Aggregation of Multiple Pseudo-Relevance Feedbacks for Image Search Reranking

Wei-Chao Lin\textsuperscript{1,2}
\textsuperscript{1}Department of Information Management, Chang Gung University, Taoyuan 333, Taiwan
\textsuperscript{2}Department of Thoracic Surgery, Chang Gung Memorial Hospital, Linkou, Taoyuan 333, Taiwan
Corresponding author: Wei-Chao Lin (e-mail: viclin@gap.cgu.edu.tw).

The work was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST 106-2410-H-182-024, and in part by Chang Gung Memorial Hospital, Linkou, Taoyuan, Taiwan under Grant NERPD2G0301.

ABSTRACT Image retrieval performance can be improved via pseudo-relevance feedback (PRF), which automatically uses the top-\textit{k} images of the initial retrieval result as the pseudo feedback. Several strategies are available for performing PRF, which lead to different search results. In this paper, we focus on image search reranking via search result aggregation as a hybrid approach. Various combinations of the original retrieval result with the results of PRF using pseudo-positive and pseudo-negative feedbacks, which are obtained by the Borda count, are compared. The results of our experiments, which are carried out on the NUS-WIDE-LIT and Caltech 256 datasets, demonstrate that search result aggregation can outperform PRF in retrieval. In particular, the combination of the original result and the result of PRF using pseudo-positive feedback outperforms the other reranking strategies by PRF using both pseudo-positive and negative feedbacks and their combination with the original result.

INDEX TERMS Image retrieval; reranking; pseudo-relevance feedback; Borda count

I. INTRODUCTION
Content-based image retrieval (CBIR) focuses on automatically extracting low-level image features, such as color, texture, and shape, to index images. In the stage of image retrieval, the indexed image features are used for the similarity measurement. In this framework, each image is represented as a visual feature vector, which is a point in a high-dimensional feature space. The similarity between two images is based on the distance between their feature vectors. Two similar images are likely close to each other in the feature space. Images that are close to the query can be regarded as having similar contents to the query.

However, the semantic gap problem is typically encountered in CBIR. The semantic gap is the gap between the low-level features that are automatically extracted and measured in terms of their similarities by computers and the high-level concepts or semantics in users’ minds. In practice, CBIR systems have difficulty effectively satisfying users’ requirements [1]. Therefore, the retrieval performances of CBIR systems are not high.

To overcome this problem, a postretrieval or image reranking step can be employed. Image reranking can be defined as follows: For a Web image search, given an initial text search, a set of images that are ordered according to descending similarity is returned. The image reranking process is executed to reorder the images according to their visual similarities to improve the initial retrieval result [2, 3].

Pseudo-relevance feedback (PRF) is a simple and widely used approach for improving initial search results. In traditional relevance feedback (RF), users must manually provide positive and/or negative feedbacks to the systems [4, 5, 6]. In contrast, according to the initial retrieval result, PRF assumes that a fraction of the top- and bottom-ranked images can be used as pseudo-positive and pseudo-negative feedbacks, respectively, to the systems. According to these feedbacks, a model can be constructed to rerank the search result set [2, 7, 8]. This approach can iteratively remodify the query vector based on the pseudo-positive and/or pseudo-negative feedbacks to move the query toward more relevant images and away from irrelevant images in the multidimensional vector space where the images are represented by vectors [9].

Although PRF avoids the user-in-the-loop process that is required in RF, the top-retrieved images that are used as the pseudo-relevance feedback set may contain noise, namely, their semantic contents are not similar to the semantic content of the query image. In addition, for some query
results, many samples in the relevant set can be highly similar or even (near) duplicates [10].

Another postretrieval approach is based on combining multiple search results via search result aggregation. The strategy behind this approach comes from the metasearch model in information retrieval, where a query is posed to multiple search engines. Next, the lists of pages that are returned by the search engines are merged and the resulting ranked list is presented to the user [11]. For a query, search results are not necessarily obtained from multiple search engines but from the responses of a single search engine or retrieval model to multiple queries, such as textual and image features. A representative example is video retrieval via multiple modalities, where image features, audio signals, face detection, and caption information can be used to improve text-based video search systems [2, 12].

In contrast to video retrieval, the objective of this paper is to present a simple but effective fusion approach for enhancing the retrieval performance of conventional CBIR systems. Particularly, the proposed aggregation approach is based on combining multiple retrieval results for image reranking, including the original result and different results produced by the PRF procedure. For PRF, using the pseudo-positive feedback alone and both pseudo-positive and pseudo-negative feedbacks at the same time can produce two different reranking results. As a result, we can obtain three different result lists corresponding to a query (i.e. the original result and two reranking results). Next, different aggregation results can be produced based on combining the two or three different result lists. In this paper, the Borda count is applied to combine different result lists for search result aggregation, which is usually used in reranking text retrieval results [13].

Therefore, the contribution of this paper is two-fold. First, a search result aggregation approach is introduced for image search reranking. Second, six different reranking strategies are compared in terms of retrieval accuracies and times over two image datasets, where the optimal strategy can be identified as one representative baseline for image search reranking.

The remainder of this paper is organized as follows: Section II overviews related literature, including studies on image reranking, pseudo-relevance feedback, and search result aggregation. Section III presents the aggregation approach and experimental results. Finally, Section IV presents the conclusions of the paper.

II. LITERATURE REVIEW

A. Image Reranking

In CBIR systems, images are initially indexed based on their visual features via an offline feature extraction process. Then, during retrieval, users can provide example(s) to the system to search for similar images as the query via the example approach. The images that have similar visual features are retrieved and they are ranked based on the level of similarity to the query image. However, the initial search results often contain noise. Image reranking, which reorders images based on related textual and/or visual features, can be employed to improve initial image retrieval results [14].

In the literature, image search reranking approaches can be classified into offline supervised-learning-based methods and online unsupervised-learning-based methods. The offline supervised-learning-based (or classification-based) approaches, which are similar to the task of ‘learning to rank’ [15], are based on training a classifier to assign a relevance score to each image.

In the example of Duan et al. [16], relevant images are clustered by using both textual and visual features; each cluster is regarded as a ‘bag’ and the images in the bag are regarded as ‘instances’. Positive and negative bags are used for classifier training and a ranking score is used to rank all the bags.

In Huang et al. [17], visual saliency and visual consistency are integrated for reranking, where visual saliency assumes that salient images are often relevant to the users’ query since they more easily catch the user’s eyes and visual consistency is based on the concept that visually similar images are closely related to the search query. The saliency model is trained to assign relevance scores to retrieved images to classify them into salient and cluttered classes.

According to Jain and Varma [18], the clicked images that correspond to a query are mostly relevant to the query. Therefore, their reranking method promotes images that are likely to be clicked to the top of the ranked list. The normalized click count for each image is predicted and it is combined with the original ranking score for image reranking.

Wang et al. [3] use an automatic offline learning scheme to associate visual semantic spaces with query keywords based on keyword expansions. The visual image features are mapped into their corresponding visual semantic spaces to obtain semantic descriptors. For online reranking, the similarities between images are measured by the semantic signatures.

In contrast, for online unsupervised-learning-based approaches, Wang et al. [19] propose the ContextRank procedure, which considers the differences in importance between target areas (the main objects) and background areas (the regions without the main objects) in images, namely, if visual words are sufficiently close within an image, their links are constructed for intra-image context. Two images that have the same visual words are constructed by combining both feature similarity and spatial consistency for inter-image context. The score of an image is the sum of the scores of the visual words in the image.

In contrast, for online unsupervised-learning-based approaches, Wang et al. [20], an ordinal reranking approach is proposed, which adjusts the initial ranking list based on cooccurrence patterns, namely, ordinal relationships between target semantics and low-level features that were extracted from the initial ranking list. Their experimental results demonstrate that this approach outperforms representative
baselines, such as RankSVM [21] and ListNet [22], in terms of video retrieval accuracy and efficiency.

In addition to dealing with only image visual features, various works focus on multiple features or modalities simultaneously for search reranking. For example, Gao et al. [8] combine textual (social tags) and visual features to improve keyword-based image search performance. Li et al. [23] use visual and geo-tags (GPS information) for example-based image retrieval. Yao et al. [24] consider visual, textual, and bag-of-words features for boosting retrieval precision.

Although the reranking search results that are obtained by supervised-learning-based reranking approaches are promising, Liu and Mei [12] have shown that these approaches cannot perform optimally and are of very high computational complexity during offline training and online computation for reranking. In addition, the retraining process is necessary if increasingly many images are continually stored in the image database, which is typically the case in Web image search. Moreover, it is highly difficult to collect sufficient training data and to train classifiers for all possible concepts or classes in Web-scale image search reranking [25]. Therefore, we focus on image reranking via online unsupervised learning in this paper. Two related approaches are described hereafter.

B. Pseudo-Relevance Feedback
In addition to the studies that are discussed above, image reranking can be simply approached via pseudo-relevance feedback (PRF), which can be regarded as an online unsupervised learning approach. It is based on the concept of query point movement: The query vector is iteratively modified by pseudo-relevant and pseudo-irrelevant feedbacks to move the query toward more relevant images and away from irrelevant images in the feature space [26].

Relevance feedback can be defined as follows: Given an image database that contains \( n \) images and suppose an interface is provided through which users can issue queries by supplying image examples. Let \( Q \) be a query example and \( I \) an image in the database, where \( Q = (x_1, x_2, \ldots, x_m) \) and \( I = (y_1, y_2, \ldots, y_m) \), from which \( m \) low-level visual features are extracted.

However, from the user’s perspective, not every retrieved image is semantically relevant to the query image \( Q \). If the retrieval result does not satisfy the user, he/she can execute the iterative RF process. After relevance feedback(s) have been provided by users, a new retrieval result that is based on a new ranking list of top similar images can be produced, which can increase the user’s satisfaction level [26, 27].

The Rocchio algorithm is a representative approach for query vector modification [26]. The query is reformulated as a modified query \( \hat{q}_m \) by the following:

\[
\hat{q}_m = \alpha \bar{q}_\alpha + \beta \frac{1}{|D_r|} \sum_{\vec{d}_r \in D_r} \bar{d}_r - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_r \in D_{nr}} \bar{d}_r
\]  

where \( \bar{q}_\alpha \) is the original query vector; \( D_r \) and \( D_{nr} \) are the relevant and nonrelevant image sets, respectively; and \( \alpha \), \( \beta \), and \( \gamma \) are the parameters that are used to adjust the related weights for components \( q_\alpha \), \( D_r \) and \( D_{nr} \). 

\( \alpha \) is the weight that is used to move the original query vector toward a specified direction and \( \beta \) and \( \gamma \) are the weights that are used to reflect the levels of importance of the relevant and irrelevant feedback sets, respectively. Then, the newly reformulated query, which is represented by the modified query vector, is issued to the system to research for similar images based on the Euclidean metric.

Due to the limitation of RF that users must provide positive/negative feedbacks, pseudo-relevance feedback (PRF) can be used to automate the manual part of RF. In PRF, it begins when the relevant images are retrieved based on an initial query. According to the initial retrieval result, a fraction of the top-\( k \) ranked images are assumed to be relevant to the query, which are regarded as pseudo-positive, and some low-ranked images are regarded as pseudo-negative. After the pseudo-positive and pseudo-negative images have been identified, the Rocchio algorithm is used to execute the RF process [9, 29].

A major advantage of PRF is its high computational efficiency for Web image search. However, although the reranking results by PRF may provide higher retrieval performance than the initial search results, the major problem with PRF is that the pseudo feedback set may contain images whose semantic contents are not similar to those of the query image, namely, some visually similar images in the feedback set are not semantically related to the query. Moreover, for Web image search, many images in the feedback set are duplicates or near duplicates due to the high similarity of their visual features.

In practice, a satisfactory search result should not only contain relevant images to the query but also cover a wide range of topics, namely, the retrieval results should be relevant and diverse; these properties correspond to the relevance-based reranking and diversified reranking problems, respectively [30]. In this paper, we mainly focus on the relevance-based reranking problem. This is because PRF is a relevance-based reranking method, which assigns higher rank to more relevant images without considering the coverage of the topics.

C. Search Result Aggregation
Aggregation of multiple search results is another approach for image reranking. It is based on combining the ordered results from multiple retrieval methods or systems via rank aggregation [31]. This approach can be regarded as metasearch in information retrieval [10], namely, the same query is issued to multiple search engines, from which various retrieval results can be obtained. Next, the ranked lists that are produced by these engines are combined. As a result, the new reranked list is likely to increase the precision and the range of topics of the resulting list.
The Borda count procedure is a conventional method for combining multiple retrieval results. It optimally satisfies all the required symmetry properties for information retrieval [11, 32].

The Borda count procedure is a generalization of majority voting. First, each voter (the retrieval system) ranks a fixed set of $c$ candidates (the set of retrieved images). For each voter, $c$ points are assigned to the top-ranked candidate, $c - 1$ points to the second-ranked candidate, and so on. The remaining points are divided evenly among the candidates that are unranked by the voter. Finally, the fixed set of $c$ candidates is ranked according to their total points and these points are used to rerank the list [13].

Suppose that there are three retrieval results, namely, $A$, $B$, and $C$, where their ranked lists of images for a query are $A = (a, c, b, d)$, $B = (b, c, a, e)$, and $C = (c, a, b, e)$.

In total, five distinct images are retrieved: $a, b, c, d,$ and $e$. The Borda count (BC) of each distinct image is computed by summing the Borda count values of the individual results $BC_A$ in the retrieval result ($A$) as follows:

$BC_A(a) = BC_A(a) + BC_A(b) + BC_A(c) = 5 + 3 + 4 = 12$

$BC_A(b) = BC_A(b) + BC_A(c) + BC_A(d) = 3 + 5 + 3 = 11$

$BC_A(c) = BC_A(c) + BC_A(e) + BC_A(e) = 4 + 4 + 5 = 13$

$BC_A(d) = BC_A(d) + BC_A(d) + BC_A(e) = 2 + 2 + 2 = 6$

$BC_A(e) = BC_A(e) + BC_A(e) + BC_A(e) = 0 + 2 + 2 = 4$

Finally, the five distinct images are reranked according to their Borda counts. The final ranked list of images is $c > a > b > e > d$.

III. AGGREGATION APPROACH

The proposed aggregation approach is based on the Borda count procedure and combines retrieval results that are obtained via various methods to produce the final reranking result. Three retrieval results are considered: the original retrieval result without reranking via PRF, the reranked retrieval result according to pseudo-positive feedback, and the reranked retrieval result according to pseudo-positive and pseudo-negative feedbacks.

Figure 1 illustrates the aggregation approach. When a user provides a query image to the retrieval system, the initial ranked retrieval result (retrieval list 1) is obtained based on the similarity between the query vector and the other feature vectors in the database. Next, top-$k$ pseudo-relevance feedback is obtained from retrieval list 1 via the Rocchio algorithm. Only the top-$k$ pseudo-positive feedback set is used, thereby resulting in a new ranked retrieval list (retrieval list 2). At the same time, the top-$k$ pseudo-positive and pseudo-negative feedback sets are used together to produce another ranked retrieval list (retrieval list 3). Finally, retrieval lists 1, 2, and 3 are combined via the Borda count procedure to produce the final reranking result (retrieval list 4).

According to the descriptions in Section 2.3, suppose that the original retrieval result (list 1) for a query is $A = (a, c, b, d)$. Then, after performing PRF for list 1 based on pseudo-positive feedback and pseudo-positive and pseudo-negative feedbacks, two retrieval results (list 2 and list 3) are obtained: $D = (b, a, c, e)$, and $E = (c, a, b, e)$. In total, four aggregation results are obtained via the Borda count procedure: list 1 + list 2, list 1 + list 3, list 2 + list 3, and list 1 + list 2 + list 3 (list 4).

The reranked results that are based on these four aggregation methods are as follows:

- list 1 + list 2 + list 3: $(a, c, b, e, d)$
- list 1 + list 2: $(a, b, c, d, e)$
- list 1 + list 3: $(a, c, b, d, e)$
- list 2 + list 3: $(a, c, b, e, d)$

For the following experiments, these single and aggregation retrieval methods are compared in terms of retrieval effectiveness and efficiency.

IV. EXPERIMENTS

A. Experimental Setup

In this paper, two datasets are used for the experiments that are discussed below. The first dataset is based on NUS-WIDE-LITE [33], which is a smaller version of the NUS-WIDE dataset. It contains many real-world images that have been downloaded from Flickr. Sixty-nine concepts (classes) are selected, which are composed of 22156 images, where each concept corresponds to at least 50 images. The second dataset is based on Caltech 256 [34]. It contains 257 object categories, which are composed of 30607 images. For each

2. Flickr is an image hosting and video hosting website (https://www.flickr.com/).
category, we randomly select 10 images as the query images. For image feature representation, the 500-D BoW feature that is based on the SIFT descriptor (BoW) [16] is extracted from each image.

In addition, we use the Euclidean distance similarity measure for the retrieval system since it is the most widely used distance function in image retrieval systems. For the pseudo-positive and pseudo-negative feedback sets, the top-20 highest ranked images and top-20 lowest ranked images, respectively, of each query are used [34, 35, 36].

Regarding the Rocchio parameters, Moschitti [37] shows that $\beta = \gamma$ is the optimal setting for text retrieval and Tsai et al. [38] compare parameter settings for PRF in images and find that $\alpha$, $\beta$, and $\gamma$ values of 1, 0.5, and 0.5, respectively, yield reasonable performance over various image feature representations. Therefore, we follow the parameter settings of [37]. Moreover, for retrieval efficiency, feedback iteration is only performed once to collect retrieval lists 2 and 3. We found that the performance gradually degrades as additional feedback iterations by PRF are executed, which may be affected by images in the pseudo feedback set whose semantic contents differ from those of the query image. Therefore, other results of PRF that are obtained via additional feedback iterations are not compared.

Finally, the top-100 retrieved images from retrieval lists 1, 2, and 3 are used for result aggregation via Borda count. This is because the reranking of all the retrieved images will incur a very large computational cost and will affect the query response time. Moreover, for each query result, users are typically not concerned with the lower ranked images. Therefore, reranking the top-100 retrieved images from various retrieval results is sufficient.

Consequently, in our aggregation approach, there are four possible combinations of three retrieval lists via Borda count: (1) list 1 + list 2, (2) list 1 + list 3, (3) list 1 + list 2 + list 3, and (4) list 2 + list 3.

**B. Experimental Results**

Figures 2 and 3 present the reranking results of various aggregation strategies on the NUS-WIDE-LITE and Caltech 256 datasets, respectively. From the results of the single retrieval methods (lists 1, 2, and 3), we observe that the original result slightly outperforms the result that was obtained via PRF in terms of the P@10 rates, which are 39.42% vs. 39% on NUS-WIDE-LITE and 35.02% vs. 34.67% on Caltech 256. Performing PRF based on pseudo-positive feedback can provide slightly better P@20 and P@50 rates for the two datasets. However, there is no significant difference between list 1 and list 2 ($p > 0.05$ in the t-test).

In contrast, for the aggregation result, the combination of list 1 and list 2 (list 1 + list 2) performs best, which provides the highest rates of P@10 and P@20 over both datasets. PRF that is based on pseudo-positive feedback (list 2) performs best in terms of P@50 and second-best in terms of P@10 and P@20 on NUS-WIDE-LITE. On Caltech 256, the combination of list 1 and list 3 performs second-best in terms of P@10, P@20, and P@50.

Regarding the results, it is explainable that the best aggregation performance can be obtained by the combination of list 1 and list 2. This is because for lists 1, 2, and 3, the top two retrieval performances are lists 1 and 2, whereas the worst result is based on list 3. Therefore, the aggregation result by lists 1 and 2 should be better than the other combinations that include list 3, i.e. list 1 + list 3, list 2 + list 3, and list 1 + list 2 + list 3.

Table 1 lists the average retrieval performances of these seven methods. The aggregation of lists 1 and 2 is the best choice for image search reranking, which significantly outperforms the others ($p < 0.05$). The exception is list 2 on the NUS-WIDE-LITE dataset, where the performance difference between list 1 + list 2 and list 2 is only 0.04%. Except for the combination of lists 1 and 2, the aggregation results do not outperform the original retrieval result or PRF via pseudo-positive feedback on NUS-WIDE-LITE. However, on the Caltech 256 dataset, the aggregation strategies, except for list 2 + list 3, outperform the three single retrieval methods. This performance difference could be affected by the image content. This may be because NUS-WIDE-LITE contains approximately half real (single object) and half abstract (nonsingle object) categories. For example, a harbor image is an abstract category that is composed of several single objects. In contrast, the categories in Caltech 256 are all single objects. Hence, search result aggregation is much more suitable for images that contain single objects.

Furthermore, we examine the average query response times of various aggregation strategies on the NUS-WIDE-LITE and Caltech 256 datasets. Each query response time is measured from the issuance of a query image example to the output of the final retrieval results. According to Figures 4 and 5, combining multiple retrieval results requires longer time than using single retrieval results. However, although the best performance with list 1 + list 2 requires 0.139 seconds on NUS-WIDE-LITE and 0.275 seconds on Caltech 256, which are longer compared to list 1, list 2, and list 3, for users this performance difference is very small because it is difficult to clearly differentiate between these computation times during retrieval. Hence, the query response times by list 1 + list 2 are sufficiently short. In addition, if we consider fewer images for reranking, the query response time will decrease substantially.

Therefore, according to the retrieval performance and efficiency, combining the original retrieval result and the result that is obtained by performing PRF based on pseudo-positive feedback can yield the best reranking result.

---

4 The software is based on Matlab 7 on an Intel Pentium 4 computer with a 2.8-GHz CPU and 20GB RAM.
FIGURE 2. Reranking performances on NUS-WIDE-LITE.

FIGURE 3. Re-ranking performances on Caltech 256.

FIGURE 4. Average query response times on NUS-WIDE-LITE.

FIGURE 5. Average query response times on Caltech 256.
VII. CONCLUSIONS

In this paper, we present a hybrid approach for image search reranking that improves the retrieval performance of CBIR systems and PRF. This approach is based on search result aggregation. Three types of single retrieval results are combined: the original retrieval result, the result of PRF by pseudo-positive feedback, and the result of PRF by pseudo-positive and pseudo-negative feedbacks.

Two datasets, namely, a small version of NUS-WIDE-LITE and Caltech 256 (c.f. Section 3.1) are used in our experiments. The experimental results demonstrate that search result aggregation can provide higher retrieval performance than the single retrieval results. The combination of the original result and the result of PRF by pseudo-positive feedback performs the best. Although the average retrieval time per query via search result aggregation is longer compared to the single retrieval methods, the query response time is still very short for reranking the top-100 images of various retrieval results, namely, 0.139 seconds for the NUS-WIDE-LITE dataset and 0.275 seconds for the Caltech 256 dataset.

Since this paper has shown the potential of search result aggregation in image retrieval, some future works should be considered in the future. That is, some related factors may affect the image reranking result. First, various distance functions, such as the cosine measure, could be compared for image retrieval. Second, supervised learning techniques, such as cotraining [39] and estimation [40], can be employed for search result cooperation. Third, different local, global, and other related features can be extracted for better feature representation [41, 42, 43].

ACKNOWLEDGMENTS

The work was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST 106-2410-H-182-024 and in part by Chang Gung Memorial Hospital, Linkou, Taoyuan, Taiwan under Grant NERP2D0301.

REFERENCES


<table>
<thead>
<tr>
<th></th>
<th>list 1</th>
<th>list 2</th>
<th>list 3</th>
<th>list 1 + list 2</th>
<th>list 1 + list 3</th>
<th>list 2 + list 3</th>
<th>list 1 + list 2 + list 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUS-WIDE-LITE</td>
<td>36.41%</td>
<td>38.13%</td>
<td>26.59%</td>
<td>38.17%</td>
<td>31.35%</td>
<td>33.84%</td>
<td>35.99%</td>
</tr>
<tr>
<td>Caltech 256</td>
<td>29.14%</td>
<td>29.5%</td>
<td>25.76%</td>
<td>31.46%</td>
<td>30%</td>
<td>28.98%</td>
<td>30.14%</td>
</tr>
</tbody>
</table>


