Effects of Visual Concept-based Post-retrieval Clustering in ImageCLEFphoto 2008

Masashi Inoue and Piyush Grover
National Institute of Informatics,Tokyo & Indian Institute of Technology, Kharagpur*
m-inoue@nii.ac.jp, pgrover@cse.iitkgp.ernet.in

Abstract

We examined the effectiveness of post-retrieval clustering that was based on the visual similarities among images to enhance the instance recall in the photo retrieval task of ImageCLEF 2008. The visual similarities are defined by the example visual concepts that were provided for the automatic photo indexing task. We tested two types of visual concepts and two kinds of clustering methods, hierarchical and modified k-means clustering. In all the runs, we used only the title fields in the search topics; we used either only the title fields or both the title and description fields of the annotations in English. The experimental results showed that hierarchical clustering can enhance instance recall while preserving the precision when certain parameters are appropriately set.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.2 Information Search and Retrieval;

General Terms

Experimentation

Keywords

information retrieval, image retrieval, clustering, k-means algorithm, evaluation

*This work has been done while the author was with National Institute of Informatics

1 Introduction

The target of this year's ImageCLEFphoto ad hoc task is to enhance instance recall: a new evaluation measure that counts the number of relevant images after topically removing same image in the ranked list. This measure 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 images of a city in the retrieval results are considered the same in terms of value for the user. Similarly, all images containing the same species of animal are treated as an image if <animal> is the criterion for the search topic. Assuming this model of user preference to be 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 clustering techniques that are based on visual content after acquiring the initial ranking that was based solely on textual annotations. By using only representative images from the clusters, we assume that the topical diversity of the images in the top range of the ranked list will increase. The experimental procedures, algorithms used, and findings will be explained in the following sections.

2 Experimental Setup

2.1 Test Collection

We used the ImageCLEFphoto 2008 ad hoc test collection that consists of 39 search topics, and $20,000$ images with structured annotations. The details of this collection are given in [4]. It consists of a monolingual collection in English and a mixed language collection in English and German. In our submitted runs, we used only the monolingual collection. In addition, our queries were all in English. Therefore, in all our submitted runs, the run names begin with **EN-EN**, which indicates that both the queries and annotations are in English.

2.2 Indexing and retrieval models for textual perspective

We used the Terrier Information Retrieval Platform[1] 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 stopword-list in the Terrier toolkit was used —all the words that were tagged as prepositions, conjunctions, particles and interjections were removed from the indexing. After removing the stop-words, the Porter's stemming algorithm [1] was used by selecting the default option **PorterStemmer** in Terrier.

We constructed two variations in indexing: First, we indexed only the <TITLE> field of the image annotations. In the second case we used both <TITLE> and <DESCRIPTION> fields for the indexing. All the retrieval

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

experiments were performed on both indexes. In the indexes, the words are assigned weights. The weights are determined by the retrieval model used. Retrieval models also specify the scoring of a particular document when given the query. In the Terrier toolkit, the ranking of documents follows the concept of divergence from randomness (DFR). The relevance score of a document d for a query $\mathbf{q}$ is given by

$$score(d, \mathbf{q}) = \sum_{t \in \mathbf{q}} qtw \cdot w(t, d), \tag{1}$$

where t is a query term in $\mathbf{q}$, and qtw is the query term weight that is given by $qtf/qtf_{\max}$. Here, qtf is the query term frequency and $qtf_{\max}$ is the maximum qtf among all the query terms. $w(t, d)$ is the weight of document d for a query term t that is determined by the DFR models.

The Terrier IR platform offers a variety of retrieval models. We selected the models and their parameters based on our pilot runs; we compared the retrieval results through observation. Therefore, the optimality of these retrieval models is not guaranteed.

When we constructed indices for the collection using only the <TITLE> field of the annotations, we used the following **IFB2 DFR** model.

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

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, 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 $c = 1.1$.

$$w(t, d) = \frac{F + 1}{n_t \cdot (tfn_e + 1)} \big(tfn_e \cdot \log_2 \frac{N + 1}{n_e + 0.5}\big) \tag{3}$$

.

The Terrier toolkit also offers an automatic query expansion functionality. We used the **Bose-Einstein 1** (Bo1) term weighing model with a parameter free approach in all of our runs.

2.3 Initial Retrieval

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 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 later run, as explained in 2.2. These runs correspond to the baseline conditions for our experiments.

2.4 Post-retrieval Clustering

2.4.1 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, 20 in our case. Similar images were removed from the results except for one representative image for each cluster. As a result, we were able to include diverse types of images in the results.

Different features can be used in determining the clusters. We used the visual concepts that were semantic concepts extracted from the raw visual signals of images. Although the appearance of 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 in 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.

2.4.2 Visual concepts

Visual concepts 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 concepts files:
(1) *deselaers-db*—annotations created by Thomas Deselaers from RWTH Aachen University following the described method [2]
(2) *clef_base_annot*—annotations created by Jean-Michel Renders from XEROX Europe following the method in [3]

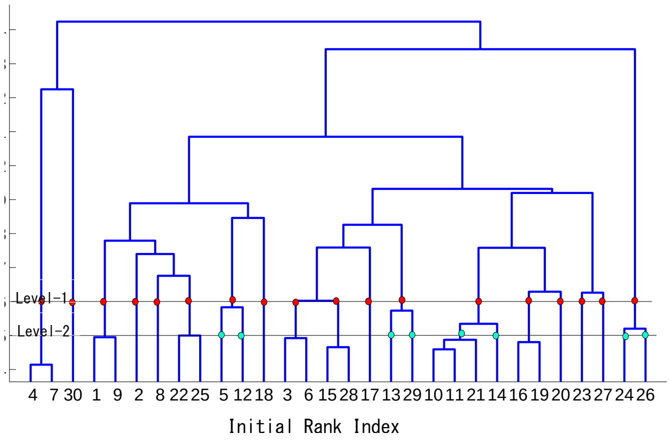

Figure 1: Dendrogram of the results from clustering top 30 images after initial text only retrieval for query number 2. Here, the y-axis represents the Euclidean distance between the two clusters based on the *clef_base_annot* visual concepts and the x-axis indicates the image indices.

Since the automatic image annotation is difficult task, they contains some errors. We use them with inherent noises. 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.

2.4.3 Hierarchical clustering approach

The first approach is based on a hierarchical clustering, in which we produced a dendrogram using the visual concepts of the initial ranking given a particular query. Figure 1 is an example of a dendrogram constructed using *clef_base_annot* visual-concepts for the first query in the task. We will explain the hierarchical clustering-based rank modification method by using the example data listed in Table 1. These data were the initial retrieval results for query 2 using the *deselaers-db* features. The first two columns of the matrix show the image or cluster indices and the third column represents the pair-wise Euclidean distance between the two images based on the visual concepts. Note that the distance 0 does not mean the two images are identical. They are indistinguishable when the visual concepts are used to calculate similarities.

Let the number of images in the initial ranking be N; then, each image is represented by its rank from 1 to N. For this particular case, the dendrogram is formed in such a way that the 116 and 127 images constitute a cluster whose distance level is 0.0 and its center is the average between 116 and 127. In the new cluster, an image of the smallest index number is regarded as the representative image because smaller index number indicates higher original relevance score.

Table 1: Example dataset for hierarchal clustering.

image index #1	image index #2	distance
116	127	0.000000
97	99	0.000000
78	80	0.000000
34	156	0.000000
19	20	0.000000
17	50	0.101948
48	90	0.142174
221	224	0.193464
76	83	0.200342

Here, the representative image of this cluster is 116 which is smaller than the other index 127. In the next step, this cluster forms a new cluster with another individual node or cluster. The new distance is calculated between the new cluster center and the neighboring new cluster center that consists of image 97 and 99.

Once the dendrogram has been constructed, we have to decide which granularity we should use to constitute a new ranked list. If we use only a higher level dendrogram, the resulting cluster may miss many useful images. However, if we draw a line at a level that is too low, the modified ranking will almost be the same as the initial ranking. The dendrogram was sliced at a certain distance level (horizontal lines in Fig. 1). For both indexing and both visual concepts, 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. In Figure 1, the circles on each level are the representative images of each cluster that lies below the circle. After the clustering and the selection of the representative image, the score is modified as

$$Score_{\text{new}} = \frac{Score_{\text{old}}}{level_{\text{no}}}. \tag{4}$$

This modification is made because we want to topically shuffle the new ranked list. The representative images of the clusters in lower levels that are visually quite similar have smaller scores and placed in the lower places of the new ranked list. 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 representative images at this distance level. If they are already chosen in the upper levels we do not do anything else, but if they are not chosen we modify this new image score to the initial retrieval score divided by $level_{\text{no}}$, which is the step number the process has passed through (here it is 2). 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.

The submitted runs corresponding to this algorithm contain either DISC-0.7-1.6 or CONT-0.7-1.6 in their names. Here *0.7-1.6* represents the threshold range.

2.4.4 K-means clustering approach

As a second approach, we applied k-means clustering to the visual concepts of the resulting images obtained by the text retrieval of a particular query. Our clustering process itself is the same as 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, we can try many random initializations, and take the most frequent result as the final candidate. However, to simplify the computation, the initial K means are assigned deterministically. The following set of equations extract initial K centers evenly from the initial ranked list of the size N with the step size R:

$$\begin{cases} R = \lfloor \frac{N}{K} \rfloor, \\ \mathbf{m}_i = \mathbf{x}_j \\ \text{ where } j = i \times R \quad \forall\ i = 1...K, \end{cases} \tag{5}$$

where $\mathbf{m}_i$ is the ith mean vector, and $\mathbf{x}_j$ is the jth visual-feature vector (visual-concept vector corresponding to the jth result). For example, if we have 383 ranked images in our initial list and $K = 10$, then the step size $R = 383/10 = 38$ and the initial cluster centers are $38, 76, ..., 380$.

Another modification lies at the representative image selection process. The process after k-means clustering is shown in Table 2. We calculate 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 are likely contains different concepts; therefore, we include all diverse images in the cluster. In the k-means method, original scores are used in sorting candidate representative images for the final ranking.

3 Experimental Results

The two evaluation measures for our submitted runs were used: precision at the 20th document (P@20) and the cluster recall at the 20th document (CR@20). P@20 is evaluated as usual. CR@20 is measured by utilizing the <cluster> fields attached with search topics. Although the relative outcomes of different runs are different for different reference points (here the 20th document in the

Table 2: K-means clustering re-ranking algorithm by using visual concepts.

```
• Cluster ranked images using k-means algorithms.
From $N$ images given in the text-only retrieval result,
$K$ clusters are constructed.

• Compute the density of each cluster.
Initialize a sparse index matrix $Z$ of size $N \times K$.
    and $Z(i,k) = 1$ if $\mathbf{x}_i \in X_{1..N}$ is allocated to cluster $k$.
for  $k = 1$ to $K$
    $s(k) = \sum_{i=1}^{N} D(i,k) * Z(i,k)$
        where $s(k)$ is the sparsity of $k$th cluster
        and $D(i,k)$ is the distance between $i$th image and $k$th cluster center.
end

• Select representative images from clusters.
The threshold value $T = \mathrm{mode}(s)$.
for $k = 1$ to $K$
    if $s(k) \leq T$
        append $\mathrm{RIOC}(k)$ to $Bucket$.
            where $\mathrm{RIOC}(k)$ find the representative image of a cluster $k$
            which is most close to the cluster center.
    else
        append all points of cluster $k$ to $Bucket$.
    end
end
$newResult = \mathrm{sort}_{\mathrm{score}}(Bucket)$
```

ranked list), as those measures are mainly used in the campaign, we interpret our results based on them. The goal of post-retrieval clustering is to enhance cluster recall. Therefore, a small drop in precision is acceptable as long as we can enhance the cluster recall sufficiently. 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 keeping the degradation of the precision at a minimum.

Table 3 shows the results on 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 fields resulted in better scores in both P@20 and CR@20. 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

Table 3: Experimental results for submitted runs by NII group: Precision at 20 and Cluster Recall at 20 are shown. The cluster recall scores using both media that are better than the text-only runs are marked with asterisks.

Run Name	P@20	CR@20
EN-EN-TXT-TITLE-AUTO	0.1397	0.1858
EN-EN-TXTIMG-TITLE-CONT-Kmeans-AUTO	0.0654	0.1201
EN-EN-TXTIMG-TITLE-DISC-Kmeans-AUTO	0.0859	0.1431
EN-EN-TXTIMG-TITLE-CONT-0.7-1.6-AUTO	0.1372	*0.1941
EN-EN-TXTIMG-TITLE-DISC-0.7-1.6-AUTO	0.1090	0.1827
EN-EN-TXT-TITDESC-AUTO	0.2090	0.2409
EN-EN-TXTIMG-TITDESC-CONT-Kmeans-AUTO	0.1115	0.2062
EN-EN-TXTIMG-TITDESC-DISC-Kmeans-AUTO	0.1090	0.1730
EN-EN-TXTIMG-TITDESC-CONT-0.7-1.6-AUTO	0.1859	*0.3027
EN-EN-TXTIMG-TITDESC-DISC-0.7-1.6-AUTO	0.1590	*0.2703

and description field runs suggest that a good initial performance may lead to bigger improvement when clustering is used.

4 Discussion

4.1 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 in which users may consider and the increase in cluster recall should be examined. Also, the conventional measure P@20 and the cluster recall are not orthogonal in evaluating ranked lists. Both of them appreciate larger number of relevant images in the top region of the ranked lists.

4.2 Query and cluster topic dependency

The clustering criteria used to calculate instance recall can be divided into two groups: geographical criteria, such as the country or city, and others such as objects. The influence of these differences as well as the topic dependencies may influence the effectiveness of the post-clustering and should be further examined.

4.3 Multilingual Retrieval

In our experiment, we used only 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 redundancy but do not actively include hidden relevant images. Existing techniques for multilingual image retrieval that rely on the visual near-identity such as [5] will not work with this post-retrieval clustering approaches because the use the visual similarity in opposite ways. If our method is used in multilingual setting, some new methods is needed to enhance initial relevant retrieved set.

5 Conclusion

We have experimentally compared two post-retrieval clustering methods relying on two types of visual concepts that were derived from the images. The experimental results of 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 topics. To make our results more reliable, we should further examine the following points: the use of perfectly created visual concepts based on the ground truth data, and a comparison between the extracted high-level visual concepts and low-level feature values themselves in the clustering.

Future research topics may include the automation of thresholding in the clustering methods. Also, the relationship between the degrees of goodness in the initial retrieval results using only text and the effectiveness of clustering when using visual features should be clarified in a more systematic way.

References

[1] Porter, M.F., “An algorithm for suffix stripping, Program”, Vol. 14, No. 3, pp 130–137 1980.

[2] Deselaers, T., Keysers, D., and Ney, H., “Discriminative Training for Object Recognition using Image Patches”, CVPR, Vol. 2, pp. 157–162, San Diego, CA, USA, June 2005.

[3] Perronin F. and Dance, C., “Fisher Kernels on Visual Vocabularies for Image Categorization”. CVPR, Minneapolis, Minnesota, US, 18-23 June 2007.

[4] Grubinger, M., Clough, P., M´uller, H., and Deselaers, T. “The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems”, In Proceedings of International Workshop OntoImage2006 Language Resources for Content-Based Image Retrieval, held in conjuction with LREC 2006, pp. 13–23, Genoa, Italy, 22 May 2006.

[5] Inoue, M., “Mining Visual Knowledge for Multi-Lingual Image Retrieval”, DMIR-07, Vol. 1, pp. 307–312, Niagara Falls, Ontario, Canada, May 21-23 2007.