frequency of the query term in the collection and $N$ is the number of documents in the whole collection. $qtf$ is the query term frequency. $qtf_{max}$ is the maximum query term frequency among the query terms. $tfn$ is given by:

$$tfn = \sum_f \left( w_f \cdot tf_f \cdot \log_2(1 + c_f \cdot \frac{avg\_l_f}{l_f}) \right), (c_f > 0) \quad (2)$$

where $tf_f$ is the term frequency of term $t$ in field $f$, $avg\_l_f$ is the average length of the field and $l_f$ is the length of $f$ in $d$. $c_f$ is a hyper-parameter for each field controlling the term frequency normalisation, and the contribution of the field is controlled by the weight $w_f$. In our experiments, we set $c_f$ and $w_f$ using training.

## 3. SEPARATE FIELD PERFORMANCE

Firstly, we assess the performance of each field separately in ranking the emails. Table 1 shows the retrieval effectiveness when each field is used for retrieval separately. In the third column of the table, the system has been trained using 25 topics that are not in the test topics. In the fourth column, the parameters of the PL2F weighting model have been trained directly to the test topics. Training for the optimal setting allows the maximum potential of each field to be assessed.

We can see that the training topics are, in general, representative of the test topics, as the results are roughly similar


*SIGIR'06,* August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

| Field | $avg\_l_f$ | Train/Test | Test/Test |
|---|---|---|---|
| All | 394.35 | **0.593** | **0.608** |
| Body | 328.59 | 0.504 | 0.536 |
| Title | 16.04 | 0.504 | 0.508 |
| Atext | 49.72 | 0.439 | 0.461 |
| Sender | 5.90 | 0.029 | 0.031 |
| Subject | 4.09 | 0.468 | 0.468 |
| List | 2.78 | 0.018 | 0.025 |
| Text | 193.02 | 0.401 | 0.437 |
| Unquoted | 144.93 | 0.424 | 0.448 |
| Quoted | 48.08 | 0.026 | 0.036 |

**Table 1: Average Length of each field ($avg\_l_f$), and performance in MRR when used separately for retrieval. Train/Test denotes when the system is trained using the training topics, and Test/Test denotes when trained using the test topics. Note that the TREC 2005 best performing official run had MRR 0.621, while the median was 0.4545. All runs were statistically different from the best run in each column at $p < 0.05$.**

between both trainings. The All field, which contains the most evidence, performs signficantly better than all other fields (Signed Rank test, $p < 0.05$). Interestingly, there are fields that achieve an MRR of 0.4 to 0.5, namely Subject, Title and Anchor text (Atext), even though these do not contain the actual message text of the email. As each of these fields contains the subject of the email, we can infer that the subject is useful for retrieval, and alone can outperform the median run of the submitted runs on this task.

When considering the fields containing the message text of the email, i.e. All, Body, Text and Unquoted, we can see that the additional evidence present in the All and Body fields increases the performance over the Text field. However, the Quoted text field is of little retrieval value, and removing Quoted text from the Text, i.e. the Unquoted field, increases retrieval performance (from MRR 0.401 to 0.424 and 0.437 to 0.448). Finally, the Sender and List fields are not useful for retrieval for these topics, perhaps due to the lack of personal involvement of the topic creators in W3C.

# 4. COMBINING FIELDS

In this section, we investigate applying pairs of fields using the PL2F weighting model (Eq. (1)). The objective is to show when two fields should be combined. Table 2 presents the performance of pairs of fields. Note that some related pairs of fields are omitted, e.g. Text and Unquoted, because the Unquoted field is entirely contained in the Text field.

From the table, we can see that several combinations of fields achieve a good MRR, including some that outperform the best official run of TREC-Ent 2005 (run uogEDates12T: MRR 0.621). In general, a field containing the unquoted text of the email, and one containing the subject must be used to achieve a high MRR.

Moreover, it is possible to deduce when fields are similar or independent. For instance, although the Atext and Subject fields perform relatively well in Table 1, combining them (as in Table 2) does not improve on the retrieval effectiveness of either alone. This indeed suggests that they contain similar evidence, which matches what we know about these two fields (they both contain terms from the subject of the emails). The combination of Atext and Title exhibits similar properties. In contrast, applying fields that contain independent evidence, for instance Sender and Subject, or List and Subject amounts to an increased performance roughly equal to the sum of the individual performances of both fields. On the other hand, note that using two independent fields, such as Title and Unquoted has led to one of the best performances, even though the achieved MRR is not equal to the sum of the individual performances.

| Fields | | Train/Test | Test/Test |
|---|---|---|---|
| Atext | Body | 0.599 | 0.618 |
| Atext | List | 0.465 | 0.493 |
| Atext | Quoted | 0.417 | 0.456 |
| Atext | Sender | 0.450 | 0.475 |
| Atext | Subject | 0.453 | 0.460 |
| Atext | Text | 0.583 | 0.611 |
| Atext | Title | 0.481 | 0.501 |
| Atext | Unquoted | **0.623** | **0.637** |
| Body | List | 0.482 | 0.544 |
| Body | Sender | 0.436 | 0.551 |
| Body | Subject | 0.571 | 0.608 |
| Body | Title | 0.605 | 0.615 |
| List | Quoted | 0.033 | 0.045 |
| List | Sender | 0.040 | 0.058 |
| List | Subject | 0.483 | 0.486 |
| List | Text | 0.398 | 0.461 |
| List | Title | 0.499 | 0.506 |
| List | Unquoted | 0.358 | 0.466 |
| Quoted | Sender | 0.059 | 0.056 |
| Quoted | Subject | 0.381 | 0.394 |
| Quoted | Title | 0.441 | 0.471 |
| Quoted | Unquoted | 0.401 | 0.455 |
| Sender | Subject | 0.509 | 0.516 |
| Sender | Text | 0.413 | 0.435 |
| Sender | Title | 0.510 | 0.523 |
| Sender | Unquoted | 0.396 | 0.445 |
| Subject | Title | 0.507 | 0.514 |
| Subject | Unquoted | 0.565 | 0.596 |
| Text | Subject | 0.527 | 0.559 |
| Text | Title | 0.590 | **0.637** |
| Title | Unquoted | 0.621 | **0.637** |

**Table 2: MRR scores for combinations of pairs of fields. Runs not statistically different from the best run in column ($p < 0.05$) are denoted with underline.**

# 5. CONCLUSIONS

Our study investigates ten possible fields that could be applied by an email search system. We show that using more evidence from each email increases the retrieval performance of an email search system. In particular, it is essential that the chosen fields contain the subject and text of an email, though the quoted text of previous emails in the thread were not shown to be useful. Moreover, our results suggest that when different sources of evidence are combined, retrieval performance can be enhanced if the chosen sources provide independent evidence. In the future, we intend to work towards automatically assessing the usefulness of fields and their combinations.

# 6. REFERENCES

[1] C. Macdonald, V. Plachouras, B. He, and I. Ounis. University of Glasgow at TREC 2005: Experiments in Terabyte and Enterprise tracks with Terrier. In *Proceedings of TREC-2005*.