COMPARISON OF CLUSTERING APPROACHES FOR SUMMARIZING LARGE POPULATIONS OF IMAGES

Yushi Jing[*]*, Michele Covell, Henry A. Rowley*

Google Inc., Mountain View, California, United States of America

ABSTRACT

This paper compares the efficacy and efficiency of different clustering approaches for selecting a set of exemplar images, to present in the context of a semantic concept. We evaluate these approaches using 900 diverse queries, each associated with 1000 web images, and comparing the exemplars chosen by clustering to the top 20 images for that search term. Our results suggest that Affinity Propagation is effective in selecting exemplars that match the top search images but at high computational cost. We improve on these early results using a simple distribution-based selection filter on incomplete clustering results. This improvement allows us to use more computationally efficient approaches to clustering, such as Hierarchical Agglomerative Clustering (HAC) and Partitioning Around Medoids (PAM), while still reaching the same (or better) quality of results as were given by Affinity Propagation in the original study. The computational savings is significant since these alternatives are 7-27 times faster than Affinity Propagation.

***Keywords*—** Web image summarization, clustering, k-medoids

1. INTRODUCTION

The goal of commercial image search engines, such as MSN, Yahoo and Google, is to retrieve and present a set of images that best represent the semantic or visual concepts and categories of the text query. For example, the top search results for the query "eiffel tower," contain that structure under various viewpoint and lighting conditions. Collectively these images can be considered as *exemplars* that *summarize* the visual concepts associated with a text query, as shown in Figure 1.

In Section 2, we review recent work in presenting visual summaries of large image collections, focusing on exemplar-based methods. These exemplar-based approaches have been found to be an efficient and intuitive method for representing the larger population of results. While statistical methods, such as principal or independent components, are often more compact descriptions of populations, exemplars provide a natural way to point to portions of a population of images, making it better for tasks that require quick human understanding. The goal of this paper is to compare and evaluate alternative clustering methods as a way to select such exemplars.

[*]Correspondence should be addressed to Yushi Jing (jing@google.com).

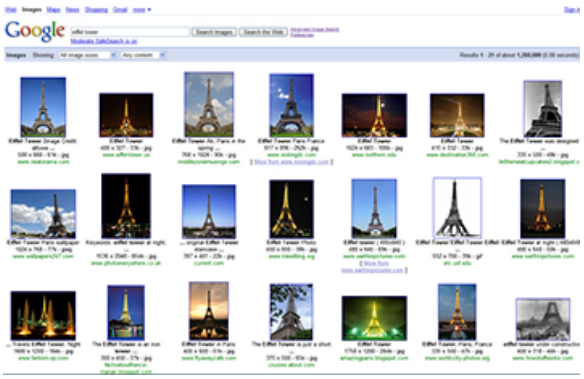

Fig. 1. The top search results can be considered as an approximation to what an ideal image summarization engine should generate given an image corpus.

Many alternative approaches to examplar selection have been proposed [3] [8] [10]. In Section 3, we discuss three commonly used exemplar selection methods: Hierarchical Agglomerative Clustering (HAC), Partition Around Medoids (PAM) [8], and Affinity Propagation [3]. These and other approaches to exemplar selection all seem promising but are difficult to evaluate and compare, due to a lack of a clear and agreed-upon definition of what constitutes a good summarization. In this paper, we propose a pragmatic evaluation of what constitutes a good summarization result, based on comparison to the top results from commercial search engines. This approach avoids the manual labelling of image populations, as it becomes prohibitively expensive when applied to very large image sets. Instead, by using top search results from commercial search engines, we can take advantage of the years of experience, design effort, and background information that these systems incorporate into their final search-result ordering. We discuss the details of the evaluation method in Section 4 and our results in Section 5.

While the quality of the cluster-based exemplar results is surprisingly good, the computational cost of the best clustering approaches is prohibitively expensive. To address this problem, we extend our basic idea, using a simple distribution-based selection filter on a larger number of cluster/exemplar pairs. This extension improves the quality of all of our clustering approaches. This improvement means that we can use the least computationally intensive approach (HAC) while keeping the exemplar-set quality as high as the most computationally inten-

978-1-4244-7493-6/10/$26.00 ©2010 IEEE

sive approach (Affinity Propagation). We describe this approach in Section 6 and present our results in Section 7.

Finally, in Section 8, we consider some of the probable causes underlying our findings and propose future extensions.

2. APPROACHES TO WEB IMAGE-COLLECTION SUMMARIZATION

Given the explosive growth of images and other multimedia information accessible online, techniques for summarization Web image has generated significant interest. Several recent studies have explored the use of online image hosting site such as Flickr [12], or from commercial search engines [11] [6] [5]. In particular, Google Image Swirl [7] summarizes and visualizes search results in the form of exemplar tree. Due to the subjective nature of task, lack of good evaluation criteria was the major drawbacks of prior approaches. This paper provides a comprehensive experiment study on the various clustering algorithm used for image summarization.

Several works [1] [4] have been proposed from the information retrieval community on image summarization and representation, using textual caption data or geo tagging for image summarization. For example, Clough et al. [1] construct a hierarchy of images using only textual caption data, and the concept of subsumption. However, none of them take advantage of the visual information in the images to fill in for bad or missing metadata.

3. CURRENT APPROACHES TO CLUSTER-BASED EXEMPLAR SELECTION

In this paper, we compare clustering-based methods as a way to select K summarizing exemplars from a population of images. Our final evaluation will ultimately be based on comparing our results to the images that (based on relevance feedback) are most generally useful for a query term. However, we need an intermediate metric for clustering against which we can operate in our unsupervised processing of new image sets. We set as this intermediate goal finding the K *exemplars* so as to best represent the full image set, using only K images taken from that set. We can formalize the measure of "best" by implicitly associating all non-exemplar images in the set with one exemplar each and considering the similarity between the full population and their associated exemplars.

Describing this mathematically, each image $I_n \forall n = 1...N$ in the full population associates itself with a single exemplar image taken from the exemplar set $C = \{c_1...c_K\}$ where c_k are the indices in the full population of the selected exemplars. We denote this association of image I_n with exemplar I_{c_k} as $c_k = L(n|C)$. Our exemplar selection process maximize a similarity function over C:

$$F(C) = \sum_{n=1}^{N} S(I_n, I_{L(n|C)}) + \sum_{n=1}^{N} \delta_n(L(n|C)) \qquad (1)$$

where $S(I_i, I_k)$ is a similarity measure between image I_i and I_k. We define the second term to insure a distinct cluster is uniquely associated with each exemplar:

$$\delta_n(k) = \begin{cases} -\infty, \text{if } n \in C \text{ and } n \neq k; \\ 0 \text{ otherwise.} \end{cases} \qquad (2)$$

Clustering consists of maximizing $F(C)$ over C. This single formulation describes a class of solutions called K-medoids since, given a partitioning of the full set into K disjoint subsets, the best exemplar for each of these subsets is the "most central" image in that set. K-medoids is somewhat similar to k-means but works directly on the similarity matrix across images, instead of operating in the pixel or image-feature space. We discuss this point further in Section 4.

Partition Around Medoids (PAM) [8] is the most-often used k-medoids method. Like k-means, it starts with an arbitrary selection of the K exemplars out of N data points. Using the image similarity matrix, it translates that subset selection into a K-way partition. Unlike K-means, the re-centering step also includes a random component: the method proposes a random non-exemplar image to replace the exemplar previously associated with that image. The proposed swap is accepted if the new partition using this modified exemplar set has a higher similarity value $F(C)$ than the original exemplar set.

Another long-standing heuristic approach to exemplar selection is to partition the data via clustering methods such as Hierarchical Agglomerative Clustering (HAC), and select exemplars from each of the clusters. HAC starts with each image in a separate cluster. The number of clusters is then reduced by "agglomerating" clusters, by merging the 2 (or more) clusters that are considered most similar. The exemplar images for the final set of clusters is selected to be the most central image, using the within-cluster distance measure. For the results reported in this paper, we used average image similarity within a candidate merged cluster as our measure of which clusters are closest.

Affinity Propagation [3], like HAC, is a bottom-up method. Unlike HAC, it has the advantage of explicitly selecting the best exemplar for each partition as part of the partitioning process. Affinity Propagation is derived as an instance of the max-sum algorithm in a factor graph describing the constraints on the labels and the energy function.

4. EXPERIMENTAL SET-UP

As mentioned in Section 1, our goal is to evaluate the quality of the summarization provided by a small exemplar set, for representing a particular population of images related to a semantic concept. This problem implicitly involves two distinct optimizations: first, finding a similarity measure $S(I_n, I_k)$ that captures the semantics of similarity for the conceptual query and, second, given that similarity measure, selecting a small set of exemplars that best summarize the full set. Separating these two pieces into distinct optimizations has the advantage allowing us to separate the problem of finding locally accurate similarity measures from the task of finding globally accurate summarizations. Locally accurate similarity measures

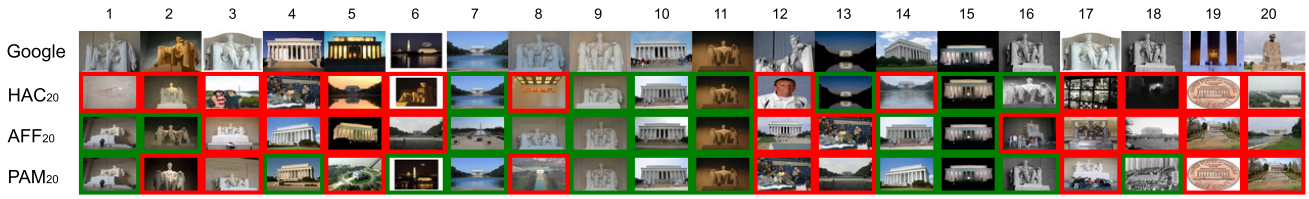

Fig. 2. Maximum bipartite matching [9] between the top 20 images returned by Google Image search for "Lincoln Memorial" and the exemplars returned by HAC_20, Affinity Propagation (AFF_{20}), and PAM_20. Bipartite pairs that are closely matched are shown with a green border; mismatched pairs have a red border.

can then be created separately, according to concept or application. This fits well with recently proposed approaches to defining distance measures that insure local relative distance orderings, without explicitly training for long-range distances. Our study is aimed at determining which heuristic exemplar-selection approach (HAC, PAM, or Affinity) provides the best globally accurate summarization, given a distance measure that shows high-quality local distance comparisons.

Since we are not currently focusing on determining locally accurate distance measures, we use visual similarities (or negative distances) computed as part of a different image-similarity application. The measure we used combines color, texture, and spatial structure measures, using a discriminative learning approach. Since the same image similarity measure is used as input to all of our clustering evaluations and since that measure was not explicitly optimized for our application, we believe that our evaluation of the relative merits of the alternative clustering approaches will carry over to its use in conjunction with other distance measures.

For this paper, we evaluate using results from a single search engine (Google image search) and we take our terms from the most popular queries, so that we can have some assurance that the results returned for these queries have been the focus of considerable engineering effort. To avoid terms for which there is no visual correlate, we limit this study to 900 popular celebrity-, location-, and product-related queries.

For each selected query, we provided the clustering algorithms with up to 1000 images that were returned in response to that query. Each clustering algorithm was run in order to select the top 20 exemplar images from this set. These selected images were then compared to the images that were returned on the top pages of the image search results. We completed the comparison using maximum bipartite graph matching [9] with the edge weights defined by the visual similarity. Using this matching process we counted how many of the paired results match with exact or near duplicates. The similarity threshold that we used was selected to as a compromise that seemed to best capture our sense of near-duplicate images. Most of the close-match pairs are semantically similar, as well as visually similar but there are some exceptions (e.g., column 18 in the figure). On average across all of the query terms, 19.8% of the image populations are covered by the image-search results, using our near-duplicate threshold. Using random selection for our 20 exemplars, we would expect a match rate of 17%.

5. RESULTS USING 20 CLUSTERS

Figure 2 shows an example of the bipartite matches found for K=20 using HAC, Affinity Propagation (AFF), and PAM. For bottom-up clustering approaches like AFF and HAC, we adjusted the parameter to obtain clustering results of approximately 20 clusters. The number of close-match pairings from the bipartite graphs between the top-20 image search results and the clustering approaches is listed at the right. On each of the bipartite pairs, close matches are marked according to whether our (locally accurate) visually similarity measure is below a predefined threshold. In Figure 2, close pairs are marked with a green border, while more distant pairs are marked with a red border.

We used this count of the close-match pairings between the image-search results and the clustering approaches, in order to determine the average matching rates. This average was computed over our set of 900 queries. These averages were: 21.5% for HAC; 33.8% for PAM; and 34.2% for Affinity Propagation.

At first glance, the percentage of close matches might seem low (all less than 40%), due to the additional (non-visual) information that goes into the ordering of images being returned by the search engine, including the content and reputation of the referring pages. In light of that additional information, the percentages of close matches is actually quite high for PAM and Affinity Propagation.

We used Wilcoxon Signed Rank test [2] to determine which, if any, of the differences in average close-match rates were statistically significant. We used this measure, instead of the more widely know Student *t* test, since the pairings are not drawn from a normal distributions as required by the Student test.

As expected from these average close-match rates, there is a statistically significant improvement from HAC both to PAM and to Affinity Propagation. The improvement provided by Affinity Propagation over PAM is also statistically significant.

Using an average close-match-rate metric for comparing our alternative approaches could be susceptible to bias from a small number of extreme outlier terms. To avoid that bias, we also analyzed our alternative approaches by counting the numbers of terms that an approach did better, worse, or the same as an alternative approach. To change these counts into a simple summary statistic, we then take the difference in the better and worse per-

Table 1. Percentage of queries that were improved/worsened by each clustering approach, compared to the alternatives. This table provides a summary of how many query terms had did better for each clustering approach than the reference approach listed in the column. See Section 4 for additional information on how these percentages were computed.

	HAC$_{20}$	**AFF**$_{20}$	**PAM**$_{20}$	**HAC**$_{40}$	**AFF**$_{40}$	**PAM**$_{40}$	**HAC**$_{80}$	**AFF**$_{80}$	**PAM**$_{80}$	**Average**	
HAC$_{20}$	0	-59%	-72%	-67%	-68%	-69%	-55%	-53%	-64%	-63%	**HAC**$_{20}$
AFF$_{20}$	59%	0	0%	-5%	-21%	-3%	10%	3%	-3%	5%	**AFF**$_{20}$
PAM$_{20}$	72%	0%	0	-7%	-17%	-5%	11%	6%	-4%	7%	**PAM**$_{20}$
HAC$_{40}$	67%	5%	7%	0	-14%	3%	18%	12%	3%	13%	**HAC**$_{40}$
AFF$_{40}$	68%	21%	17%	14%	0	16%	28%	22%	16%	25%	**AFF**$_{40}$
PAM$_{40}$	69%	3%	5%	-3%	-16%	0	16%	7%	1%	10%	**PAM**$_{40}$
HAC$_{80}$	55%	-10%	-11%	-18%	-28%	-16%	0	-8%	-13%	-6%	**HAC**$_{80}$
AFF$_{80}$	53%	-3%	-6%	-12%	-22%	-7%	8%	0	-8%	0%	**AFF**$_{80}$
PAM$_{80}$	64%	3%	4%	-3%	-16%	-1%	13%	8%	0	9%	**PAM**$_{80}$

centages for each pair of methods. Using this measure with HAC as the baseline, Affinity Propagation did better on a net of 59% of the full 900-term set and PAM did better on a net of 72% of the full 900-term set. However there was no net term-contest difference between Affinity Propagation and PAM, when they were compared directly to one another: for many of those terms on which HAC is the worst, Affinity Propagation does better than PAM, bringing the term-contest comparison back into balance. This term-contest information, taken in conjunction with the earlier average close-match statistics, show that the Affinity Propagation is doing very well on a subset of queries (as shown by the high average close-match statistic) but that PAM is providing a more consistent across-query improvement to the quality (as shown by the higher win rate in the term contests).

Based on these results, Affinity Propagation and PAM provide the best exemplar selection, being significantly better than HAC in terms of both average close-match performance and term contests. However, these improvements are achieved through the use of significant computational resources. On average, Affinity Propagation and PAM take 160 sec and 22 sec per term, respectively, while HAC takes 6 sec per term. These differences in clustering time are a significant barrier for any large-scale application. In the next section, we propose an extension, mainly aimed at improving the performance of the simpler clustering methods without increasing their computational cost.

6. DISTRIBUTION-BASED FILTERING OF CLUSTERING RESULTS

Our HAC clustering results are disappointing, being only 24% better than random selection. When we examined the clusters and exemplars that were supporting this approach for some of the queries, we found that this poor performance could be explained by way the later levels of the HAC process were agglomerating clusters: the final clusters had non-compact support. One contributing factor is that the image-similarity measure that we are using was created to accurately reflect local distances. When we get to the top levels of the cluster hierarchy, the image distances are beyond the accurate range for our similarity measure.

Since the early clusters do correspond to intuitive groupings, we modified our approach to cluster-based exemplar selection. We start in the same way as in the earlier experiments, but, instead of forcing the clustering to reduce to only 20 clusters, we clustered our input set into K clusters (giving K exemplars) for $K > 20$. We sort this set of K exemplars based on the size of the supporting cluster, with the exemplars that correspond to the largest clusters getting the highest priority. We then use the top 20 exemplars from this sorted list in the same experimental set up that was described in Section 4.

7. RESULTS USING FILTERED CLUSTER-BASED EXEMPLARS

Tables 1 and 2 give our average close-match counts from the bipartite graphs and the results from term-count contests, respectively. In Table 1, the final non-label column gives an average term-contest statistic across all non-identity pairs of methods.

We include for comparison the xxxt$_{20}$ results. As discussed in the previous section, the xxx$_K$ for $K > 20$ results are based on the bipartite graphs between the top image search results and the top-population-cluster exemplars. Since all of the approaches select their 20 images before creating a bipartite graph, the performance across the approaches *are* directly comparable, even across the different size designations: there is no built in advantage for the filtered approaches.

All of the filtered clustering approaches give encouragingly high close-match rates and much better term-contest win rates. Affinity Propagation is still better than the alternatives, by a statistically significant margin. However, the filtering brings the performance of all of the xxx$_{40}$ up to or above the performance of pure 20-cluster Affinity Propagation. This is an important achievement, since it allows to achieve that level of performance without paying the computational cost of Affinity Propagation.

8. CONCLUSIONS AND FUTURE WORK

Our results suggest that, at least for a subset of concepts that have a strong visual component, cluster-based selection of exemplars can get surprisingly close to the same distribution of top images as do rankings that incorporate a large amount of

Table 2. Average rate of close-match images between each of the clustering and filtering approaches and the top-20 image-search results.

	HAC	**AFF**	**PAM**
xxx_{20}	21.5%	34.2%	33.8%
xxx_{40}	34.9%	36.6%	34.4%
xxx_{80}	32.3%	33.3%	34.5%

non-local, non-visual data about each image. Within the set of cluster-based methods that we looked at Affinity Propagation and PAM did significantly better than HAC but at significantly higher computational cost. While all of the approaches benefited by increasing the number of exemplars/clusters that were generated and then filtering based on ranked cluster size, HAC benefited the most. The quality of the filtered HAC exceeds that of all of the unfiltered methods, while using less than 4% of the computation needed for Affinity Propagation and and less than 20% of that for PAM.

This study leaves several questions for future work. First, similar to affinity propagation, spectral clustering and kernel methods can be useful to cluster features bounded by non-linear decision space, and we hope to conduct such experiment in the future. Another question is whether the exemplar approach is applicable to non-search-term related collections of images (such as collections of photos taken by friends and family). Finally, we should evaluate how much of the improvement seen with the filtered exemplar supersets (e.g. $K = 40$) is due to the specific image similarity measure that we used and how much is fundamental to diverse image collections.

9. REFERENCES

[1] P. Clough and D. Petrelli. Using concept hierarchies in text-based image retrieval: A user evaluation. *Lecture Notes in Computer Science (LNCS)*, 4022:297–306, 2006.

[2] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[3] D. Dueck and B. Frey. Non-metric affinity propagation for unsupervised image categorization. In *Proc. 11th IEEE International Conference on Computer Vision (ICCV)*, 2007.

[4] A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries for large collections of geo-referenced photographs. In *Proc. of the 15th international conference on World Wide Web*, pages 853–854, New York, NY, USA, 2006. ACM.

[5] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(11):1877–1890, November 2008.

[6] Y. Jing, S. Baluja, and H. Rowley. Canonical image selection from the web. In *Proc. 6th International Conference on Image and Video Retrieval (CIVR)*, pages 280–287, 2007.

[7] Y. Jing, H. Rowley, and M. Covell. Visualizing image search via google image swirl. In *Proc. NIPS workshop on Statistical Machine Learning for Visual Analytics*, 2009.

[8] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, 1990.

[9] L. Laszlo and M.D. Plummer. *Matching Theory*. North-Holland, 1986.

[10] A. Tardos D.B. Shmoys M. Charikar, S. Guha. A constant-factor approximation algorithm for the k-median problem. *Journal and Computer and System Science*, 65(1):129–149, 2002.

[11] P. Perona R. Fergus and A. Zisserman. A visual category filter for Google images. In *Proc. 8th European Conference on Computer Vision (ECCV)*, pages 242–256, 2004.

[12] I. Simon, N. Snavely, and S. M. Seitz. Scene summarization for online image collections. In *Proc. 11th International Conf. on Computer Vision (ICCV)*, 2007.

[13] Yangqiu Song, Wenyen Chen, Hongjie Bai, Chih-Jen Lin, and Edward Y. Chang. Parallel spectral clustering. In *ECML 2008*, pages 374–389.