

Applying the Annotation View on Messages for Discussion Search

Ingo Frommholz

University of Duisburg-Essen

`ingo.frommholz@uni-due.de`

In this paper we present our runs for the TREC 2005 Enterprise Track discussion search task. Our approaches are based on the view of replies as annotations of the previous message. Quotations in particular play an important role, since they indicate the target of such an annotation. We examine the role of quotations as a context for the unquoted parts as well as the role of quotations as an indicator of which parts of a message were seen as important enough to react to. Results show that retrieval effectiveness w.r.t. the topicality of email messages can be improved by applying this annotation view on email messages.

1 Introduction

When replying to an email in a discussion list, users usually quote the part of the previous mail they refer to. An email body can thus be divided into two parts: the quotations and the new, unquoted part which contains the comments of the email author. The process of replying to emails is similar to the process of annotating documents: users usually select the document or passage they are referring their annotation to and write a comment to the chosen part. On the other hand, recent annotation prototypes like the one presented in [Frommholz et al., 2003] have used nested annotations as a building block for discussion. These examples make the relationship between emails in discussion lists and annotations clear. In this paper, we argue that email messages being replies to other messages can be regarded as annotations of (parts of) the previous email. As we will discuss, this view on emails as annotations motivates our approach for discussion search in the 2005 TREC Enterprise Track (TRECent). Our focus is on the role of quotations (the annotated parts) as a context for retrieving relevant emails. In many approaches for discussion search, quotations are often ignored, but we think they contain valuable information for the task at hand.

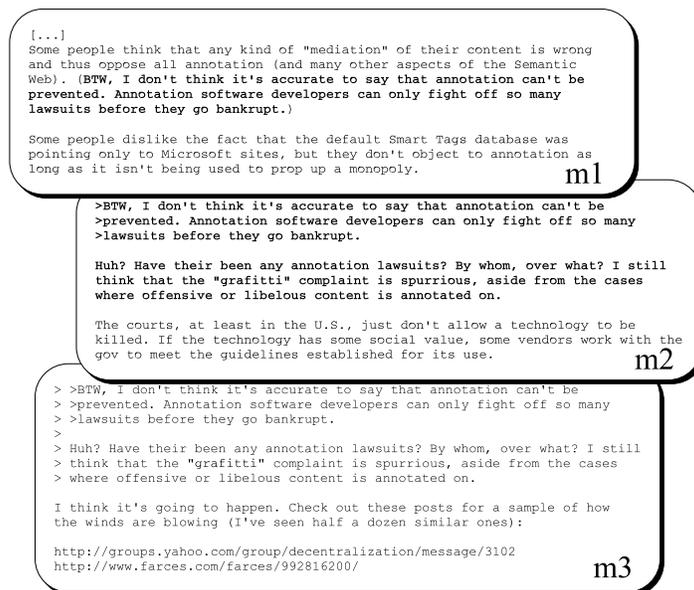


Figure 1: An example discussion thread

2 The Annotation View on Email Messages

Figure 1 shows some example messages m1, m2 and m3¹. m2 replies to m1, and m3 is a reply to m2. Replies, such as m2 and m3, usually contain two different parts in their body:

- the *quotations*, which are passages of the original text. Quotations are identified by quotation characters which prefix each line of the quotation². Such a quotation character can be '>' or ': ' or probably other characters or regular expressions; combinations of them usually define the *quotation depth*. In m3, quotations belonging to m1 have the depth 2 and are identified by two quotation characters ('> >'), whereas quotations belonging to m2 are identified by the single quotation character '>';
- the *new* part containing new content by the author of the message.

In replies, new parts contain annotations (here: textual, shared comments) of passages of previous messages. These passages are determined by the corresponding quotations. Quotations are thus the annotation targets. The distinction between new parts of messages and quotations, and their role as annotations and annotation targets, respectively, motivates our further retrieval approaches.

¹The examples were taken from the W3C email collection; no typos were corrected.

²Unfortunately, there are exceptions to this rule.

3 Discussion Search Approaches

We will now describe our approaches and experiments for discussion search. The goal of these approaches is to retrieve partially relevant messages of the test collection; a message is partially relevant if their new part is topically relevant; we make no efforts to identify whether these messages are fully relevant (i.e. contain pros and cons). All our retrieval functions are modelled with probabilistic predicate logic using probabilistic datalog [Fuhr, 2000]. For all our experiments we used the HySpirit retrieval framework³, which is an implementation of probabilistic datalog [Fuhr and Rölleke, 1998].

3.1 Baseline

In our baseline experiment we make no distinction between quotations and new parts, so the whole email message is regarded as one single document. We use the probabilistic binary predicate `fullmessageterm` to describe documents; the probability of this predicate reflects the term-frequency-based weight of document-term pairs. Example: `0.4 fullmessageterm("approach", "lists-117")` means the document "lists-117" is indexed with the stemmed term "approach" with a probability of 0.4. Additionally, we calculate weights for each term in the termspace based on the inverse document frequency and use this value as the probability for the unary predicate `termspace`. To determine the document frequency of terms, we do not distinguish whether the term appears in a new part or quotation of the message. Example: `0.81 termspace("pushbutton")` is the weight for the term "pushbutton" in the termspace. The `termspace` predicate is used to weight query terms with the datalog rule

```
wqterm(T) :- qterm(T) & termspace(T).
```

where the `qterm` predicate contains the query terms (after stemming and stopword elimination). Our baseline retrieval strategy is now implemented by the rule

```
retrieve(D) :- wqterm(T) & fullmessageterm(T,D).
```

We chose this as the baseline to study if the separation into quotations and new text and the application of context and highlight quotations (see below) does make sense w.r.t. retrieval effectiveness, or if a document should be regarded as a whole, without further distinction between quotations and new parts. The name of our baseline run is **du05baseline**.

Example 1. Let us assume we have the query "web deal" and the email "lists-1" contains, among others, these term, so we have the following knowledge base:

```
0.3 fullmessageterm("deal", "lists-1"). 0.2 fullmessageterm("web", "lists-1").
0.4 termspace("deal").                 0.1 termspace("web").
    qterm("deal"). qterm("web").
```

³<http://qmir.dcs.qmul.ac.uk/hyspirit.php>

HySpirit would now calculate $P(\text{wqterm}(\text{"deal"})) = P(\text{qterm}(\text{"deal"}) \wedge \text{termospace}(\text{"deal"})) = 1 \cdot 0.4 = 0.4$, $P(\text{wqterm}(\text{"web"})) = 0.1$ and (with the inclusion-exclusion formula; see [Fuhr, 2000] for further details how HySpirit calculates probabilities for probabilistic event expressions.)

$$\begin{aligned}
P(\text{retrieve}(\text{"lists-1"})) &= P(\text{wqterm}(\text{"deal"}) \wedge \text{fullmessageterm}(\text{"deal"}, \text{"lists-1"}) \\
&\quad \vee \text{wqterm}(\text{"web"}) \wedge \text{fullmessageterm}(\text{"web"}, \text{"lists-1"})) \\
&= P(\text{wqterm}(\text{"deal"})) \cdot P(\text{fullmessageterm}(\text{"deal"}, \text{"lists-1"})) \\
&\quad + P(\text{wqterm}(\text{"web"})) \cdot P(\text{fullmessageterm}(\text{"web"}, \text{"lists-1"})) \\
&\quad - (P(\text{wqterm}(\text{"deal"})) \cdot P(\text{fullmessageterm}(\text{"deal"}, \text{"lists-1"})) \\
&\quad \cdot P(\text{wqterm}(\text{"web"})) \cdot P(\text{fullmessageterm}(\text{"web"}, \text{"lists-1"}))) \\
&= 0.4 \cdot 0.3 + 0.1 \cdot 0.2 - 0.4 \cdot 0.3 \cdot 0.1 \cdot 0.2 = 0.1376
\end{aligned}$$

as the retrieval status value of “lists-1”.

3.2 Alternative Baseline

Email messages in TREC_{Cent} are considered (partially) relevant if their new parts were (partially) relevant. This leads us to an alternative baseline where quotations in email are not considered. Documents in this case thus consist of new parts of messages only. When neglecting the quotations in messages, we cannot use the `termospace` predicate above, but introduce another predicate `termospacequo`; to calculate its idf-based probability, terms are considered as appearing in a document only when they appear in the new part. We also introduce another predicate `term` whose probability is based on new parts only (in contrast to `fullmessageterm`). The retrieval function for the alternative baseline is therefore

```

wqterm(T) :- qterm(T) & termospacequo(T).
retrieve(D) :- wqterm(T) & term(T,D).

```

The name of the alternative baseline experiment, which we did not submit to TREC_{Cent} 2005, is `du05noquot`.

3.3 Context Quotations

As we discussed above, unquoted parts in replies can be regarded as annotations. One of the intrinsic features of annotations is that they depend on the object they refer to, which is the annotation target [Agosti et al., 2004]. In emails, the annotation target is the part of the previous message which is quoted by the email⁴. Such annotation targets contain contextual information which is needed to understand the associated comment. Considering the example in Fig. 1, taking only the user comment in m3 into account we have no indication that the topic of this comment is actually “annotation lawsuits”. This information can be derived from the annotation target. We call quotations *context quotations* when we use them as a context in the way described above. Our

⁴If the whole message is quoted or there is no quotation at all, we infer that the unquoted part of the email belongs to the whole previous message.

discussion shows that we might draw conclusions about the respective comment's topicality from the aboutness of context quotations. We will therefore regard context quotations as separate virtual documents which are not to be retrieved but support the retrieval function in making a relevance decision. From each email reply, two documents are created: the main document consisting of the new parts of the email and the attached virtual document consisting of the quotations (or the whole previous email if there is no quotation). To integrate context quotations into the retrieval function means that we look at the term weights of both the virtual and main document. New parts are represented by the `term` predicate introduced above, whereas the terms of quotations define the predicate `quoteterm`. Main and virtual documents are connected with another predicate `link` whose probability defines the strength of this connection. For example, `0.9 link("quotation")` models a strong link between new parts and quotations. The integration of the quotation terms and new terms is realised by the rule

```
about(T,D) :- term(T,D).
about(T,D) :- link("quotation") & quoteterm(T,D).
```

This means that a message is about a certain term if the new part or the quotation is about the term (depending on the strength of the connection between the quotation and unquoted part). The final retrieval function is

```
wqterm(T) :- qterm(T) & termspace(T).
retrieve(D) :- wqterm(T) & about(T,D).
```

We submitted two runs using context quotations: in **du05quotstrg** we used a strong link between quotations and new parts, whereas in **du05quotweak** we used a weak one (with `0.3 link("quotation")` in the second `about` rule).

Example 2. Let us assume we have the following datalog facts:

```
0.3 term("deal", "lists-1").      0.05 term("web", "lists-1").
0.4 quoteterm("web", "lists-1").
0.4 termspace("deal").           0.1 termspace("web").
0.9 link("quotation").
    qterm("deal"). qterm("web").
```

so we only have weak evidence that the new part is about “web”, but more evidence that the quotation is about this term. We get $P(\text{about}(\text{"deal"},\text{"lists-1"})) = 0.3$ and $P(\text{about}(\text{"web"},\text{"lists-1"})) = 0.05 + 0.9 \cdot 0.4 - 0.05 \cdot 0.9 \cdot 0.4 = 0.392$ and finally $P(\text{retrieve}(\text{"lists-1"})) = 0.4 \cdot 0.3 + 0.1 \cdot 0.392 - 0.4 \cdot 0.3 \cdot 0.1 \cdot 0.392 = 0.1445$.

Looking at the `wqterm` rule above, we later found that the use of the `termspace` predicate here is not the best choice, since the probability of each `termspace` predicate depends on the inverse document frequency based on whole email messages, whereas the probability of `term` depends on new parts only. This is kind of inconsistent, so we decided to perform alternative runs with

```
wqterm(T) :- qterm(T) & termspacenoquot(T).
```

The names of the alternative runs are **du05qaltstrg** and **du05qaltweak**, respectively, depending on the link strength of quotations to new parts.

3.4 Highlight Quotations

When a user annotates a (part of a) document, it is assumed that she found it interesting enough to react to it. This means the annotation target is implicitly highlighted and considered important by the annotation author. Examining the quotations and the quotation levels of emails, we can identify such highlighted parts of previous messages. A highlight quotation of a message m in another message m' is the part of m which is quoted by m' , where m' is a (direct or indirect) successor of m in the discussion thread. As shown in Fig. 1, the quotation of m2 is a highlight quotation of m1. Even m3 contains a highlight quotation of m1 (the part of the quotation belonging to m1). Our claim is that this part of m1 is even more important, since it appears in m2 as well as in its successor, m3. In the example, the highlight quotations found in m2 and m3 show that m1 is a relevant message when discussing the case of annotation lawsuits. It also shows that m1 might be a good entry point in the discussion thread for the given topic. It has to be noted, however, that if a part of a message is not highlighted (i.e. quoted by other messages) in the sense described above, this does not mean that this part is not important. Possibly everyone agrees with this part, but does not state this agreement explicitly, which is a common observation in email discussions. On the other hand, messages often contain parts which are not relevant for the further discussion or even not for the overall topic of a message, and it is our claim that these parts are mostly not subject to highlight quotations. Just as for context quotations, it is also useful to know what highlight quotations are about, since with this information we are able to conclude which topics of a message were interesting for others. Each highlight quotation is therefore as well regarded as a virtual document whose aboutness influences the relevance of the messages they are part of. So in contrast to context quotations, these virtual documents are connected to the unquoted part of previous messages, which are the main documents in this case. Using `retrieve` and `qlink` as above, we define the “about” rule as

```
about(T,D) :- term(T,D).
about(T,D) :- link("highlight") & highlightquoteterm(T,H,D).
```

Highlight quotation virtual documents are described with the predicate `highlightquoteterm`. Term frequencies are based on these documents. In contrast to the `quoteterm` predicate, this is a 3-ary predicate, since messages can have more than one highlight quotation and these are thus identified by the document ID and the ID of the document containing the highlight quotation. Arguments are therefore the term, the ID of the highlight document, and the document ID. Example: `0.62 highlightquoteterm("deal","lists-015","lists-117")` means that the term “deal” in “lists-117” is highlighted by “lists-015”, i.e. it appears in the quotation of the latter message. The term weight of “deal” in the highlight quotation is 0.62. We performed two experiments with different connection strengths of main and virtual documents using highlight quotations: **du05highstrg** (`0.9 link("highlight")`) and **du05highweak** (`0.3 link("highlight")`).

Example 3. In this example, we have the following knowledge base:

```

0.6 highlightquoteterm("deal", "lists-2", "lists-1").
0.5 highlightquoteterm("deal", "lists-3", "lists-1").
0.3 term("deal", "lists-1"). 0.05 term("web", "lists-1").
0.4 termspace("deal").      0.1 termspace("web").
0.3 link("highlight").
    qterm("deal"). qterm("web").

```

HySpirit translates the `highlightquoteterm` predicate in the tail of the second “about” rule into a probabilistic disjunction w.r.t. possible bindings of the variable H. When applying the second “about” rule above with T=“deal” and D=“lists-1”, HySpirit calculates⁵

$$\begin{aligned}
P(\text{link}(\text{"quotation"}) \wedge \text{highlightquoteterm}(\text{"deal"}, \text{"lists-2"}, \text{"lists-1"}) \vee \\
\text{link}(\text{"quotation"}) \wedge \text{highlightquoteterm}(\text{"deal"}, \text{"lists-3"}, \text{"lists-1"})) = \\
0.3 \cdot 0.6 + 0.3 \cdot 0.5 - 0.3 \cdot 0.6 \cdot 0.3 \cdot 0.5 = 0.303
\end{aligned}$$

and thus for `about("deal", "lists-1")` the probability $0.3 + 0.303 - 0.3 \cdot 0.303 = 0.5121$ and 0.05 for `about("web", "lists-1")`. As a final result, we get $P(\text{retrieve}(\text{"lists-1"}))$ is $0.4 \cdot 0.5121 + 0.1 \cdot 0.05 - 0.4 \cdot 0.5121 \cdot 0.1 \cdot 0.05 = 0.2088$.

As with context quotations, we also performed alternative runs with

```
wqterm(T) :- qterm(T) & termspacenoquot(T).
```

The names of these runs are **du05haltstrg** (with 0.9 `link("highlight")`) and **du05haltweak** (with 0.3 `link("highlight")`), respectively.

3.5 Combination of Context and Highlight Quotations

We also wanted to examine the retrieval effectiveness when both context and highlight quotations are involved. This means we had to create combined `about` rules as

```

about(T,D) :- term(T,D).
about(T,D) :- link("highlight") & highlightquoteterm(T,H,D).
about(T,D) :- link("quotation") & quoteterm(T,D).

```

for the following runs:

⁵We use extensional semantics here, which means `link("quotation")` is counted twice.

Name	Connection Strength
<code>du05highquotstrgstrg</code>	0.9 link("highlight") 0.9 link("quotation")
<code>du05highquotstrgweak</code>	0.9 link("highlight") 0.3 link("quotation")
<code>du05highquotweakstrg</code>	0.3 link("highlight") 0.9 link("quotation")
<code>du05highquotweakweak</code>	0.3 link("highlight") 0.3 link("quotation")

We again used

```
wqterm(T) :- qterm(T) & termspacenoquot(T).
```

to weight query terms.

4 Creating Documents from Emails

To create the virtual and main documents and the required probabilistic predicates introduced above, we had to parse emails and discussion threads in order to extract quotations and new parts and assign the quotations to their corresponding source. If a message contains several quotations and new parts (e.g., a quotation followed by a comment followed by another quotation followed by a comment), we merged the quotations and new parts, so we always gain two documents (one for the merged quotations and one for the merged new parts). In this way, we lose the information about which quotation belongs to which comment; nevertheless, the basic components we are looking at are the merged new parts of emails, and not passages, so we did not consider such a finer granularity in the link structure. The merged quotations of an email message constitute the context quotation virtual document. To extract highlight quotation virtual documents, we had to further examine the quotations of a message w.r.t. their quotation depth. Virtual documents made of quotations of depth 1 were assigned to the predecessor of the email, while virtual documents made of quotations of depth 2 were assigned to the predecessor of the predecessor, etc. In the example in Fig. 1, the quotation of m2 would be assigned to m1, but the quotation of m3 would be assigned to m1 (“BTW...go bankrupt.”) and to m2 (“Huh?...is annotated on.”). So m1 would have two virtual documents assigned from highlight quotations which, in this example, would have the same content, adhering to the fact that the same passage of m1 was quoted twice.

5 Indexing

Documents (email messages and also the virtual documents discussed above) are represented as bags of words, after performing stopword elimination and stemming. For the predicates `term`, `fullmessageterm`, `quoteterm` and `highlightquoteterm` we calculated their corresponding probabilities $P(t|d)$ as

$$P(t|d) := \frac{tf(t, d)}{avgtf(d) + tf(t, d)}$$

where $tf(t, d)$ is the term frequency of term t in document d and $avgtf(d)$ is the average term frequency of d , calculated as

$$avgtf(d) = \frac{\sum_{t \in d^T} tf(t, d)}{|d^T|}$$

with d^T being the document representation of d (i.e. the bag of words of d). Our calculation of $P(t|d)$ only depends on document-specific values which are available on-the-fly when the document is indexed (no collection-dependent values like the average document length are needed). For `termspace` and `termspacenoquot` we computed the $P(t)$ as their respective probability, which is interpreted as an intuitive measure of the probability of t being informative, based on the inverse document frequency of a term t

$$idf(t) = -\log\left(\frac{df(t)}{numdoc}\right)$$

with $df(t)$ as the number of documents (whole emails for `termspace` or only new parts for `termspacenoquot`) in which t appears and $numdoc$ as the number of documents in the collection. Based on the inverse document frequency, we calculated $P(t)$ as

$$P(t) := \frac{idf(t)}{maxidf}$$

with $maxidf$ being the maximum inverse document frequency.

6 Results

6.1 Official Runs

We participated in the discussion search task of TRECcent. The goal of our experiments was to see whether we gain better results w.r.t. retrieval effectiveness by applying our annotation view on emails instead of just considering whole emails (the baseline run) or new parts (the alternative baseline run). Our focus was on the topicality of emails (partial relevance in TRECcent discussion search). The test collection was the W3C email corpus provided by TREC. After submitting our TRECcent runs, we discovered a bug in the calculation of $idf(t)$ (we accidentally computed $numdoc$ values which were too high). We fixed the bug and repeated the experiments. The new, correct results are presented here.

All results were produced with `trec_eval -a -M1000 <qrels> <resultfile>`⁶. Figure 2 shows the interpolated precision values for the recall points 0.1, 0.2, . . . 1, averaged over all queries. Table 1 shows some other values returned by `trec_eval`. With respect to context quotations, the results indicate that quotations and new parts should be separated in an email. Nevertheless, the good results of `du05quotstrg` against `du05baseline` indicate that there is a strong connection between these parts, supporting the observation that in many cases the aboutness of the quotation context determines the aboutness of the new parts. Using highlight quotations can push relevant documents into the first ranks, but other relevant documents seem to be ranked worse. Effectiveness can benefit from regarding quoted parts as highlighted parts, but non-relevant documents with lots of replies might be over-emphasised against relevant documents which might not be subject to that many replies.

⁶`trec_eval` is available at http://trec.nist.gov/trec_eval/trec_eval.7.3.tar.gz

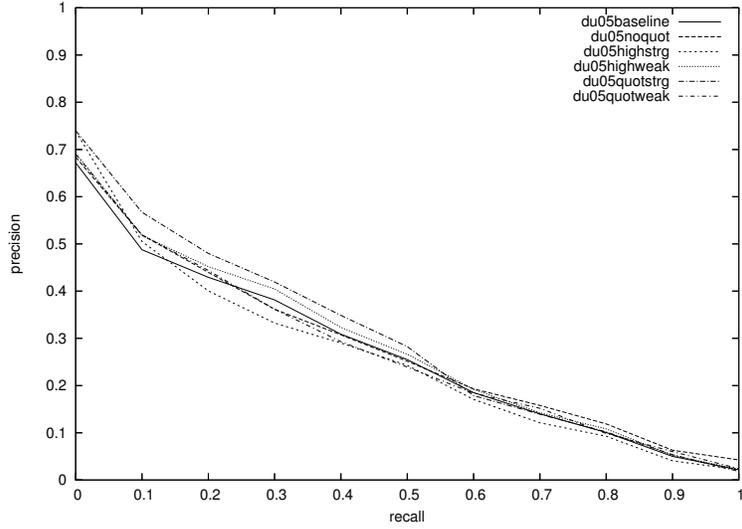


Figure 2: Interpolated recall-precision averages of official runs

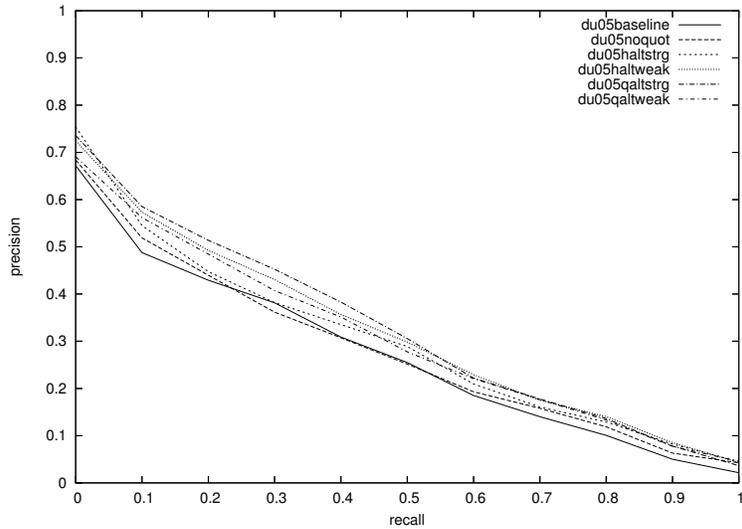


Figure 3: Interpolated recall-precision averages of alternative runs

6.2 Alternative Runs

We repeated the experiments with context and highlight quotations later with taking `termspacenoquot` instead of `termospace` for the `wqterm` rule (except in `du05baseline`). We gained

Run	baseline	noquot	quotweak	quotstrg	highweak	highstrg
Relevant Documents	3445					
Returned Relevant	2315	2338	2397	2463	2367	2319
Mean Avg. Precision	0.2565	0.2599	0.2565	0.2816	0.2691	0.2460
Prec. 5 docs retr.	0.4678	0.4441	0.4610	0.5085	0.4814	0.5186
Prec. 10 docs retr.	0.3966	0.4102	0.4186	0.4390	0.4237	0.4288
Prec. 15 docs retr.	0.3729	0.3955	0.3887	0.4	0.4	0.3706
Prec. 20 docs retr.	0.3458	0.3695	0.3653	0.3602	0.3619	0.3381
Prec. 30 docs retr.	0.3068	0.3220	0.3215	0.3198	0.3181	0.3017
Prec. 100 docs retr.	0.1844	0.1988	0.1937	0.1993	0.1912	0.1808
Prec. 200 docs retr.	0.1251	0.1303	0.1312	0.1338	0.1293	0.1267
Prec. 500 docs retr.	0.0673	0.0683	0.0704	0.0723	0.0694	0.0668
Prec. 1000 docs retr.	0.0392	0.0396	0.0406	0.0417	0.0401	0.0393

Table 1: Some results of the official runs

even better results against both baselines in these alternative experiments when incorporating context and highlight quotations. Figure 3 shows the interpolated recall-precision averages of these alternative runs (not submitted). Table 2 shows the mean average precisions (MAP) of the alternative runs.

Run	MAP
du05baseline	0.2565
du05noquot	0.2599
du05qaltweak	0.2889
du05qaltstrg	0.3094
du05haltweak	0.2726
du05haltstrg	0.2615

Table 2: Mean average precision of alternative runs

6.3 Combined Runs

Figure 4 shows the recall-precision averages of the combined runs. We can see that combining context and highlight quotations again slightly enhances retrieval effectiveness, which is no surprise considering our discussion above. Table 3 shows the mean average precisions for the combined runs.

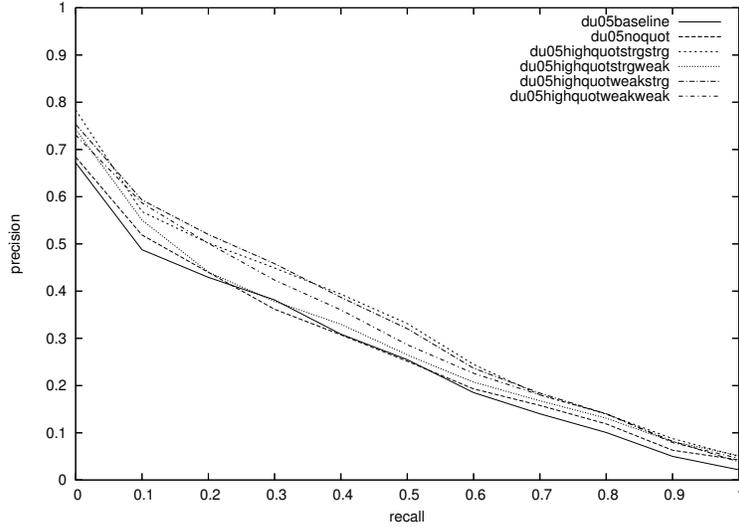


Figure 4: Interpolated recall-precision averages of combined runs

Run	MAP
du05highquotstrgstrg	0.3165
du05highquotstrgweak	0.2816
du05highquotweakweak	0.2992
du05highquotweakstrg	0.3174

Table 3: Mean average precision of combined runs

References

- [Agosti et al., 2004] Agosti, M., Ferro, N., Frommholz, I., and Thiel, U. (2004). Annotations in digital libraries and laboratories – facets, models and usage. In Heery, R. and Lyon, L., editors, *Research and Advanced Technology for Digital Libraries. Proc. European Conference on Digital Libraries (ECDL 2004)*, Lecture Notes in Computer Science, pages 244–255, Heidelberg et al. Springer.
- [Frommholz et al., 2003] Frommholz, I., Brocks, H., Thiel, U., Neuhold, E., Iannone, L., Semeraro, G., Berardi, M., and Ceci, M. (2003). Document-centered collaboration for scholars in the humanities - the COLLATE system. In Constantopoulos, P. and Sølvsberg, I. T., editors, *Research and Advanced Technology for Digital Libraries. Proc. European Conference on Digital Libraries (ECDL 2003)*, Lecture Notes in Computer Science, pages 434–445, Heidelberg et al. Springer.
- [Fuhr, 2000] Fuhr, N. (2000). Probabilistic Datalog: Implementing logical information retrieval for advanced applications. *Journal of the American Society for Information Science*, 51(2):95–110.

[Fuhr and Rölleke, 1998] Fuhr, N. and Rölleke, T. (1998). HySpirit – a probabilistic inference engine for hypermedia retrieval in large databases. In *Proceedings of the 6th International Conference on Extending Database Technology (EDBT)*, pages 24–38, Heidelberg et al. Springer.