Generic and Spatial Approaches to Image Search Results Diversification

Monica Lestari Paramita, Jiayu Tang, and Mark Sanderson

Department of Information Studies, University of Sheffield, UK,
{lip07mlp, j.tang, m.sanderson}@sheffield.ac.uk

Abstract. We propose a generic diversity and two novel spatial diversity algorithms for (image) search result diversification. The outputs of the algorithms are compared with the standard search results (which contains no diversity implementation) and found to be promising. In particular, the geometric mean spatial diversity algorithm manages to achieve good geographical diversity while not significantly reducing precision. To the best of our knowledge, such a quantitive evaluation of spatial diversity algorithms for context based image retrieval is new to the community.

1 Introduction

Most of the IR algorithms implement an independent ranking approach, where each document is ranked based on its relevance ignoring other documents in the search results [1]. Even though this approach works quite well in most cases, independent ranking suffers when users enter ambiguous or ill-specified queries. Users who enter the query "bat" might implicitly expect information from a particular context, such as animal species, sports equipment, or even a company (British American Tobacco), but the user's particular preference is often unknown by the search engine. This problem can in part be tackled by implementing diversity in the results.

In this paper, we describe both generic and spatial algorithms for increasing the diversity in standard search results. Spatial diversity is a way to present diverse results to users by presenting documents from as many different locations as possible. Location refers to the place where images were captured. Although there has been some research on diversity in general and certain domain specialisations such as person search [2], we believe that the study of spatial diversity is little explored and the quantitive evaluation of both generic and spatial diversity algorithms for image retrieval (described in this paper) is the first time of its kind in the literature.

2 Background

A decade ago, Carbonell and Goldstein [4] appear to be the first to identify the need for diversity. They developed the Maximal Marginal Relevance (MMR) algorithm, which was intended to find documents which were "both relevant to

the query and contains minimal similarity to previously selected documents"[4]. Following on from this, a number of researchers [3, 1] implemented diversity algorithms. Zhai et al [3] mentioned that "a relevant document may be useless to a user if the user has already seen another document with the same content".

Some other studies [5, 6], however, raised the importance of another more specific kind of diversity - spatial diversity. Unlike standard diversity which is applied to the document content, spatial diversity is applied to the location of documents. This topic was examined by [7] who combined spatial information retrieval and diversity to retrieve a spatially diverse result. However, testing of the algorithms was somewhat limited. The implementation of spatial diversity is a new topic which was barely studied in the past, in part due to a lack of appropriate test collections. The use of geographical information to retrieve images in a search engine has been researched by [12] and [11], particularly in automatic extraction and searching within unlabelled images. However none of the studies were dedicated to analyse the importance of spatial diversity in the retrieval process.

2.1 Similarity vs. Diversity

To present results which are both diverse and relevant, an appropriate combination of similarity and diversity must be achieved [8]. Thus, similar documents could be eliminated and relevant documents from different contexts could be added to the results. This approach will increase the probability that users find relevant information regardless of the context they are searching in [1]. Nevertheless, [8] realized that combining both similarity and diversity could be problematic. If diversity is being prioritized, an aspect of similarity would be sacrificed, which means irrelevant documents might reach a high rank.

2.2 Diversity Evaluation

In the recent campaign, ImageCLEFPhoto 2008, participants were encouraged to implement diversity in the retrieval process [9]. The relevant images of each test collection topic were manually grouped into topic clusters (or subtopics) by the organizers. For example, in geographic query, different clusters were represented by different locations of where the images were captured. The diversity of results was assessed by examining the number of topic clusters represented in the top K. The organisers used S-recall [3] as the measurement, as follows

$$\text{S-recall at K} = \frac{|\cup_{i=1}^{K} subtopics(d_i)|}{n_A} \quad (1)$$

S-recall at K, which is also referred as "cluster recall" represents percentage of retrieved subtopics in the top K documents.

3 Diversity Techniques

In the following, we will present our approaches to generic diversity and spatial diversity, both of which increase diversity by re-ranking standard search results.

3.1 Generic Diversity

The generic diversity approach we adopted is based on document clustering. It assumes that each document in the ranked list belongs to a particular sub-topic. In order to diversify the results, the ranked list is re-ordered so that the top documents come from different sub-topics. Since each sub-topic can be considered as a cluster of documents, a variety of clustering algorithms can be utilised. In this work, we explored text based diversity using the Carrot[2][1], which is an open source search results clustering engine. Besides, we choose Lingo in Carrot[2] as the clustering algorithm, which is based on singular value decomposition (SVD). Once the results are clustered by Carrot[2], we use the following procedure to re-rank the results:

1. Denote the ranked list from the search engine as L, a temporary list as T, and a final re-ranked list as R. T and R are empty.
2. Add first document in L to group T, and remove the document from L.
3. Find the document in L which has the highest rank in L and belongs to a cluster not existing in T. Add the document to T and remove it from L.
4. If L is empty, append T to R and exit the procedure.
5. If the number of document in T equals to the total number of clusters, append T to R, and empty T.
6. Do process 3, 4 and 5.

3.2 Spatial Diversity

Two spatial diversity algorithms, which calculate diversity score based on relevance score and spatial distance, were analysed in the following. Diversity scores were then used for re-ranking the documents.

The modified Van Kreveld's Algorithm. In order to promote diversity in results for queries such as "Castles near Koblenz", Van Kreveld [7] proposed an algorithm that takes spatial distance between documents into account while ranking them. However, Van Kreveld's algorithm prioritizes images with shorter distance from the query to be retrieved in higher ranks. Since this study is intended to find images which are distributed as widely as possible within a relevant area, we proposed a modified version, which calculates a diversity score for document j as the following.

$$div\_score_j = \sum_{i \in R} 1 - e^{-\lambda \sqrt{rel(j)^2 + dist(i,j)^2}} \tag{2}$$

where rel is the relevance score generated by Lucene[2] and $dist$ calculates the geographical distance between the locations of document j and i, and i represents a document in the re-ranked list R. Van Kreveld states that the constant λ "is a constant that defines the base e^{λ} of the exponential function". The value of λ was set to 0.5.

[1] http://project.carrot2.org/ [Visited 03/10/08]
[2] http://lucene.apache.org [Visited 03/10/08]

The Geometric Mean Algorithm. Another algorithm tested was one based on the Geometric Mean, which was inspired by the GMAP algorithm of [10] used for aggregating scores across topics. Unlike the previous algorithm, where both relevance and distance score were given the same weight, here relevance was weighted more. This was intended to reduce irrelevant documents located far away from other documents being placed in high ranks. In this study, the constants a and b were set to 1 and 3 respectively (the values were chosen after some empirical testing), ε contains a very small value to avoid $log(0)$.

$$div\_score_j = exp\Big(\frac{\sum_{i \in R} log(dist(i,j)^a rel(j)^b + \varepsilon)}{noOfRankedDocs}\Big) - \varepsilon \tag{3}$$

Once we calculated the diversity scores using the above two algorithms, the following procedure was used for re-ranking documents, using the same terminology as that in Section 3.1.

1. Denote the ranked list from the search engine as L, and a final re-ranked list as R. R is empty.
2. Add first rank document in L to R, and remove it from L.
3. For each document in L, evaluate the diversity score to every document in R.
4. After all documents are scored, choose the document with the highest diversity score, add it to R and remove it from L.
5. Repeat process 3 and 4 iteratively until all of the documents have been added to R.

4 Experiment

We chose Lucene as the search engine for generating standard search results, which were then diversified by the generic and spatial diversity algorithms.

4.1 Experiment Environment

Image Collection. Experiments were conducted on the IAPR TC-12 image collection [9] which comprised of 20,000 images. This collection was used in the ImageCLEFPhoto 2008 campaign. The images came with annotations which described the image content and the location of where the image was taken.

Gazetteer. Spatial diversity requires information about location coordinates, so that distance between locations could be calculated correctly. This work uses the gazetteer from GeoNames which lists all cities in the world with a population greater than 5,000. However, some additional information had to be inputted manually in order to locate all the places in the collection, by using Wikipedia and Google Maps.

Indexing. The text of each image was split into several fields, which were indexed as a normal text document except for the location field, which was parsed to identify location types (eg "city", "country", etc).

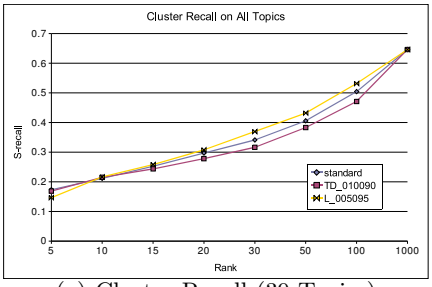
Query. There were 39 topics/queries in ImageCLEFPhoto 2008. Every topic was supplied with a tag "cluster". It contained the cluster type based on which the organizers expected the results to be diversified. For example, if the "cluster" was city, diversity was measured on how many different cities were returned in the top results. For our generic diversity experiment, all topics were used. A subset of 22 geographical topics required support for spatial diversity and were thus used for our experiment on spatial diversity. Spatial diversity algorithms used the "cluster" tag as a reference to the geographical granularity, based on which the algorithms choose, for example, whether to calculate distances between cities or countries. Each query contains information about the title, narrative and a list of sample relevant images. Our search engine used information from the topic title and narrative to build the query. Not all words in the narrative were added. Sentences containing the phrase "not relevant" were filtered out.

4.2 Results

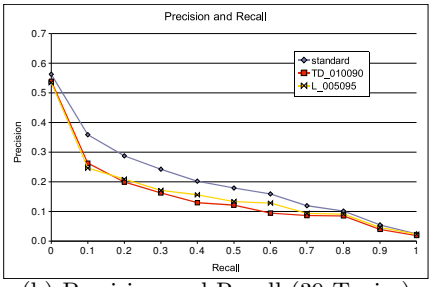
In this section, we present the results of our experiment and compare different diversity approaches. Precision and S-recall will be used for comparing the performances of different approaches, i.e. standard search without diversity, generic diversity and spatial diversity.

Comparison in Generic Diversity. Firstly, we compare the results generated by generic diversity approaches with the standard results without diversity, on all the 39 topics. For the generic diversity approach described in 3.1, we experimented with two groups of runs. The first applied clustering to the "title" and "description" field of the document. The second applied the same clustering to the 17 non-geographic related topics, but restricted cluster to the "location" field for the 22 geographic related topics.

For each group of runs, we also varied the two main parameters of the clustering algorithm (Lingo) in Carrot^2, resulting in four runs - (0.05, 0.95), (0.10, 0.90), (0.15, 0.85) and (0.20, 0.80). The first parameter of each set is the Cluster Assignment Threshold, determining how precise the assignment of documents to clusters should be. A lower threshold assigns more documents to clusters and less to "Other Topics", which contains unclassified documents. The second parameter is the Candidate Cluster Threshold, determining how many clusters Lingo will try to create. Higher values give more clusters. As a result, we have 8 runs in total for generic diversity. Due to limited space, we choose only the best run from each group for comparison, which are (0.10, 0.90) from the first group and (0.05, 0.95) from the second group. They are denoted as TD_010090 and L_005095 respectively.



(a) Cluster Recall (39 Topics)

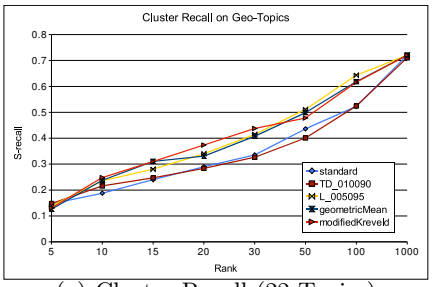


(b) Precision and Recall (39 Topics)

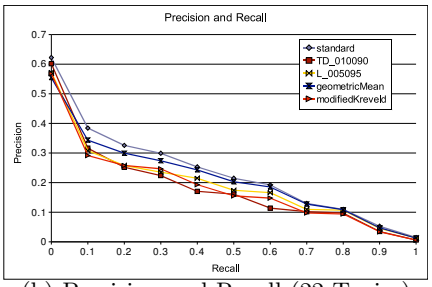
Fig. 1. Evaluation of the generic diversity algorithm on all 39 topics

Figure 1(a) compares the two generic diversity runs with the standard non-diversified run. As can been seen, the L_005095 run managed to improve cluster recall after rank 10, although at rank 5 it performed a little worse than the standard run. The TD_010090 run, however, achieved worse cluster recall than the standard run overall, which ran counter to our expectations. It means that the TD_010090 run promoted too many non-relevant documents to the front of the list. The results also imply that Carrot[2] did not manage to find the correct clusters using the "title" and "description" fields. Clustering based on the "location" field (the second group of runs), however, was able to improve diversity. Figure 1(b) shows the precision and recall curves of the three runs. As expected, re-ranking compromised precisions to some extent.

Comparison in Spatial Diversity. Now we compare spatial diversity with generic diversity and the standard search, on the 22 geographic related topics, 5 different runs were compared, including two spatial diversity runs and the three runs from the previous section.



(a) Cluster Recall (22 Topics)

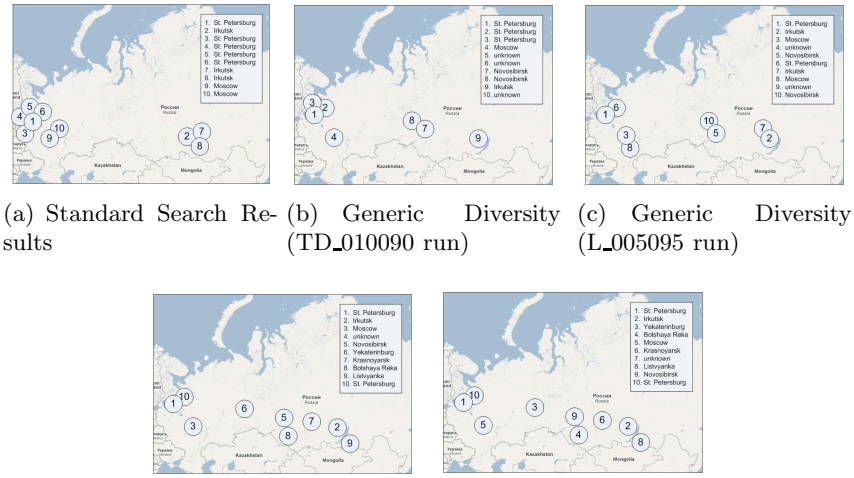


(b) Precision and Recall (22 Topics)

Fig. 2. Evaluation of spatial and generic diversity algorithms on 22 geographical topics

Figure 2(a) compares cluster recall generated by each run. The L_005095 run and the two spatial diversity runs all achieved noticeably better S-recall than the standard run. The modified Van Kreveld algorithm seemed to provide the highest overall diversity. The precision and recall curves are depicted in Figure 2(b). We also calculated Mean Average Precision (MAP) for each run, the results of which are standard (0.1989), CP_010090 (0.1589), L_005095 (0.1664), geometric mean (0.19) and modified Kreveld (0.1546). It can be seen that the standard run still performed the best. All the other diversity runs could not achieve the same precision level as the standard run. However, the geometric mean algorithm managed to achieve very close performance to the standard run in precision, when compared with the other runs. We have conducted a significance test (two tailed T-test) for every pair of runs. By using significance level 0.05, we found that all runs have a p value lower than the significance level except geometric mean which has a value of 0.2836. It implies that the difference between the geometric mean algorithm and the standard run was insignificant. Considering geometric mean's apparent improvement over the standard run in cluster recall, it seems to be fair to say that the geometric mean spatial diversity algorithm outperforms all other runs on the 22 topics overall.

In Figure 3, we mapped the top 10 documents returned for one of the queries from each run, to demonstrate how standard, generic diversity and spatial diversity choose documents differently.



(a) Standard Search Results (b) Generic Diversity (TD_010090 run) (c) Generic Diversity (L_005095 run)

(d) Modified Kreveld's Algorithm (e) Geometric Mean

Fig. 3. Standard and diversified result for Topic 11: "black and white photos of Russia"

5 Conclusions and Future Work

We have proposed both generic and spatial diversity algorithms for diversifying the results of a standard search engine. Based on both precision and cluster recall evaluations, the geometric mean method performs the best in finding both relevant and diverse results on geographic topics. In the future, more relationships will be developed so that systems could handle more types of geographical queries, such as "near", "south of", "30 km from", etc.

6 Acknowledgements

Work undertaken in this paper is supported by the EU-funded TrebleCLEF project (Grant agreement: 215231) and by Tripod (Contract No. 045335).

References

1. Chen, H., Karger, D.R.: Less is More: Probabilistic Models for Retrieving Fewer Relevant Documents. In: SIGIR '06. (2006) 429 – 436
2. Artiles, J., Gonzalo, J., Sekine, S.: The SemEval-2007 WePS Evaluation: Establishing a Benchmark for the Web People Search Task. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic (June 2007) 64–69
3. Zhai, C.X., Cohen, W.W., Lafferty, J.: Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In: SIGIR '06, Toronto, Canada (2003) 10 – 17
4. Carbonell, J.G., Goldstein, J.: The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In: SIGIR '98, Melbourne, Australia, ACM (1998) 335–336
5. Clough, P., Joho, H., Purves, R.: Judging the Spatial Relevance of Documents for GIR. In: ECIR. (2006) 548–552
6. Purves, R.S., Clough, P., Jones, C.B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A.K., Vaid, S., Yang, B.: The Design and Implementation of Spirit: a Spatially Aware Search Engine for Information Retrieval on the Internet. International Journal of Geographical Information Science **21**(7) (2007) 717–745
7. van Kreveld, M., Reinbacher, I., Arampatzis, A., van Zwol, R.: Distributed Ranking Methods for Geographic Information Retrieval. In: Proceedings of the 20th European Workshop on Computational Geometry. (2004) 231–243
8. Smyth, B., McClave, P.: Similarity vs. Diversity. In: Proceedings of the International Conference on Case-Based Reasoning, Springer (2001) 347–361
9. Arni, T., Tang, J., Sanderson, M., Clough, P.: Creating a Test Collection to Evaluate Diversity in Image Retrieval. In: Proceedings of the Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments, held at SIGIR (2008)
10. Robertson, S.: On GMAP: and Other Transformations. In: Conference on Information and Knowledge Management, Virginia, USA (2006) 78–83
11. Rattenbury, T., Good, N., Naaman, M.: Towards Automatic Extraction of Event and Place Semantics From Flickr Tags. In: Proceedings of SIGIR. (2007) 103–110
12. Naaman, M., Paepcke, A., Garcia-Molina, H.: From Where to What: Metadata Sharing for Digital Photographs with Geographic Coordinates. On The Move to Meaningful Internet Systems: CoopIS, DOA, and ODBASE (2003) 196–217