What else is there? Search Diversity Examined

Mark Sanderson, Jiayu Tang, Thomas Arni, Paul Clough

Dept. of Information Studies, University of Sheffield, Regent Court, Sheffield, S1 4DP, UK
{m.sanderson, j.tang, t.arni, p.d.clough}@shef.ac.uk

Abstract. This paper describes a study on diversity in image search results. One of the first test collections explicitly built to study diversity – the ImageCLEFPhoto 2008 collection – was used in an evaluation exercise in the summer of 2008. Analyzing 200 of the runs submitted by 24 research groups enabled the relationship between precision and result diversity to be studied. In addition, the level of diversity present in search results produced by retrieval systems built without explicit support for diversity was computed. The remaining potential to improve on diversity was calculated and finally, a significant preference by users for diverse search results was shown.

1 Introduction

A relatively overlooked topic of information retrieval research is diversity in search results. Despite the seemingly wide adoption of this technique in certain parts of the commercial web search sector, relatively little research has been published and there are almost no test collections available to evaluate methods. As an initial step to overcome the lack of a suitable benchmark for studying this topic, the ImageCLEFPhoto 2008 test collection was created to foster research in promoting diversity in search results. In addition to being a task in the CLEF 2008 evaluation campaign, the collection and run data was additionally analyzed to enable a broad study of search diversity to be conducted, which is the subject of this paper. We start with a brief survey of research in diversity in Section 2, followed by a description of the design and construction of the collection in Section 3. Next details of the run data submitted to the ImageCLEFPhoto 2008 task of CLEF are outlined, followed by the series of experiments conducted on the data in Sections 4 and 5. Finally conclusions and directions for future work are described in Section 6.

2 Literature Review

The underlying principle of most retrieval systems is to rank documents in the order of their similarity to the query. However such an approach fails to consider how similar relevant documents should be retrieved; neither does it consider the potential for queries with different interpretations, where documents relevant to distinct interpretations might need to be retrieved at the same time. Spärck Jones et al. [1]

discussed the need to consider such retrieval and more recently, Sanderson [2] demonstrated the extent of queries in search logs that have multiple interpretations.

Some research on devising search algorithms that promote diversity has been conducted: Maximal Marginal Relevance (MMR) [3]; Maximal Diverse Relevance from Zhai [4] and follow on work from Chen and Karger [5]. A common theme to the work was almost a complete lack of a test collection with queries that required diverse search with relevance judgments that describe links between documents. The TREC interactive tracks in TREC 6-8 created relevance judgments with topic clusters, however only 20 topics were created [6, 7]. More recently Clarke et al [8] adapted a question-answering collection to be used as a retrieval collection with topic clusters in its relevance judgments. To the best of our knowledge, these collections represent the totality of resources available to the research community. We describe the adaptation of an existing test collection to support measurement of diversity.

3 Building a Test Collection to Test Diversity

ImageCLEFPhoto [9] is a sub-task of ImageCLEF, which itself is part of CLEF (the Cross Language Evaluation Forum). In 2008, it was decided to make diversity the main research focus of the task. ImageCLEFPhoto used the IAPR TC-12 collection for the past three years [10, 11, 12], and it was extended to allow diversity measurement, by grouping the relevance judgments of existing topics into clusters that reflect relationships between relevant images in the collection.

A subset of topics was identified from the collection's existing 60 that were judged to be in need of diverse results. For the majority of the topics, the clustering was clear. For example, if a topic asked for "images of beaches in Brazil", clusters were formed based on location; if a topic asked for "photos of animals", clusters were formed based on animal type. Out of the 60 existing topics it was judged that 39 were appropriate to use in the 2008 evaluation. The topics were classified into two classes: Geographical (22) and Miscellaneous (17). On average there are 8 clusters per topic with an average of 62 relevant images per topic. Detail on the processes of topic selection and cluster assessment can be found in [12].

Participating groups in ImageCLEFPhoto were asked to return runs, containing a ranked list of images for each of the 39 topics. Of the 24 groups who submitted runs, one submitted a substantial number; consequently, a post hoc ten run limit per group was imposed. This resulted in a set of 200 runs (not all groups submitted ten). Details of the methodologies used by the participating groups can be found elsewhere [13].

Evaluation was based on precision measured at a fixed rank of 20, P20, and a diversity measure based on a statistic proposed by Zhai et al. [13]: *Cluster Recall* (CR) [14] measured at rank *K* defined as follows:

$$CR(K) = \frac{clusters(K)}{tc}$$

where *clusters* is the number of different clusters in the document ranking up to and including the document at rank *K*, and *tc* is the total number of clusters in a particular

topic. When a single evaluation measure was needed, the classic harmonic mean *f* was used. It is defined as follows:

$$f = 2\frac{P \cdot CR}{P + CR}$$

4 Data Analysis

A series of analyses of the run data set were conducted, which are now described. The first analysis was simply to scatter plot the runs based on the P20 and CR20 of each submitted run. As can be seen in Figure 1, the two measures were well correlated (correlation coefficient was 0.8). This initial result might suggest that diversity is not worth considering. However, the next experiment showed this not to be the case.

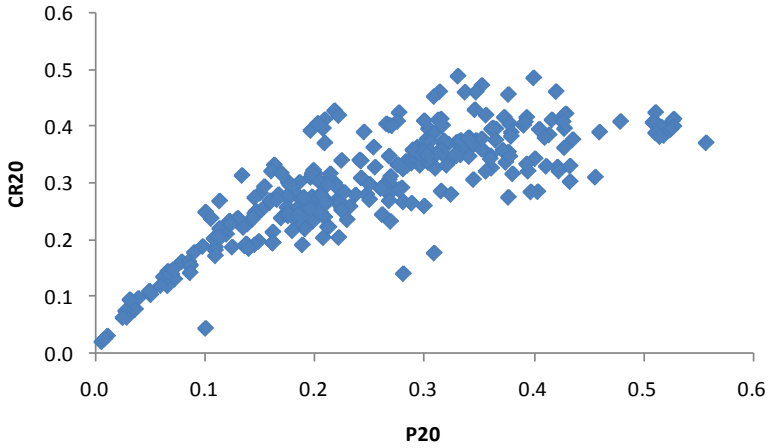

Fig. 1. Precision at rank 20 (P20) versus cluster recall at rank 20 (CR20)

Because the documents, topics and images marked as relevant in the ImageCLEFPhoto 2008 test collection were as the same as those used in 2006 and 2007, it was possible for a comparison to be made between the 2008 runs and those of the previous years. In both past years, diversity was not evaluated; so participating groups had little motivation in building a retrieval system that supported it. However because topic cluster judgments were added, it was possible to calculate cluster recall on each 2006 and 2007 run and compare to the 2008 runs so as to understand the difference between retrieval systems that supported diversity and those in the previous years that did not.

Measuring average CR20 in 2006, 2007 and 2008 revealed that in 2008 it was substantially higher: 0.30 compared to 0.20 in 2007, 0.21 in 2006. However, since the same relevance judgments were used across the three years, some form of learning effect might have impacted on the results. There was evidence for this as there was an overall increase in the precision of runs submitted in 2008 compared to 2006 and

2007. Given, that the previous analysis showed that high precision is likely to lead to high cluster recall, simply comparing cluster recall values across the two years was not sufficient. Therefore, a different data analysis was devised.

Comparisons between the 3 years were restricted to pairs of runs from different years that had the same value of precision. Such pairs were identified and the CR on each was compared. For 2007, 37 pairs were found (24% of all 2007 runs) and for 2006, 35 pairs (25%). Across the 2007-2008 comparisons, in 34 pairs (92%), the cluster recall in 2008 was higher than in 2007. For the 2006-2008 comparisons 27 pairs (77%) of the 2008 cluster recall scores were higher. Sign and t-tests revealed that both comparisons were significant at the 0.05 level. Even after controlling for the increase in precision across the years, cluster recall increased substantially and significantly in 2008.

4.1 The Relationship of Cluster Recall to Precision

Participants in the 2008 task reported that applying diversification strategies was often at the expense of reduced precision [15, 16] and an examination of the graph in Figure 1 provide some support for this concern, where it can be seen that runs with the very highest precision did not have the highest cluster recall.

Therefore, we studied the ten best runs from each of the 24 participating groups as determined by *f*. We conducted a pair-wise analysis, this time on a per topic basis. For each topic, pairs of runs that had an equal P20 were identified and those that had a maximal difference in CR20 were selected. Thus the two runs represented extremes of diversification, but at equal precision. For each of the 39 topics, up to 20 unique pairs at 18 different precision values were identified. A run for a particular topic was used only once in any pairing. The 18 P20 values were from 0.1 to 1.0 increased in 0.05 steps. A total of 2,339 unique pairs were found. The arithmetic mean over the cluster recall difference of the runs in one pair is 0.24 with a standard deviation of 0.14.

Table 1. Overview of arithmetic means of all pairs at various precision levels

Precision at rank	Cluster Recall (CR20)	Mean Precision	Std. Dev
5	Lower	0.49	0.33
	Higher	0.51	0.33
10	Lower	0.43	0.26
	Higher	0.44	0.26
15	Lower	0.40	0.23
	Higher	0.41	0.23

We studied Precision at ranks 5, 10, and 15 for each of the pairs. Perhaps surprisingly, Table 1 showed that the more diverse runs (Higher CR20) had a slightly higher precision than the less diversified result runs (Lower CR20). A paired two-tailed t-test revealed that all differences were significant: P5, $p<0.0005$; P10, $p=0.014$; P15, $p=0.042$. The average difference in CR between the 2,339 pairs was 0.33. The P5 differences for the subset of pairs with a CR difference below the mean were examined and a small significant difference was found in precision measured at rank

5, $p<0.0005$. The results show that for runs with the same P20 value but with a high cluster recall, the relevant documents occurred at a higher average rank than runs with a low cluster recall. This is a result that runs counter to the concerns that promoting diversity in ranks lowers precision.

4.2 Estimating the Potential to Improve Diversity

Although it was clear from the 2006/7-2008 analysis that support for diversity in retrieval systems was improved significantly in 2008, it was judged interesting to know what potential there was for improving cluster recall further. Therefore, two analyses were applied to the run data from 2007 and 2008. For all the submitted runs for each topic, each relevant document was replaced by another relevant document taken from the qrels file. The precision at rank 20 for each run was thus unaffected after the replacement process, but the cluster recall was altered. Two types of relevant document replacement were tried:

1. Randomly replace each relevant document in a run with a different relevant document taken from the qrels file. The aim behind this was to establish a random baseline for diversity across all runs regardless of their precision value.
2. All relevant documents were replaced with other relevant documents from the qrels file such that the maximum possible CR20 for that run would result. Here some idea of the upper bound on cluster recall for the runs would be calculated.

The results of these analyses are shown in Table 2. As expected, the magnitude in improvement for CR20 for both replacement strategies was largest for the 2007 runs. It was notable that the random replacement strategy outperformed the original diversified runs in 2007 and in 2008 by 23.5% and 14.8% respectively. From these results, it would appear that there is still considerable potential for improving methods to promote diversity.

Table 2. Changes in CR20 values using two different randomisation strategies

Year	Original	Random	Improvement	Max. CR	Improvement
2007	0.23	0.29	23.5%	0.36	56.2%
2008	0.31	0.35	14.8%	0.45	47.0%

5 User Experiment

The key motivation for promoting diversity in ImageCLEFPhoto 2008 was the belief that a diverse yet relevant result set is more likely to satisfy a user's need. However, as far as we know, little research has been published showing user preference for diverse search results. Carbonell & Goldstein [3] and Song et al. [17] conducted user experiments, but with 5 users involved in each. Thus we used the 2008 run data to conduct such a similar but larger experiment.

We randomly selected a set of 25 topics (proportionally to the number of topics in each category and sub-category) and for each topic we searched in the 2008 runs for pairs where each topic had an equal P20 and a maximal difference in CR20. Because the earlier experiments showed that precision paired runs with different cluster recall had consistent differences in the ranking of the relevant documents, it was decided to remove all non-relevant images from the runs. The search results of each paired run were displayed side-by-side, with the retrieved images and captions arranged in a simple grid. To remove order effects, the runs with the low and high CR20 were randomly chosen to be on the left or right side of the side-by-side display.

A total of 31 persons were involved in the user experiment, most of whom were students from the University of Sheffield. For each topic, participants were instructed to select the result set that in their opinion was the better result for the given query. If participants had no preference for one of the two given results they could chose that both result sets are "equally good" or they could choose that "neither" were any good. The intention of the survey was not revealed to the participants. The more diversified result set was randomly presented on the left or right result set. The users were not informed that non-relevant images were removed.

5.1 Results

Across the 25 topics and all users, 54.6% preferred the more diversified results set and 19.7% preferred the less diversified set; 17.4% thought the both result sets were equal and 8.3% chose "neither". This agrees with the results from previous studies, e.g. in the study by Carbonell and Goldstein [3], 80% of users preferred diverse results over those generated from a "standard" ranking. An examination of individual user behavior reflected the globally calculated preferences. Only one user preferred the less diversified set more often than the more diversified. In all other cases users consistently favored the more diversified results more often than the less diversified across the 25 topics. However, not in all those cases was the more diversified option the most selected. Three persons judged the majority of the result set to be equally good and one person chose the "neither" option the most. Nevertheless, a paired two-tailed t-test showed there was a significant difference measured between the participant preferences for the result set with a high cluster recall over the one with a low cluster recall at rank 20 ($p<0.0005$).

On a topic-by-topic basis the results were very similar to the user behavior results. In only four topics did the majority of users favour the less diversified result set. In the remaining 21 topics the more diversified result set were always the most chosen option. As with the user overview, there was a significant difference between the more diversified versus the less diversified results ($p<0.05$). This study provided strong evidence that users do care about diversity. From a user's perspective, the more diversified result set was significantly preferred over the less diversified set.

In a post experiment questionnaire, participants were asked how often they used an image search engine: 9.7% used image search engines on a daily basis; 41.9% weekly; 29.0% monthly; 16.1% very seldom and 3.2% never use images search engines. In an open question on important properties of image search engines, most stated (in some form of words) that relevance was critical, additionally, users

explicitly responded that they like a variety of relevant images retrieved, avoidance of duplicate images or a wide range of relevant images.

6 Conclusions and Future Work

This paper reported on an analysis of the 2008 ImageCLEFPhoto task and its principle focus on studying diversity in search results. The challenge for participants was to maximize both the number of relevant images, as well as the number of relevant image clusters represented within the top 20 results. The new task attracted a large number of submissions, which allowed a number of experiments to be conducted on the submitted run data set. The key results of the experiments were:

1. A comparison between the 2008 and 2006/7 versions of the task showed that retrieval systems not explicitly built to maximize diversity (as typified in the 2006/7 tasks) had significantly lower CR than the systems that explicitly supported diversity. In other words "standard" retrieval systems do not by default support diversity well.
2. Although there was a concern that building a retrieval system that maximizes diversity was likely to impact on precision, there was little evidence to support this. However, experiments to establish upper bound and random baselines for diversity showed that there is much potential to improve diversity in the future.
3. Finally, a user experiment showing pairs of search results at different levels of cluster recall produced significant evidence showing that users preferred the search results that were more diverse.

The results shown here strongly suggest that support for diversity is an important and currently largely overlooked aspect of information retrieval. However, these results were derived from one image test collection and it is reasonable to question how applicable the results will be in other IR domains such as full text document search, or when studied in different searching tasks. Because of the lack of testing resources, there are limited examples to draw conclusions from. However, the small user study of Carbonell and Goldstein [3] showing preference for diversity was conducted on a "TREC-like" full text document collection indicating that preferences for diversity is not just a feature of image search. Clearly further work is required to confirm such tentative views.

The future work planned is first to better understand the needs of users with respect to diversity and then to create a substantially larger collection for ImageCLEFPhoto 2009 to enable a broader range of diversity experiments to be conducted.

Acknowledgments

Thanks to Anand Ramamoorthy and to the valuable comments from the anonymous reviewers. Work was partly supported by the following EU-funded projects: TrebleCLEF (Grant agreement #215231) and MultiMatch (Contract #IST-033104).

References

1. Spärck-Jones, K., Robertson, S. E. and Sanderson, M. (2007) Ambiguous requests: implications for retrieval tests, systems and theories, ACM SIGIR Forum, v.41 n.2, p.8-17.
2. Sanderson, M. (2008) Ambiguous queries: test collections need more sense. In Proc. ACM SIGIR 499-506.
3. Carbonell, J. and Goldstein, J. (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proc. ACM SIGIR, 335-336.
4. Zhai, C. (2002) Risk Minimization and Language Modeling in Text Retrieval, PhD thesis, Carnegie Mellon University.
5. Chen, H. and Karger, D. R. (2006) Less is more: probabilistic models for retrieving fewer relevant documents. In Proc. ACM SIGIR, 429-436.
6. Hersh, W. R. and Over, P. (1999) Trec-8 interactive track report. In Proc. of TREC-8.
7. Over P. (1997) TREC-5 Interactive Track Report. In: Proc. TREC-5 29-56.
8. Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008) Novelty and diversity in information retrieval evaluation. In Proc. ACM SIGIR, 659-666.
9. Arni, T., Clough, P., Sanderson, M., Grubinger, M. (2008) Overview of the ImageCLEFPhoto 2008 Photographic Retrieval Task, in CLEF 2008 Working Notes
10. Grubinger, M., Clough, P., Müller, H. and Deselaers, T. (2006) The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems, In Proc. of International Workshop OntoImage'2006 Language Resources for Content-Based Image Retrieval, held in conjunction with LREC'06, Genoa, Italy, pp. 13-23.
11. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A. and Müller, H. (2007), Overview of the ImageCLEF 2006 Photographic Retrieval and Object Annotation Tasks, Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, LNCS Vol. 4730, 2007, 579-594.
12. Grubinger, M., Clough, P., Hanbury, A. and Müller, H. (2007) Overview of the ImageCLEFPhoto 2007 photographic retrieval task. In the Working Notes of the 2007 CLEF Workshop, Budapest, Hungary, 19-21 September 2007.
13. Zhai, C. X., Cohen, W. W., and Lafferty, J. (2003) Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In Proc. ACM SIGIR, 10-17.
14. Arni, T., Tang, J., Sanderson, M., Clough, P. (2008) Creating a test collection to evaluate diversity in image retrieval, in the Proceedings of the Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments, held at SIGIR 2008
15. O'Hare N., Wilkins P., Gurrin C., Newman E., Jones G and Smeaton A. (2008) DCU at ImageCLEFPhoto 2008, In Working Notes of CLEF 2008 Workshop.
16. Chang Y. and Chen H. (2008) Increasing Relevance and Diversity in Photo Retrieval by Result Fusion, In Working Notes of CLEF 2008 Workshop.
17. Song, K., Tian, Y., Gao, W., and Huang, T. (2006) Diversifying the image retrieval results. In Proc. ACM Multimedia, 707-710.