Query Types and Visual Concept-based Post-retrieval Clustering

Masashi Inoue[1] and Piyush Grover[2]

[1] National Institute of Informatics
m.inoue@acm.org
[2] Indian Institute of Technology, Kharagpur*
pgrover@cse.iitkgp.ernet.in

Abstract. In the photo retrieval task of ImageCLEF 2008, we examined the influence of image representations, clustering methods, and query types in enhancing result diversity. Two types of visual concept vectors and hierarchical and partitioning clustering as post-retrieval clustering methods were compared. We used the title fields in the search topics, and either only the title field or both the title and description fields of the annotations were in English. The experimental results showed that one type of visual concept representation dominated the other except under one condition. Also, it was found that hierarchical clustering can enhance instance recall while preserving the precision when the threshold parameters are appropriately set. In contrast, partitioning clustering degraded the results. We also categorized the queries into geographical and non-geographical, and found that the geographical queries are relatively easy in terms of the precision of retrieval results and post-retrieval clustering also works better for them.

1 Introduction

The target of this year's ImageCLEFphoto ad hoc task is to enhance the topical diversity in retrieved results. Usually, instance recall is measured by counting the number of correctly retrieved relevant documents. To measure the topical diversity of retrieved images, the instances are assumed to be the topic. This change in measurement is intended to partially reflect a user's potential needs, which is that many users look at many different choices in terms of the objects or topics given in the retrieval results. For example, if a search topic is associated with the <city> criterion, all the images of the same city in the retrieval results are considered the same in terms of value for the user. Similarly, all images whose subjects are the same species of animal, they are treated as the same if <animal> is the criterion for the search topic. Assuming this model of a user's preference is true, the results should be diverse, which includes as many different objects or topics as possible. To address this problem, we examined the utility of

* This work was conducted while the author was with the National Institute of Informatics

clustering techniques that are based on visual content after acquiring the initial ranking that was based solely on textual annotations. We can assume that the topical diversity of the images in the top range of the ranked list will increase by using only representative images from the clusters. The experimental procedures, algorithms used, and experimental results, will be explained and discussed in the following sections.

2 Experimental Setup

2.1 Initial Retrieval

We used the ImageCLEFphoto 2008 ad hoc test collection that consists of 39 search topics, and 20,000 images with structured annotations for this research. The design of the task is explained in [1]. It consists of a monolingual collection in English and a mixed language collection in English and German. We used only the monolingual English collection and all of our queries were in English.

As the retrieval engine, we used the Terrier Information Retrieval Platform[3] for all the textual processing including the pre-processing of the image annotations, indexing, and the matching between queries and indexes. As for the pre-processing, the default stop-word word list and the Porter's stemming in the Terrier toolkit were used.

We tested two variations in the indexing: First, we indexed only the <TITLE> field of the image annotations. In the second case we used both the <TITLE> and <DESCRIPTION> fields for the indexing. All the retrieval experiments were performed on both indexes. In the indexes, the words are assigned weights. The weights are determined by the retrieval model used. The retrieval models also specify the scoring of a particular document when given the query. In the Terrier toolkit, the ranking of documents follows the framework of divergence from randomness (DFR).

The Terrier IR platform offers a variety of retrieval models. To obtain a reasonable baseline retrieval system, we selected the models and their parameters based on our pilot runs. In the pilot runs, we did not use any formal training collection, but we compared the retrieval results on the test collection through the manual inspection of the relevance regarding the several top ranked images. In this process, all topics provided for 2008 were used, but we did not tune the models for each query but the same models and parameter values were used throughout the queries. Therefore, these retrieval models and parameter values returns some reasonable results but not optimal for the test collection. When we constructed indices for the collection using only the <TITLE> field of the annotations, we used the following **IFB2 DFR** model.

$$w(t,d) = \frac{F+1}{n_t \cdot (tfn+1)} \big(tfn \cdot \log_2 \frac{N+1}{F+0.5} \big) \tag{1}$$

[3] http://ir.dcs.gla.ac.uk/terrier/

where tf is the within-document frequency of t in d, N is the number of documents in the entire collection, F is the term frequency of t in the entire collection, and n_t is the document frequency of t. tfn is the normalised term frequency. This is given by **n**ormalisation 2:

$$tfn = tf \cdot \log_2(1 + c \cdot \frac{\bar{l}}{l_d}),$$

where l_d is the document length of d, which is the number of tokens in d, $\bar{l}$ is the average document length in the collection, and c is a tuning parameter. We set the parameter to $c = 2.5$.

When we used the <TITLE> and <DESCRIPTION> fields of the image annotations for the indexing, we used the following **In_expC2 DFR** model with $c = 1.1$.

$$w(t,d) = \frac{F+1}{n_t \cdot (tfn_e + 1)} \big(tfn_e \cdot \log_2 \frac{N+1}{n_e + 0.5} \big) \tag{2}$$

.

Our retrieval task consists of two main stages. In the first stage we obtained the retrieval results by using only the indexed data, which is the text retrieval, and the <TITLE> field of the queries in the topic file. The submitted runs corresponding to the text only retrieval were named as follows:
1. EN-EN-TXT-TITLE-AUTO.res
2. EN-EN-TXT-TITDESC-AUTO.res
where TITLE means only the <TITLE> fields were used and TITDESC corresponds to the runs in which both the <TITLE> and <DESCRIPTION> fields were used. Both of them were automatic runs with automatic query expansion by the BE1 model. For the former run, the IFB2 model was used and, the In_expC2 model was used for the latter run. These runs correspond to the baseline conditions for our experiments.

2.2 Post-retrieval Clustering

Diversification by clustering The initial ranking obtained using only the text contains many duplicate or near duplicate images in terms of their topics. Thus, the retrieved images were clustered to include diverse image sets in the limited window size of the retrieval results, which was 20 in our case. Topically similar images in clusters were represented by the most representative image and did not appear in the final ranked list. As a result, we were able to include diverse types of images on the screen.

Different features can be used in determining the clusters. We used the visual concept vectors that were the semantic concepts extracted from the raw visual signals of the images. These concepts were prepared for the VCDT 2008 task [3]. Although the appearance of the images does not directly correspond to the clustering topical criteria, as we have already used text features in obtaining the initial scores for the documents, we may use another feature of the documents to

compensate for the lack of detail in the ranking. We applied two simple clustering approaches to the results obtained from the text retrieval to diversify the final results.

visual concept vectors visual concept vectors are different from raw visual signals, but they are the semantic entities represented by word tokens that correspond to the visual content in images. Therefore, later on, they can be used as an extra vocabulary. The concepts are extracted using various image processing and pattern recognition techniques. We used two visual concept vectors files:

1. *DISC*—annotations created by Thomas Deselaers from RWTH Aachen University following the described method [2]
2. *CONT*—annotations created by Jean-Michel Renders from XEROX Europe following the method in [6]

The first concept set is labeled DISC because their values are discrete and each image contains concepts represented as binary values. The second concept set is labeled CONT because their values are continuous and each image contains concepts probabilistically. Since automatic image annotation is a difficult task, it contains some errors. We use them with inherent noise.

Hierarchical clustering approach The first approach is based on a hierarchical clustering in which we produced a dendrogram using the visual concept vectors of the initial ranking given a particular query. Here, we explain the clustering process. All retrieved images that have some relevance scores are clustered. The process is further explained in [5] using an example. Let the number of images in the initial ranking be N; then, each image is represented by its rank from 1 to N. The Euclidean distance between two images represented by concept vectors was used to create the image pairs regardless of the initial index. In the next step, this cluster forms a new higher level cluster with another individual node or cluster. Cluster centers are defined as the mean value of the concept features for member images. The new distance is calculated between the new cluster center and the neighboring new cluster center.

Once the dendrogram has been constructed, we have to decide which granularity we should use to constitute a new ranked list. The dendrogram was sliced at a certain distance level. For both indexing and both visual concept vectors, we changed the distance values for the threshold value from 1.6 to 0.7 at a step size of 0.1. We select the representative images in the clusters at these 9 different levels from the higher values to the lower ones. These parameter values were selected based on the manual inspection of the retrieval results for the 2008 queries. We fixed the the values that returns seemingly reasonable results for all queries. Once we have set the threshold, in the final clusters, images with the smaller index number are regarded as the representative images because the smaller index number indicates a higher original relevance score. In our example, since we start this merging process from a distance level of 1.6 and come down to 0.7, we first make clusters and obtain the representative images for all the

clusters at a distance level of 1.6. They will be included in the modified rankings, but their positions have not yet been determined at this point. In the next step, as we come down to a distance level of 1.5, we select the representative images at this distance level. If they are not chosen already, we modify this new image score to the initial retrieval score divided by *level*, which is the step number the process has passed through (here it is 2). This score adjustment is made because we want to topically shuffle the new ranked list. The representative images of the clusters in the lower levels that are visually quite similar to the images that are already placed in the new ranked list have smaller scores and are placed in the lower rankings. Similarly, we continue going down until we reach a distance level of 0.7. After getting all the representative images up to the last level (here the 9th level) and their scores have all been modified, we sort the list according to the new scores and obtain the final modified ranked list for a particular query. We used a threshold value ranging *0.7-1.6* for all our experimental runs. The step size and ranges were determined by conducting a manual inspection of the clustered results.

K-means clustering approach As a second approach, we applied k-means clustering to the visual concept vectors of the all resulting images obtained by the text retrieval of a particular query. Our clustering process itself is the same as an ordinary k-means clustering. If we randomly assign the initial K means, the final result will also contain randomness and then it becomes difficult to compare the differing conditions. To avoid such randomness, a modification of initialization of k-means clustering was made. K initial cluster centers were evenly allocated in the initial ranked list. Another modification lies in the representative image selection process. We use the densities of the clusters. If a cluster is dense, we assume that the cluster contains near identical images homogeneously; thus, only representative images are included in the final ranking. On the other hand, if clusters are sparse, they likely contain different concepts; therefore, we include all the diverse images in the cluster. In the k-means method, original scores are used in sorting candidate representative images for the final ranking. The details of these procedures are explained by using the pseudo codes in [5].

3 Experimental Results

The two evaluation measures for our submitted runs that were used were precision at the 20th document (P@20) and cluster recall at the 20th document (CR@20). The goal of post-retrieval clustering is to enhance cluster recall. Therefore, a small drop in precision is acceptable as long as we can sufficiently enhance the cluster recall. Degradation may happen because very relevant images of the same categories are removed from the ranked list. To summarize this, we want to improve CR@20 while minimizing the degradation of the precision.

Table 1 shows the results of the two measures. A clear difference in the upper half of the table (<TITLE> only) and the lower half of it (<TITLE> and <DESCRIPTION>) can be seen. More information given in the description

Table 1. Precision at 20, Cluster Recall at 20, and F-measure are shown. The cluster recall scores for both media that are better than the text-only runs are marked with boldface.

Run Name	P@20	CR@20	F-measure
EN-EN-TXT-TITLE-AUTO	0.1397	0.1858	0.1620
EN-EN-TXTIMG-TITLE-CONT-Kmeans-AUTO	0.0654	0.1201	0.0858
EN-EN-TXTIMG-TITLE-DISC-Kmeans-AUTO	0.0859	0.1431	0.1063
EN-EN-TXTIMG-TITLE-CONT-0.7-1.6-AUTO	0.1372	**0.1941**	0.1599
EN-EN-TXTIMG-TITLE-DISC-0.7-1.6-AUTO	0.1090	0.1827	0.1365
EN-EN-TXT-TITDESC-AUTO	0.2090	0.2409	0.2238
EN-EN-TXTIMG-TITDESC-CONT-Kmeans-AUTO	0.1115	0.2062	0.1447
EN-EN-TXTIMG-TITDESC-DISC-Kmeans-AUTO	0.1090	0.1730	0.1337
EN-EN-TXTIMG-TITDESC-CONT-0.7-1.6-AUTO	0.1859	**0.3027**	0.2303
EN-EN-TXTIMG-TITDESC-DISC-0.7-1.6-AUTO	0.1590	**0.2703**	0.2002

fields resulted in better P@20 and CR@20 scores. Also, between the two clustering methods, the modified k-means algorithm was not effective. Although it is not systematic, the difference between the title field only runs and the title and description field runs suggest that a good initial performance may lead to bigger improvement when clustering is used.

4 Discussion

4.1 Query and cluster topic dependency

The clustering criteria used to calculate the instance recall can be divided into two groups: geographical criteria, such as the country or city, and others such as the objects. The geographical categorization is based on the official clustering criteria. The geographical criteria include the name of country, name of city, or just location. Geographical criteria dominate about 60% of criteria among all 39 topics. The query numbers for each category are listed in Table 2. The topic dependencies may influence the effectiveness of the post-clustering. Table 3 shows the difference in precision at 20 values for different categorizations. Since the CONT feature usually works better than the DISC feature and only hierarchical clustering could enhance the instance recall as discussed in Sec. 3, we only examined the CONT-0.7-1.6 conditions here. The queries that are associated with the geographical clustering criteria achieved a higher precision in the initial retrieval and after clustering. A similar tendency was observed in the cluster recall values. Actually, in non-geographical topics, clustering damaged the cluster recall scores but enhanced the precision scores for the TITLE only condition. When both TITLE and DESCRIPTION fields were used, cluster recall had been improved in both geographical and non-geographical queries; however, compared with the notable gain in geographical queries, the change in non-graphical ones can be considered marginal. The reasons why images of geographical topics can

Table 2. Categorization of queries based on clustering criteria.

	Query Number
Geographical:	2 6 10 11 12 13 15 17 18 19 21 24 28 34 40 41 43 44 50 53 54 55 58
Non-geographical:	3 5 16 20 23 29 31 35 37 39 48 49 52 56 59 60

Table 3. This table shows a performance comparison among the query groups defined in Table 2 under the CONT-0.7-1.6 condition. The changes in retrieval effectiveness before (text-only: t/o) and after clustering (clstd) are shown in terms of the precision (PR) and cluster recall (CR) values at 20th rankings. The scores after clustering that are better than the text-only runs are marked with boldface.

TITLE only

Query groups	P@20 (t/o)	P@20 (clstd)	CR@20 (t/o)	CR@20 (clstd)
All queries	0.1397	0.1372	0.1858	**0.1941**
Geographical queries	0.1500	0.1413	0.1878	**0.2080**
Non-geographical queries	0.1250	**0.1313**	0.1828	0.1742

TITLE & DESCRIPTION

Query groups	P@20 (t/o)	P@20 (clstd)	CR@20 (t/o)	CR@20 (clstd)
All queries	0.2090	0.1859	0.2409	**0.3027**
Geographical queries	0.2130	0.1983	0.2522	**0.3523**
Non-geographical queries	0.2031	0.1488	0.2247	**0.2315**

be clustered well by visual content should be examined in the future. The higher initial precision due to the existence of proper names for geographical queries may explain part of this phenomenon. Another possible hypothesis is that the geographical topics are associated with landmarks that are easier to identify visually.

4.2 Multilingual Retrieval

In our experiment, we used only a monolingual corpus. When the target collection images are annotated in different languages, the initial ranked list given by the text retrieval contains few relevant images. The post-retrieval clustering methods used here eliminate any redundancy found in the top region of the ranked list, but do not actively search for lower ranked hidden relevant images. If our method is used in the multilingual setting, some new methods are needed to enhance the initial relevant retrieved set. Existing techniques for multilingual image retrieval that rely on visual near-identity such as [4] can be used together with this post-retrieval clustering approach because they use the visual similarity in opposite ways.

4.3 Evaluation Measures

The new evaluation measure used in this year's experiments is a cluster recall whose relevance to the ad hoc tourist photo retrieval task has not yet been clarified. The relationship between the utility that users may choose and the increase in cluster recall should be examined. Also, the conventional P@20 measure and the cluster recall are not orthogonal in evaluating ranked lists. Both of them count the number of relevant images in the top region of the ranked lists.

5 Conclusion

We have experimentally compared two post-retrieval clustering methods relying on two types of visual concept vectors that were derived from the images. The experimental results of a monolingual retrieval showed that the use of hierarchical clustering can enhance the instance recall such that the top ranked images are diverse in terms of the topics. Also, we found that the clustering criteria that are assigned to search topics influence the improvement of scores. Generally, the benefit of post-clustering is observed when images are clustered with geographical perspectives. To make our results more reliable, we should further examine the following points: the use of perfectly created visual concept vectors based on the ground truth data, and a comparison between the extracted high-level visual concept vectors and the low-level feature values themselves in the clustering. Future research topics may include the automation of thresholding in the clustering methods that is now manually set by results inspection. The categorization of queries in other criteria such as whether they are context-oriented or content-oriented might be interesting.

References

1. Thomas Arni, Paul Clough, Mark Sanderson, and Michael Grubinger. Overview of the ImageCLEFphoto 2008 photographic retrieval task. In *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, September 2008 (printed in 2009).
2. T. Deselaers, D. Keysers, and H. Ney. Discriminative training for object recognition using image patches. In *CVPR*, volume 2, pages 157–162, San Diego, CA, USA, June 2005.
3. Thomas Deselaers and Allan Hanbury. The visual concept detection task in ImageCLEF 2008. In *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, September 2008 (printed in 2009).
4. Masashi Inoue. Mining visual knowledge for multi-lingual image retrieval. In *DMIR-07*, volume 1, pages 307–312, Niagara Falls, Ontario, Canada, May 21-23 2007.
5. Masashi Inoue and Piyush Grover. Effects of visual concept-based post-retrieval clustering in imageclefphoto 2008. In *9th Workshop of the Cross-Language Evaluation Forum*, 2008.
6. F. Perronin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, Minneapolis, Minnesota, US, 18-23 June 2007.