Pitt at TREC 2005: HARD and Enterprise

Daqing He, Jae-wook Ahn
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA 15260
{dah44,jaa38}@pitt.edu

Abstract

The University of Pittsburgh team participated in two tracks for TREC 2005: the High Accuracy Retrieval from Documents (HARD) track and the Enterprise Retrieval track.

The goal of Pitt's HARD study in TREC 2005 was to examine the effectiveness of applying Self Organizing Maps (SOM) as a visual presentation tool and as a clustering tool in the context of HARD tasks, especially its role in clarification form generation. Our experiment results demonstrate that SOM can be used as a clustering tool to generate terms for query expansion based on interactive relevance feedback. It produced significant improvement over the baseline when measured by R-Prec. However, its effectiveness of being a visualization tool for users to make relevance feedback still needs careful examination and further studies.

Our goal in this year's enterprise search track was to study the effect of query expansion based on an expansion corpus in retrieving emails from an email corpus. The expansion corpus consisted of the WWW, People and ESW sub-collections of the W3C test collection. The results indicate that query expansion based on the expansion corpus can achieve significant improvement over the no expansion baselines. However, there is no significant difference to the simpler query expansion approach using blind relevance feedback. Interestingly the terms used in these two query expansion approaches are different, with averagely only 6 term overlap among 20 possible terms. Further study is needed for examining the effect of combining these two approaches.

1 HARD track

1.1 Overview

Searching for information is increasingly common in people's life. Modern techniques based on "free text" indexing and ranked retrieval have proven to be scalable and robust. Batch mode information retrieval (IR), which essentially studies retrieval algorithms, receives a great deal of attention. However, the initiative of searching for information, ultimately, lies to human. It is people who pose the questions, interpret what they read, and determine when their needs have been met. Especially in modern retrieval process, end users, who are not necessary search experts nor domain experts, leverage easy access to full text to support increasingly focused exploratory searches via iterative refinement [Marchionini1995]. Therefore, the ultimate goal of retrieval systems is not to generate the best possible ranked list for a given search query, but to provide the best information access mechanisms to users so that they can easily find needed information, and have a pleasant search experience.

High Accuracy Retrieval from Documents (HARD) is a track in TREC trying to address the problem of studying interaction between human users and retrieval systems, but at the same time keeping the TREC tradition of examining retrieval algorithms and achieving cross-site comparison [Allan2003].

The goal of Pitt's HARD study in TREC 2005 is to examine the effectiveness of using Self Organizing Maps (SOM) as a visual presentation tool and as a clustering tool in the context of HARD tasks, especially its role in clarification forms generation.

1.2 Self Organizing Maps

Self Organizing Map (SOM) is a technique invented by Professor Teuvo Kohonen which reduces the dimensions of data through the use of self-organizing neural networks [Kohonen2000]. It has been used for data visualization and natural language processing tasks [Lagus et al.1996, Lin1997]. The way that SOM reduces high data dimensions is by producing a map of usually two dimensions which plot the similarities of the data by grouping similar data items together. So SOM is a natural tool for visualizing and clustering data based on their similarities.

1.3 Training and Baselines

We adopted Indri 2.0 as our retrieval system. Indri is a state-of-art retrieval system developed by University of Massachusetts Amherst and Carnegie Mellon University (http://www.lemurproject.org/indri/). It has similar rich query context as Inquery system that was also developed by University of Massachusetts Amherst.

The training was performed on 50 HARD03 topics on AQUAINT collection, which contains 320380 documents. The search queries were extracted from the title and the description part of the HARD topics for the run HDTRAN-PLAIN, and plus the terms from the named entities in the narrative part of the topic statement automatically identified by BBN's Identifinder [1] for the run HDTRAN2-PLAIN. Our training indicated that HDTRAN-PLAIN achieved better Mean Average Precision (MAP) and R-Precision (R-Prec) than HDTRAN2-PLAIN. This is why almost all our further studies were based on the former way of obtaining topic information (see Table 1).

Runs	selected docs	selected terms	weight	MAP	R-Prec
HDTRAN-PLAIN				0.3647	0.3818
HDTRAN2-PLAIN				0.3324	0.3437
HDTRAN-BRF2020@0.3	20	20	0.3	0.3709	0.3919
HDTRAN-BRF2020@0.5	**20**	**20**	**0.5**	**0.4168**	**0.4238**
HDTRAN-BRF2020@0.8	20	20	0.8	0.4127	0.4190
HDTRAN-BRF1520@0.5	15	20	0.5	0.4154	0.4200
HDTRAN-BRF2520@0.5	25	20	0.5	0.4149	0.4235
HDTRAN-BRF2015@0.5	20	15	0.5	0.4102	0.4191

Table 1: The training on 50 HARD topics. The parameters of the run in bold font was selected as the parameters for the evaluation baseline HDEVAL for generating clarification forms.

We also examined the blind relevance feedback (BRF) mechanism in Indir system. As shown in Table 1, BRF with the right parameters can achieve greatly improvement on MAP and R-Prec over the runs without (say HDTRAN-PLAIN). Therefore, we used the BRF setting of HDTRAN-BRF2020@0.5 for generating the baseline runs HDEVAL and HDEVAL2 for our initial HARD submission [2], and used HDEVAL for generating our clarification forms.

[1] http://www.bbn.com/speech/docs/datasheets/idnt-022103.pdf

[2] The difference between HDEVAL and HDEVAL2 is the same as that between HDTRAN-PLAIN and HDTRAN2-PLAIN.

1.4 Clarification Forms

Unlike the HARD topics in past two years, there is no metadata associated with this year's HARD topics [He and Demner-Fushman2003, He et al.2004]. However, the requirements to the clarification were relaxed to allow more complex form structures. Our focus on clarification forms was at the usage of SOM. We explored two different but commonly usages of SOMs.

1.4.1 CF1: Graphic Based Approach

Our first approach to the clarification forms was using SOM to generate a graphic representation of the returned documents. The maps were designed in such way that all similar returned documents would be grouped into the same or adjacent cells on the map. Therefore, effectively, we visualize the "clusters" of returned documents [3].

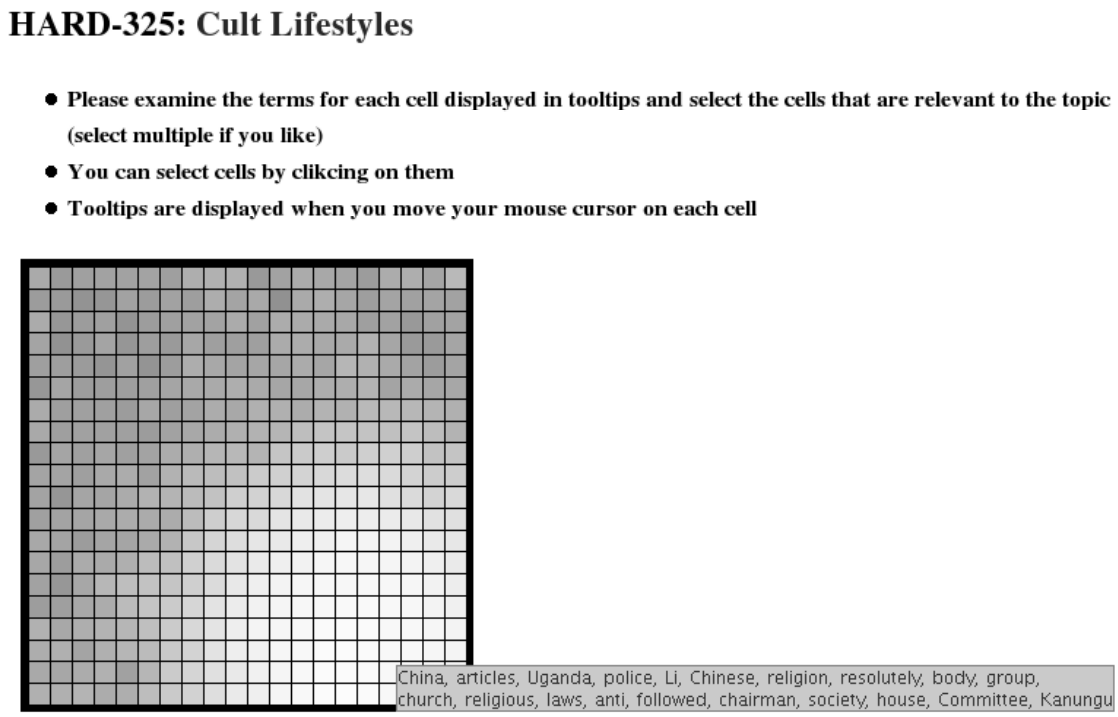

Figure 1: A Graphic Based Clarification Form. The gray scale indicates the similarity of a cell to its surrounding cells. The text is the top 20 terms associated with a cell.

To construct the graphic based clarification forms, for each topic, we selected top 400 returned documents from the HDEVAL. Based on the content of these documents, we extracted a 1000 word vector as the representation of each documents. The weights of the terms were calculated using BM25. The software used for generating SOMs was SOMPAK version 3.1 developed by the SOM Programming Team of the Helsinki University of Technology Laboratory of Computer and Information Science [Kohonen et al.1996].

The SOM training was divided into two phases. Firstly, there was the ordering phase during which the vectors for the map cells are generated and ordered. In this phase, the initial neighborhood radius was set at 10 and decreased to one during the training. The initial learning rate was 0.05 and it also decreased to zero.

[3] The map does not actually show the clusters of documents. We will come back to this point in later discussion.

Secondly, there was the fine tune of the values of the vectors. In this phase, the initial neighborhood radius was set at 3 and the initial learning rate was 0.02. All 400 documents were fed sequentially for training for each of the phases. The topology of the maps generated here was rectangular and the neighborhood function was Gaussian.

The SOMs generated in the above steps were displayed with a Java applet to visualize the map and the terms for each cell. We used the gray scales of the cells to represent its similarity with its surrounding eight cells. The similarity was calculated based on the cosine value of the vectors of the cells. The method was that the similar a cell is with its surrounding cells, the brighter its gray scale will be. Therefore, if a cell was identical to its surrounding cells, it was filled with white color, and if it is totally different to its surrounding cells, its color is black. Anything in between will have a gray scale corresponding to it. Users were able to move the cursor on each cell to view the top 20 most representative terms associated with the cell. If he liked the terms, he could select the cell, and then move to review another cell. More than one cell could be selected for this task. Figure 1 shows a SOM map for topic HARD-325.

1.4.2 CF2: Text Based Approach

Our second approach to the clarification forms was using a more traditional text based presentation. However, different to previous methods, we used SOM as a clustering tool to cluster documents into clustering based on their content similarities. The document representation and the SOM construction were identical to the methods used for graphical base CFs.

From the maps, we extracted two sets of terms and their associated weights in the SOM. The first one is from the cell of the map that is most similar the query. We obtained this information by calculating the cosine value between the vector of each cell and that of the query. Top 20 terms and their associated weights in the vector of the cell were extracted to be put into the clarification forms.

Although SOM can be used to group similar documents together, it does not represent the concept of clusters in the map. It only has cells, in which contain similar documents. Cells that are adjacent to each other are more similar than cells that are apart, but it is not guaranteed that two adjacent cells would belong to the same cluster. For our text based clarification forms, we have to present terms from different clusters of documents, so we need to construct the clusters of documents based on the cells of documents in the map. Our method was that each cell was treated as a unit, and they are grouped into clusters of cells by using single link clustering method. The similarity between cells was calculated based on cosine, and the threshold was set at 0.85. Based on the clustering results, we then selected top 20 terms and their weights from the top 6 biggest clusters of cells. The reason to choose clusters based on their size was because we assume that the cells containing some relevant documents would be similar to each other, whereas cells contain irrelevant documents would contain rather random content. Therefore, the clusters with many cells have higher chance to contain relevant documents than those with few cells. Figure 2 shows an example of the text based clarification forms.

1.5 Incorporating Clarification Forms

Both graphic based and text based clarification forms would return a set of terms based on the selection of annotators. The major technique we used for incorporating the results from clarification forms was term based query expansion.

In CF1 – the graphic base clarification forms, we not only knew the terms selected, but also their weights in the SOMs, so we explored the usage of not only those terms but also their weights in the query expansion. For terms selected in CF2 – the text based clarification forms, we did not obtained weights from SOMs for each individual expanded terms. To differentiate the original query from the expanded terms, we allocated 0.75 relative weight to the former, and 0.25 to the latter across all the topics and runs from all clarification forms.

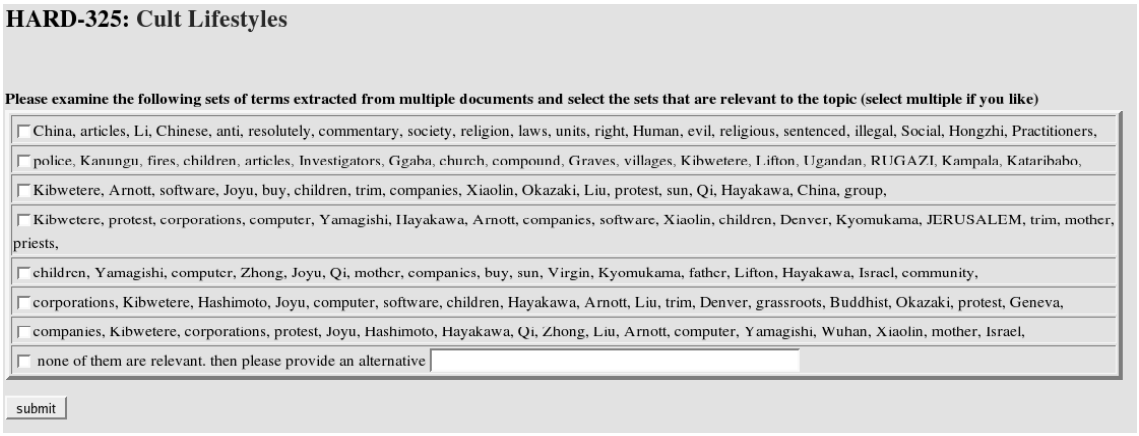

Figure 2: A Text Based Clarification Form

We also explored evidence combination method for combining the results from multiple runs. We used the "CombSUM" method for combining several rank lists into one [Fox and Shaw1994]. This method reranks the documents based on the sum of their normalized scores in each rank list. This method assumes equal weights for all rank lists. We combined the rank lists from four different runs. The first one is the baseline run HDEVAL used for generating clarification forms. We also included another baseline run HDEVAL2, whose difference to HDEVAL was that the queries of the former also included the terms related to the named entities marked up in the topic narrative parts in the topic statements. The rest two are the two runs (HDEVAL-CF1WW and HDEVAL-CF2) based on the clarification forms CF1 and CF2, respectively.

1.6 Results and Discussion

The final evaluation results gave us a different view of the two baseline runs. This time, HDEVAL2 achieved better results than HDEVAL when looking at both MAP and R-Prec. However, their difference is not statistical different (for example, p = 0.782 in the Wilcoxon Matched-Pairs Signed-Ranks Test when comparing to the average precision of the two runs)[4].

Runs	CF	QE terms with weights	CFQE+BRF	MAP	R-Prec
HDEVAL	n/a	n/a	n/a	0.2577	0.2903
HDEVAL2	n/a	n/a	n/a	0.2637	0.2981
HDEVAL-CF1WW	CF1	yes	no	**0.2202**	0.2743
HDEVAL-CF2NOB	CF2	no	no	0.2765	**0.3296**
HDEVAL-CF2B225	CF2	no	yes, 20,20,0.5	0.2908	**0.326**
HDEVAL-COMB	n/a	n/a	n/a	**0.2771**	0.3242

Table 2: The HARD evaluation results

Using CF1, the graphic based clarification forms, actually hurt the performance. Comparing to the retrieval effectiveness of the run HDEVAL, that of HDEVAL-CF1WW decreased about 15% relatively when measured by MAP, and 5.5% relatively when measured by R-Prec. The decrease in MAP is statistical significant (p = 0.0084). However, this bad result should not rule out the usefulness of the graphic based clarification form using SOMs. Some feedbacks from the NIST annotators about

[4]the Wilcoxon Matched-Pairs Signed-Ranks Test is the statistical significant test used throughout this report. Therefore, we will not mention it every time.

the graphic based clarification forms were mostly concentrated on unclear instructions, and too many similar cells, which makes the selection of the cells really hard. We need further studies to identify the exact failure of the graphic based clarification forms.

Our CF2, the text based clarification forms generated positive result without problems. When examining via R-Prec, the preferred measure in HARD, HDEVAL-CF2NOB achieved 14% relative increase over the baseline HDEVAL. When measured by MAP, the increase is smaller, but it is still 7% relative increase. The increase measured by R-Prec is statistical significant with the p value as 0.01. The increase measured by MAP is not statistical significant (p = 0.227).

In general, as shown in the training, BRF is a valid query expansion technique that would bring in positive results. One interesting question to ask here is whether or not BRF still make sense after the query expansion using terms from clarification forms. This was the purpose of the run HDEVAL-CF2B225. Again taking the advantage of the BRF mechanism implemented in Indri search engine, we set the BRF parameters as extracting 20 terms from top 20 documents and use 0.5 as the relative weight allocation between the query and new BRF terms (that is the two have the same weight). We obtained somewhat mixed signals. When looking at MAP, BRF after query expansion based on clarification forms achieved another 5% relative increase, but it generated 1% relative decrease when measured by R-Prec. None of these difference are statistical different. However, comparing to the baseline HDEVAL, HDEVAL-CF2B225 still achieved 12% relative increase in R-Prec, and the increase is statistical significant with the p value as 0.03. Even though HDEVAL-CF2B225 achieved 13% relative increase in MAP against HDEVAL, the difference is not statistical significant (p = 0.08).

The evidence combination run did not perform well. It indeed achieved better results than the two baseline runs HDEVAL and HDEVAL2, which are the two runs used in the combination. In addition, the improvement over HDEVAL is statistical significant not only measured by R-Rrec (p = 0.001), but also measured by MAP (p = 0.029). However, it does not performed as well as the best of the four runs HDEVAL-CF2B225, although the difference is not significant (p=0.32).

2 The Enterprise Track

2.1 Overview

The goal of the Enterprise track is to study the issues related to searching documents of an enterprise (organization). The tasks defined for this year's experiments are expert search and email search. The latter has two subtasks: search for known emails in the collection (called "known item search") and search for emails discussing certain topics (called "discussion topic search"). The collection used for this year's experiment is W3C test collection, which was based on the crawls at W3C.com websites in June 2004. The W3C test collection consists of several sub-collections, of which "email" sub-collection is the target corpus that the returned emails should be drawn from. The email sub-collection contains 198,394 emails. The other three W3C sub-collections used in our studies were "www", "esw", and "people" corpora, which contains 45975, 19605, and 1016 documents respectively.

Term mismatch is a major problem in information retrieval [Crestani2002]. This indicates that due to complexity and flexibility of language use, the terms existing in the documents and that in the query could be different even though the two terms might express the same concept. This problem becomes worse when the documents are short. This is because short documents do not provide enough space for the various expressions of a concept to be presented in the documents.

Searching for emails usually faces even worse term mismatch problem. This is because the average emails size tends to be smaller than average articles. For example, the average size of emails in W3C collection is 9.8kb, much shorter than the average size of the Web pages in W3C collection, which is 23.8kb.

As discussed in [Singhal and Pereira1999], a common approach to overcome the term mismatch

problem is through expansion. Only by adding more representative terms from relevant or potential relevant documents, can retrieval system get better chance to observe all possible terms associated with the expression of a concept. One key factor that affects the effectiveness of using expansion techniques is that the expansion corpus should be topically related to the query and the target collection. Such requirement may not be easily satisfied in some other retrieval tasks, however we hope that it will be the case for our enterprise retrieval task.

The target collection is the W3C email sub-collection (referred as email corpus in this report), and all the search topics were developed on this email corpus. The expansion corpus we used was the combination of www, esw, and people sub-collections (referred as the expansion corpus). We hope that the fact that both corpora belong to the same organization – W3C.com– would means that the expansion corpus can be viewed as the comparable collection to the email corpus which would effectively reduce the term mismatch problem in our email retrievals.

Therefore, our goal in this year's enterprise search track is to study the effect of query expansion using the expansion corpus in retrieving emails for our email corpus. The other research question is about the comparison between query expansion with the expansion corpus and that with blind relevance feedback on top returned emails. We studied both research questions in the two subtasks: known item search and discussion topic search.

Same as in our participation to HARD track, the retrieval system we used was Indri 2.0 developed by University of Massachusetts Amherst and Carnegie Mellon University. We made two modifications to the Indri system to make it be able to index emails documents and Web documents, and to make the Indri system to print out the terms and their weights used by Indri system for its query expansion based on blink relevance feedback.

During indexing, we found that the email corpus contained some empty or duplicated emails. After cleaning, there are total 173146 documents in the email corpus. Our cleaning also included reformatting the email documents to remove unnecessary information. The result format is shown in Figure 3. All fields in the emails were indexed.

The expansion corpus in total contains 66596 documents, with average document length at 19.3kb. We used an in-house developed script to strip off html tags, and reformat the documents into a standard trec format. All documents in the expansion corpus were kept, and their new format is shown in Figure 4.

2.2 Known Item Search

The known item search (KIS) task assume that the searcher knows that an email in the collection contains the information that he/she wants, and the task for the retrieval system is to locate and return the email at the top of the rank list.

The search topics for KIS is different to normal TREC topics in that it contains only the title part, which is usually a few keywords (see Figure 5).

Our query expansion algorithm for KIS is straightforward. It involves the five steps presented in Figure 6.

We explored various parameters for the query expansion algorithm using the 25 available training topics. Our training study demonstrated that the retrievals obtained the best results with the query expansion from the top 20 terms selected from top 20 documents. Our training study also demonstrated that the relative weight allocation between the original query and the expanded terms did affect the performance, and it seems that the Indri system achieved the top performance (measured by Mean Reciprocal Precision (MRP)) when the allocation is about 0.7 to 0.8 for the original query, and 0.3 to 0.2 for the expanded terms (see Table 3). This is reasonable consider KIS is a high precision search, so expanded terms, which are not confirmed in their relevance to the search, should not be weighted too much.

```
<DOC>
<DOCNO>lists-000-0012197</DOCNO>
<RECEIVED>Fri Feb 13 03:31:25 2004</RECEIVED>
<ISORECEIVED>20040213083125</ISORECEIVED>
<SENT>Fri, 13 Feb 2004 17:31:24 +0900</SENT>
<ISOSENT>20040213083124</ISOSENT>
<NAME>Daigo Matsubara</NAME>
<EMAIL>daigo@w3.org</EMAIL>
<SUBJECT>new list</SUBJECT>
<ID>nm8isib2xer.wl@natto.w3.mag.keio.ac.jp</ID>
<TO>copras-public@w3.org</TO>
<TEXT>
charset="US-ASCII"
expires="-1"
List_Name: copras-public
Requester_Email: rigo
ListPurpose: This list is the public mailing-list for the copras
EU-Project. This project is set up to coach IST-Projects under
IST-Framework 6 into the standardization process. This list will be
used to interface publicly with IST-Projects
Maintaining_Activity: Copras EU-Project
--
Daigo Matsubara / W3C Systems Team / mailto:daigo@w3.org
</TEXT>
</DOC>
```

Figure 3: An example of reformatted document in the email corpus

Our submission runs on the final 125 evaluation topics showed the similar picture. As shown in Table 2.3, using the expansion corpus do achieve significant improvement over no expansion baseline run using the two relative weight allocation obtained in training. The Wilcoxon Matched-Pairs Signed-Ranks Test was used for statistical significant test, and the p value for the difference between the "KITRAN-BASE" and "KITRAN-WWW-QE@0.7", and "KITRAN-WWW-QE@0.8" are both at 0.00058.

However, when the blind relevance feedback based on the top retrieve emails from the email corpus have an appropriate relative weight allocation (say 0.99 for the original query vs 0.5 for the original query), it can produce comparable improvement over the non-expansion base run (KIEVAL-BASE) as the WWW query expansion did. As shown in Table

We then examined the effect of query expansion by WWW expansion corpus and that by BRF approach, and found that these two approaches not only achieved similar effect on overall retrieval

Runs	weight for original query	weight for the expanded terms	MRP
KITRAN-BASE			0.3622
KITRAN-WWW-QE@0.5	0.5	0.5	0.3400
KITRAN-WWW-QE@0.6	0.6	0.4	0.3987
KITRAN-WWW-QE@0.7	0.7	0.3	0.4193
KITRAN-WWW-QE@0.7	0.8	0.2	0.4017

Table 3: The effect of different relative weight between the original query and the expanded terms measured by Mean Reciprocal Precision of the Searches on 25 known item training topics.

```
<DOC>
<DOCNO>www-000-10696137</DOCNO>
<TITLE>Context I</TITLE>
<CREATOR>Don Box</CREATOR>
<TITLE>box.ppt</TITLE>
<TEXT>
http://www.w3.org/2000/03/xp65435/box.ppt
what is soap
don box
sun/netscape bof
january 25, 2000

soap philosophy

First invent no new technology

SOAP simply codifies existing practice of using HTTP+XML as an
   application protocol

... ... ...
<</TEXT>
</DOC>
```

Figure 4: An example of reformatted document in the expansion corpus

```
<top>
<num> Number: KI33 </num>
<title> WSD WG f2f ws reference examples </title>
</top>
```

Figure 5: A example of known item topic

effectiveness, but also expressed similar impact to individual topics. This is reflected in the following observations:

- Among 125 search topics, there were 49 topics have MRP changes greater or equal than 0.02 between the run with query expansion and the run without. 32 of them have the difference between the change of the KIEVAL-WWW-QE@0.8 run and that of the KIEVAL-BRF QE@0.99 run smaller than or equal to 0.01.
- Among the 76 topics that had MRP changes smaller than 0.02, only 32 topics were due to the base run has better MRP score than the average 0.2537. All other 44 topics were eight has 0 or lower MRP scores than 0.2537.

However, the terms used for query expansion in KIEVAL-BRF-QE@0.99 and KIEVAL-WWW-QE@0.8 runs only share in average 6 common terms among 20 possible terms across 125 topics. This is consistent with Harman's study showing that different query expansion mechanisms select different terms. We have not done studies on the effect of combining the two set of terms.

2.3 Discussion Topic Search

The scenario for discussion topic search is that the user is searching for emails discussing the pros or cons of an argument. Although it would be nice to be able to identify whether the emails are for

1. step 1: select a query Q1 from the batch query file;
2. step 2: using Indri search engine to search the expansion corpus for query Q1;
3. step 3: based on predefined parameters, the Indri system will return top N terms selected from top M documents in the research result of step 2;
4. step 4: based on predefined weight allocation, the expanded query Q2 is a combination of original query Q1 and the expanded terms from step 3.
5. step 5: go to step 1.

Figure 6: The query expansion algorithm

Runs	weight for original query	weight for the expanded terms	MRP
KIEVAL-BASE			0.2577
KIEVAL-WWW-QE@0.7	0.7	0.3	0.3325
KIEVAL-WWW-QE@0.8	0.8	0.2	0.3353
KIEVAL-BRF-QE@0.5	0.5	0.5	0.2442
KIEVAL-BRF-QE@0.99	0.99	0.01	0.3413

Table 4: The Mean Reciprocal Precision of Searches on 125 Known Item topics.

or against an argument, it is not the task of discussion topic search to figure that out, at least not in this year's task. The task is to return emails that contain such discussions. There were 50 topics ("issues"), generated by all the participants in the discussion topic search, and the target collection is the W3C email corpus. A correct relevant email is defined as an email which contributes a pro or con relating to the topic, in new (not quoted) text.

We adopted the similar approach in the discussion topic search as that in the known item search. We incorporated the expansion corpus used in known item search for the query expansion in the discussion topic search, and used the similar relative weight allocation that was used in known item search.

We also explored the usage of email threading in retrieving discussion topic emails. We treated two emails sharing the same thread if one is the antecedent of the other in the thread chain, or the two share the same antecedent in the thread chain. We used the threading information to adjust the ranking of documents in two ways. The first one is aggressive approach, where if a document in the rank list shares the same thread with a document in the top N position in the rank list, it will be moved to the top position (the run based on this approach is DTEVAL-BIG) . The other approach is more conservative. The documents that share the same thread with a document in the top N position of the rank list will get a predefined booster. The run DTEVAL-SML1 used top 10 document with the booster 0.8, whereas the run DTEVAL-SML2 looked at top 50 documents with the booster factor 0.8.

Our analysis shows that the aggressive is a bad strategy, whereas the conservative strategies did achieved improvement over the run that did not using the thread information. The improvement is statistical significant when measured by MAP (p=0.040) for map, and by Precision at 10 (p=0.044).

3 Conclusion

In this paper, we have presented the studies we conducted in the participation to TREC 2005 in the tracks of HARD and enterprise. The goal of Pitt's HARD study in TREC 2005 is to examine the

Runs	MAP	R-Prec	P@10
DTEVAL	0.1846	0.1906	0.2831
DTEVAL-BIG	0.028	0.0932	0.0169
DTEVAL-SML1	**0.2184**	0.2241	**0.3271**
DTEVAL-SML2	0.1949	0.1911	0.3203

Table 5: The results of discussion topic searches on the evaluation topics.

effectiveness of using Self Organizing Maps (SOM) as a visual presentation tool and a clustering tool in the context of HARD tasks, especially its role in clarification forms generation. Our experiment results demonstrate that SOM can be used as a clustering tool to generate terms for user's selection and thus for query expansion. It produced significant improvement over the baseline when measured by R-Prec. However, it needs careful design and the user should be given well prepared instructions when using SOM as a visualization tool for users to make relevance feedback. Our further work on HARD studies is at analyzing the failure of the graphic based clarification forms, and exploring a more elaborated approach for evidence combination that we studies in our CLEF 2005 experiment [He and Ahn2005].

Our goal in this year's enterprise search track is to study the effect of query expansion using the expansion corpus in retrieving emails for our email corpus. The expansion corpus consisted of the WWW, People and ESW sub-collections of the W3C test collection. The results indicate that query expansion based on the expansion corpus can achieve significant improvement over the baselines. However, there is no significant difference to the simpler query expansion approach based on blind relevance feedback. Interestingly the terms used in these two query expansion approaches are different, with averagely only 6 term overlap among 20 possible terms. Our Further study in Enterprise track is to finish the analysis of discussion topic search and the examination of the effect of combining the two approaches for query expansion.

Acknowledgements

The authors would like to thank James Allen for organizing HARD track, Ian Soboroff, Nick Craswell, and Arjen P. de Vries for coordinating Enterprise track.

References

[Allan2003] James Allan. 2003. Hard track overview in trec 2003 high accuracy retrieval from documents. In *The Twelfth Text Retrieval Conference*. to appear.

[Crestani2002] Fabio Crestani, 2002. *Using semantic and phonetic term similarity for spoken document retrieval and spoken query processing*, pages 363–375. Physica-Verlag GmbH, Heidelberg, Germany, Germany.

[Fox and Shaw1994] E.A. Fox and J.A. Shaw. 1994. Combination of multiple searches. In *Proeddings of the 2nd Text REtrieval Conference (TREC-2)*, pages 243–252.

[He and Ahn2005] Daqing He and Jae-wook Ahn. 2005. Pitt at CLEF05: Data Fusion for Spoken Document Retrieval. In *Proceedings of Cross-Language Evaluation Forum 2005*.

[He and Demner-Fushman2003] Daqing He and Dina Demner-Fushman. 2003. HARD Experiment at Maryland: From Need Negotiation to Automated HARD Process. In *Proceedings of Text REtrival Conference (TREC) 2003*.

[He et al.2004] Daqing He, Dina Demner-Fushman, Douglas W. Oard, Damianos Karakos, and Sanjeev Khudanpur. 2004. Improving passage retrieval using interactive elicition and statistical modeling. In *Proceedings of TREC 2004*.

[Kohonen et al.1996] Teuvo Kohonen, Jussi Hynninen, Jari Kangas, and Jorma Laaksonen. 1996. SOMPAK: The Self-Organizing Map Program Package. Technical Report Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.

[Kohonen2000] Teuvo Kohonen. 2000. *Self-Organizing Maps*. Springer, 3rd edition.

[Lagus et al.1996] K. Lagus, T. Honkela, S. Kaski, and T. Kohonen. 1996. Self-organizing maps of document collections: A new approach to interactive exploration. In E. Simoudis, J. Han, and U Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 238–243, Menlo Park, California. AAAI Press.

[Lin1997] Xia Lin. 1997. Map displays for information retrieval. *Journal of the American Society for Information Science*, 48:40–54.

[Marchionini1995] Gary Marchionini. 1995. *Inforamtion seeking in electronic environments*. Cambridge Series on Human-Computer Interaction. Cambridge University Press.

[Singhal and Pereira1999] Amit Singhal and Fernando C. N. Pereira. 1999. Document expansion for speech retrieval. In *Research and Development in Information Retrieval*, pages 34–41.