Measuring Pseudo Relevance Feedback & CLIR

Paul Clough
University of Sheffield
Western Bank
Sheffield, UK
+44 114 222 2664
p.d.clough@sheffield.ac.uk

Mark Sanderson
University of Sheffield
Western Bank
Sheffield, UK
+44 114 222 2648
m.sanderson@sheffield.ac.uk

ABSTRACT
In this poster, we report on the effects of pseudo relevance feedback (PRF) for a cross language image retrieval task using a test collection. Typically PRF has been shown to improve retrieval performance in previous CLIR experiments based on average precision at a fixed rank. However our experiments have shown that queries in which no relevant documents are returned also increases. Because query reformulation for cross language is likely to be harder than with monolingual searching, a great deal of user dissatisfaction would be associated with this scenario. We propose that an additional effectiveness measure based on failed queries may better reflect user satisfaction than average precision alone.

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

General Terms
Algorithms, Performance.

Keywords
Pseudo Relevance Feedback, CLIR, Evaluation.

1. INTRODUCTION
Previous cross language (CL) and monolingual IR research has shown that on average across topics, PRF can help improve retrieval performance [1][6][7][8]. In PRF, the top *n* documents are assumed relevant and used in iterative retrieval cycles, e.g. for query expansion. In CLIR, PRF can be used prior or post translation (or both) for pre/post-translation query expansion (see, [1][6]). This strategy works well with many relevant documents retrieved in the initial top *n*, but is less successful when the initial retrieval effectiveness is poor, which is commonly the case in CLIR where initial retrieval performance is affected by translation accuracy (see, e.g. [4]).

Retrieval effectiveness is commonly measured using either average precision across a series of recall values or at a fixed rank. Using these measures, PRF appears beneficial in most CLIR experiments, as using PRF seems to consistently produce higher average precision than baseline systems. This implies users would prefer them, but the technique is rarely deployed in actual IR systems.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007

As part of an experiment looking at the effects of relevance feedback on cross language image retrieval [3] we found that using average precision, PRF appeared to increase retrieval performance. However, when considering the number of queries that return no relevant images, PRF actually makes retrieval performance worse. Such a contradiction suggests that, in this case at least, average precision is not reflecting user preferences. Dunlop discusses user-centred evaluation measures for information retrieval in [5] arguing that factors other than just system performance must be taken into account during evaluation.

2. METHODOLOGY
Our experimental setup and methodology is described in more detail in [4], but can be summarised as follows. Using textual captions associated with a photographic collection for image retrieval, the top *n* captions were selected for query expansion using the `Lemur` language model for IR [8]. The ImageCLEF[1] ad hoc test collection was used for evaluation [2] comprising 28,133 historic photographs and fifty user queries representative of typical CL image requests. Queries in German, French, Italian, Dutch, Spanish and Chinese were translated into English (the target language) using the `Systran` machine translation system (see [3] for an analysis of translation errors using this resource) and default `Lemur` feedback parameters were used during the evaluation ($\alpha = \lambda = 0.5$).

Retrieval effectiveness was measured across all fifty topics using the following: mean average precision (MAP), precision at 100 (P_{100}), normalised precision at 100 ($P_{norm100}$), the number of perfect topics and the number of bad topics. P_{100} measured the proportion of the top 100 retrieved which were relevant[2] and $P_{norm100}$ the proportion of relevant documents found in the top 100. The number of topics in which *all* relevant images were found in the top 100 were called *perfect*, and topics with *no* relevant in the top 100 were referred to as *bad* topics. These last two measures consider what we feel is important to searchers: (1) that returning at least one relevant document will satisfy a user's search, and (2) that returning no relevant documents will cause user dissatisfaction. We contend that minimising bad topics better reflects, than average precision, a user's view of retrieval effectiveness.

[1] ImageCLEF 2004 – see http://ir.shef.ac.uk/imageclef2004/

[2] We presume users are willing to search at least one hundred images to locate relevant images, which from our experience in user studies of this form of retrieval [4] is justified.

3. RESULTS

The first table summarises initial retrieval effectiveness without feedback at rank position one hundred. Results vary dramatically across the languages reflecting the quality of query translation. However, which is worse depends on the evaluation measure used. For example, based on MAP, German would appear better than Spanish. However, the number of bad topics indicates that 18% of German topics have no relevant in the top one hundred compared with 8% for Spanish; the latter we believe more satisfying for the user. Similarly, German has much higher MAP than Chinese, but has one more bad topic than Chinese: again we would contend that a Chinese searcher would be more satisfied with their system compared to a German searcher.

	% mono MAP	Avg P_{100}	Avg $P_{norm100}$	#perfect topics	#bad topics
Mono	0.5514	0.18	0.81	22	1
German	73.3%	0.13	0.65	19	9
French	75.5%	0.16	0.69	18	4
Italian	72.9%	0.14	0.66	14	7
Dutch	69.0%	0.11	0.58	15	9
Spanish	71.5%	0.15	0.65	16	4
Chinese	50.7%	0.12	0.54	13	8
	Average	0.14	0.66	16.7	6.0

Although fewer bad topics compared to German, $P_{norm100}$ for Chinese indicates on average that 54% of relevant images were found in the top 100, compared to 65% for German. This highlights the importance of evaluation based on more than one measure to obtain a more accurate picture of retrieval effectiveness.

	% diff MAP	% diff P_{100}	% diff $P_{norm100}$	#perfect topics	#bad topics
Mono	1.9%	0.3%	0.6%	23	1
German	3.8%	2.4%	-0.8%	20	9
French	3.4%	-1.0%	-0.4%	20	4
Italian	4.2%	3.1%	2.8%	16	7
Dutch	2.4%	6.3%	3.1%	15	11
Spanish	0.9%	6.0%	4.5%	15	3
Chinese	6.8%	19.9%	0.7%	14	9
Average	3.3%	5.3%	1.5%	17.6	6.3

The second table shows the results after PRF selecting thirty terms for query expansion from the top ten documents and one feedback iteration. In contrast to MAP reflecting purely rank position changes of all relevant documents, the $P_{norm100}$ score provides a different reflection of effectiveness change as this score is not affected by re-ranking, i.e. a higher $P_{norm100}$ score only results from more relevant being found in the top one hundred after the feedback cycle. The results show the evaluation measures are somewhat contradictory, particularly when the P_{100} and $P_{norm100}$ scores increase after feedback, but the number of bad topics also increases (e.g. for Dutch and Chinese) and only once decreases. By examining this range of measures, we find that pseudo relevance feedback polarises topics: apparently improving ones that were already retrieving relevant documents, but harming a few previously effective topics so much that now no relevant are retrieved.

4. CONCLUSIONS

In this poster, it was shown that effectiveness based on a more user-centered evaluation measure: the number of perfect and bad topics reflected an alternative view on the "quality" of a retrieval system across topics. This was shown for different forms of cross language retrieval system and for pseudo relevance feedback. From the results, we contend that analysing performance with average precision only is not necessarily the best summary of evaluation effectiveness, particularly with respect to the user. We believe that the simple count of topics in which all relevant were found, and in particular topics in which no relevant were found within a top ranked set (i.e. the top one hundred) are a useful additional measure of retrieval effectiveness. In this task, we found that pseudo relevance feedback worsened cross language image retrieval effectiveness overall based on the number of bad topics, counter to previous results based on mean average precision.

5. FUTURE WORK

We plan to investigate more closely the relationship between perfect and bad topics, average precision and user satisfaction to confirm our belief that bad topics are a key factor in a user's view of the utility of a retrieval system. In addition, we believe it would be beneficial to re-asses previous results for methods such as stemming, stop word removal, alternative weighting schemes, etc. to determine whether results based on average precision are a suitable indicator of effectiveness, or whether the simpler measure of perfect/bad topics is more reflective of user preferences.

6. REFERENCES

[1] Ballesteros, L., Croft, B. Resolving Ambiguity for Cross-Language Retrieval. In *Proceedings of SIGIR 1998*. 64-71.

[2] Clough, P.D., Sanderson, M. The CLEF Cross Language Image Retrieval Track, In *Proceedings of CLEF 2003*.

[3] Clough, P.D., Sanderson, M. Assessing Translation Quality for Cross Language Image Retrieval. In *Proceedings of CLEF 2003*.

[4] Clough, P.D., Sanderson, M. The Effects of Relevance Feedback in Cross-Language Image Retrieval. In *Proceedings of ECIR 2004*. 238-252.

[5] Dunlop, M. Reflections on MIRA: Interactive Evaluation in Information Retrieval. In *Journal of the American Society for Information Science*. Vol. 51(14). (2000). 1126-1274.

[6] McNamee, P., Mayfield, J. Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources. In *Proceedings of SIGIR 2002*. 159-166.

[7] Mitra, M., Singhal, A., Buckley, C. Improving Automatic Query Expansion. In *Proceedings of SIGIR 1998*. 206-214.

[8] Zhai, C., Lafferty, J. Model-Based Feedback in the KL-Divergence Retrieval Model. In *Proceedings of CIKM 2001*. 403-410.